Implicit Regularization in Deep Matrix Factorization

Sanjeev Arora [†‡] **Nadav Cohen** [§] Wei Hu [†] Yuping Luo [†]

[†] Princeton University [‡] Institute for Advanced Study [§] Tel Aviv University

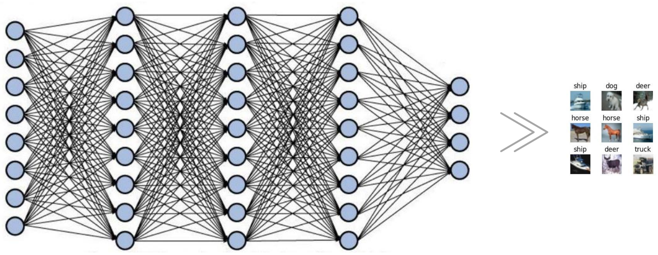
Neural Information Processing Systems (NeurIPS) 2019

Supported by: NSF, ONR, Simons Foundation, Schmidt Foundation, Mozilla Research, Amazon Research, DARPA, SRC

Mystery

DNNs generalize with no explicit regularization even when:

$$\#\text{ of learned weights} \gg \#\text{ of training examples}$$

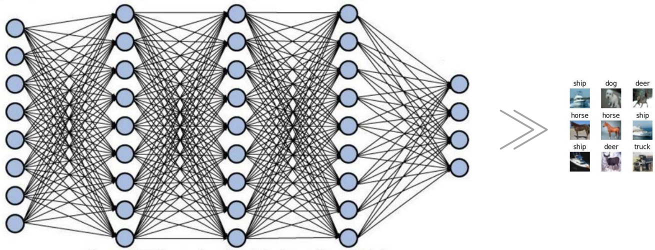


Implicit Regularization in Deep Learning

Mystery

DNNs generalize with no explicit regularization even when:

$$\# \text{ of learned weights } \gg \# \text{ of training examples}$$



Conventional Wisdom

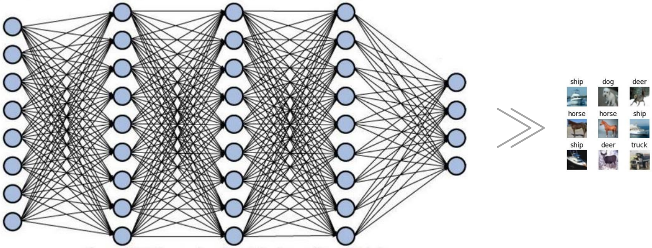
Gradient-based optimization induces an implicit regularization

Implicit Regularization in Deep Learning

Mystery

DNNs generalize with no explicit regularization even when:

$$\# \text{ of learned weights } \gg \# \text{ of training examples}$$



Conventional Wisdom

Gradient-based optimization induces an implicit regularization

Question

Can we mathematically understand this effect in concrete settings?

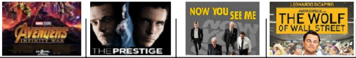
Matrix completion: recover low rank matrix given subset of entries

				
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

Netflix Prize

Setting: Matrix Completion

Matrix completion: recover low rank matrix given subset of entries



Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

Netflix Prize

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Setting: Matrix Completion

Matrix completion: recover low rank matrix given subset of entries

				
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

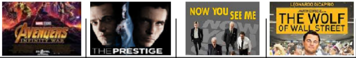
[Netflix Prize]

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Setting: Matrix Completion

Matrix completion: recover low rank matrix given subset of entries



Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

Netflix Prize

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Minimize ℓ_2 loss + nuclear norm regularization:

$$\min_W \sum_{(i,j)\in\Omega}(W_{ij} - b_{ij})^2 + \lambda \cdot \|W\|_{nuclear}$$

Setting: Matrix Completion

Matrix completion: recover low rank matrix given subset of entries

				
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

Netflix Prize

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Minimize ℓ_2 loss $+$ nuclear norm regularization:

$$\min_W \ \sum_{(i,j)\in\Omega}(W_{ij} - b_{ij})^2 + \lambda \cdot \|W\|_{nuclear}$$

Provably "optimal"[1]

[1] Cf. Candes & Recht 2008

Setting: Matrix Completion

Matrix completion: recover low rank matrix given subset of entries

				
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

Netflix Prize

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Minimize ℓ_2 loss $+$ nuclear norm regularization:

$$\min_W \; \sum\nolimits_{(i,j)\in\Omega}(W_{ij} - b_{ij})^2 + \lambda \cdot \|W\|_{nuclear}$$

Provably "optimal"[1] $\leftarrow$ if observations are sufficiently many

[1] Cf. Candes & Recht 2008

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network[1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \ldots W_N} \sum_{(i,j) \in \Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network[1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \ldots W_N} \sum_{(i,j) \in \Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

$$\uparrow$$

No explicit regularization!

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network[1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \ldots W_N} \sum_{(i,j) \in \Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

$\uparrow$

No explicit regularization!

Past Work *(Gunasekar et al. 2017)*

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network[1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \dots W_N} \sum_{(i,j) \in \Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

No explicit regularization!

Past Work *(Gunasekar et al. 2017)*

For depth 2 only:

$$\min_{W_1, W_2} \sum_{(i,j) \in \Omega} \left((W_2 W_1)_{ij} - b_{ij} \right)^2$$

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network [1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \dots W_N} \sum_{(i,j)\in\Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

$\uparrow$

No explicit regularization!

Past Work *(Gunasekar et al. 2017)*

For depth 2 only:

$$\min_{W_1, W_2} \sum_{(i,j)\in\Omega} \left((W_2 W_1)_{ij} - b_{ij} \right)^2$$

- Experiments: recovery often accurate

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network[1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \ldots W_N} \sum_{(i,j) \in \Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

$\uparrow$

No explicit regularization!

Past Work *(Gunasekar et al. 2017)*

For depth 2 only:

$$\min_{W_1, W_2} \sum_{(i,j) \in \Omega} \left((W_2 W_1)_{ij} - b_{ij} \right)^2$$

- Experiments: recovery often accurate
- Conjecture: implicit regularization = nuclear norm minimization

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

Deep Matrix Factorization

Deep Learning Approach ("**deep matrix factorization**")

Parameterize by depth N linear neural network[1] and minimize ℓ_2 loss with gradient descent (GD):

$$\min_{W_1 \dots W_N} \sum_{(i,j) \in \Omega} \left((W_N W_{N-1} \cdots W_1)_{ij} - b_{ij} \right)^2$$

$$\uparrow$$

No explicit regularization!

Past Work *(Gunasekar et al. 2017)*

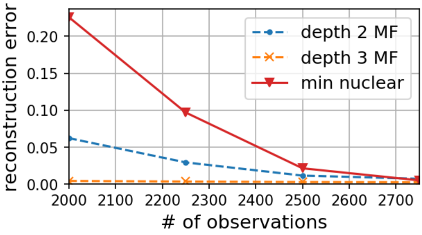
For depth 2 only:

$$\min_{W_1, W_2} \sum_{(i,j) \in \Omega} \left((W_2 W_1)_{ij} - b_{ij} \right)^2$$

- Experiments: recovery often accurate
- Conjecture: implicit regularization = nuclear norm minimization
- Theorem: conjecture holds for certain restricted setting

[1] Cf. Saxe et al. 2014, Kawaguchi 2016, Advani & Saxe 2017, Hardt & Ma 2017, Laurent & Brecht 2018, Gunasekar et al. 2018, Ji & Telgarsky 2019, Lampinen & Ganguli 2019

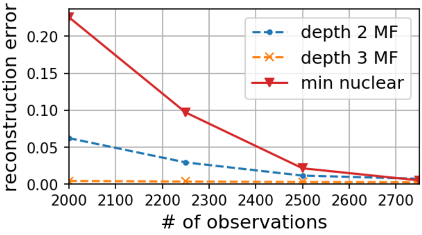
Our Results



Experiments

Depth $\geq$ 3 outperforms depth 2 outperforms nuclear norm minimization

Our Results



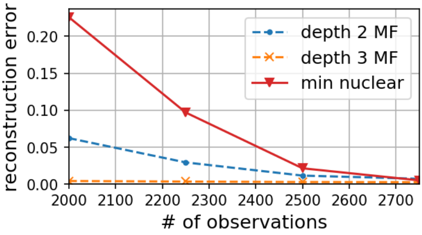
Experiments

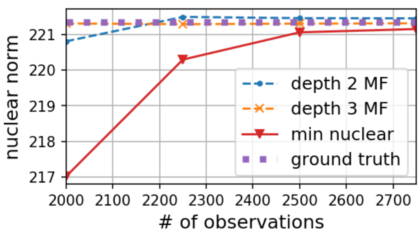
Depth ≥ 3 outperforms depth 2 outperforms nuclear norm minimization

Theory & Experiments

Evidence that:

Our Results





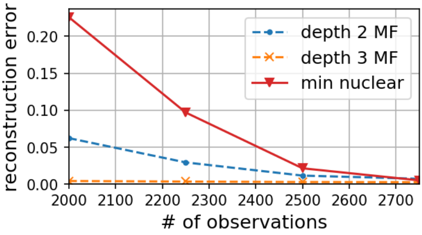
Experiments

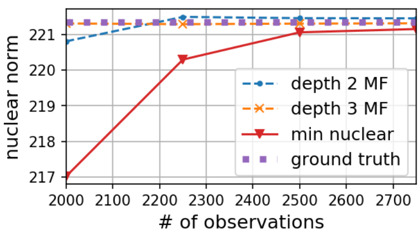
Depth ≥ 3 outperforms depth 2 outperforms nuclear norm minimization

Theory & Experiments

Evidence that:

- Implicit regularization $\neq$ nuclear norm minimization





Experiments

Depth $\geq$ 3 outperforms depth 2 outperforms nuclear norm minimization

Theory & Experiments

Evidence that:

- Implicit regularization $\neq$ nuclear norm minimization
- Capturing implicit regularization via single norm may not be possible

Our Results (cont')

Our Results (cont')

Theory & Experiments

Trajectory analysis for GD over deep matrix factorizations:

Our Results (cont')

Theory & Experiments

Trajectory analysis for GD over deep matrix factorizations:

- Depth makes singular vals move slower when small, faster when large

Theorem

With depth N (and small init) each singular val $\sigma_r(t)$ evolves $\propto {\sigma_r}^{2-2/N}(t)$

Our Results (cont')

Theory & Experiments

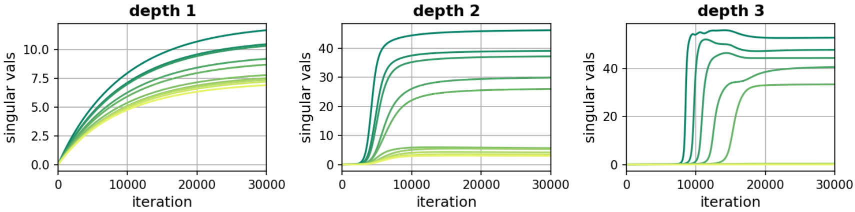
Trajectory analysis for GD over deep matrix factorizations:

- Depth makes singular vals move slower when small, faster when large

Theorem

With depth N (and small init) each singular val $\sigma_r(t)$ evolves $\propto \sigma_r^{2-2/N}(t)$

- Leads to larger gaps between singular vals



Theory & Experiments

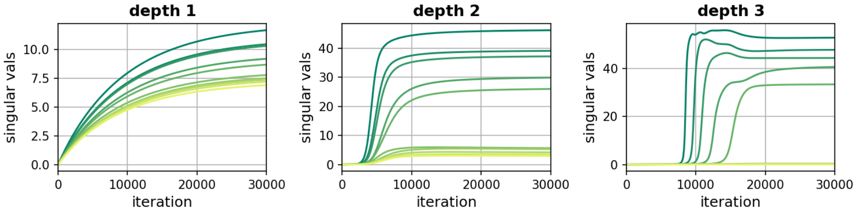
Trajectory analysis for GD over deep matrix factorizations:

- Depth makes singular vals move slower when small, faster when large

Theorem

With depth N (and small init) each singular val $\sigma_r(t)$ evolves $\propto \sigma_r{}^{2-2/N}(t)$

- Leads to larger gaps between singular vals $\implies$ lower rank!



Our Results (cont')

Theory & Experiments

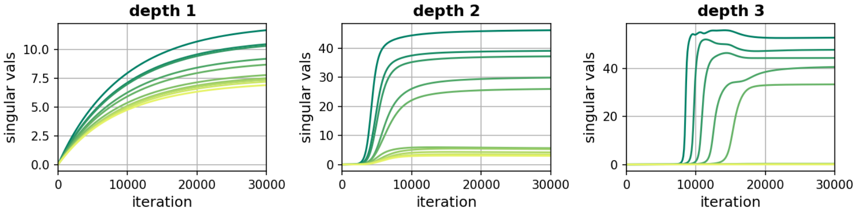
Trajectory analysis for GD over deep matrix factorizations:

- Depth makes singular vals move slower when small, faster when large

Theorem

With depth N (and small init) each singular val $\sigma_r(t)$ evolves $\propto \sigma_r^{2-2/N}(t)$

- Leads to larger gaps between singular vals $\implies$ lower rank!



See our poster: Thu 10:45AM–12:45PM, #245

Our Results (cont')

Theory & Experiments

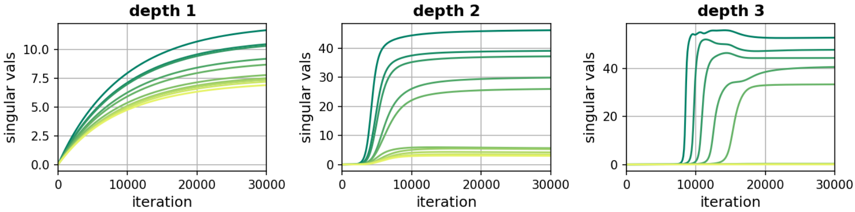
Trajectory analysis for GD over deep matrix factorizations:

- Depth makes singular vals move slower when small, faster when large

Theorem

With depth N (and small init) each singular val $\sigma_r(t)$ evolves $\propto \sigma_r^{2-2/N}(t)$

- Leads to larger gaps between singular vals $\implies$ lower rank!



See our poster: Thu 10:45AM–12:45PM, #245

THANK YOU!