Implicit Regularization in Deep Learning May Not Be Explainable by Norms

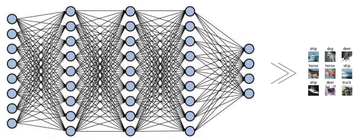
Noam Razin **Nadav Cohen**

Neural Information Processing Systems (NeurIPS) 2020

Tel Aviv University

I) Implicit Regularization in Deep Learning (DL)

DNNs generalize w/o explicit regularization when # of weights $\gg$ training set size



Conventional wisdom: gradient descent (GD) induces an implicit regularization

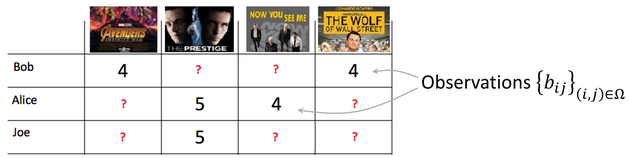
Motivating example: in linear regression, when # weights $>$ # training examples GD initialized at 0 converges to min ℓ_2 norm solution

Widespread hope: GD in DL finds solutions minimizing some **norm**

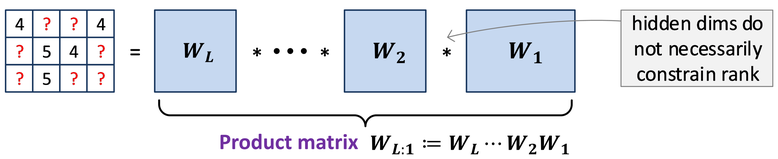
$$\operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\| \text{ s.t. } \mathbf{w} \text{ is global min}$$

II) Case Study: Matrix Factorization

Matrix completion: recover low-rank matrix given subset of entries



Parameterize solution as linear neural network and minimize ℓ_2 loss with GD:



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

III) Open Question

Does the implicit regularization in matrix factorization minimize a norm?

Opposing Conjectures

- Gunasekar et al. 2017: nuclear norm[1] is minimized
- Arora et al. 2019: no norm is minimized

[1] Sum of singular values

IV) Our Main Contribution: Resolving Open Question

Theorem

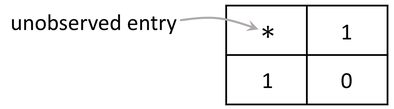
There exist matrix factorization settings where:

- ***All* norms (and quasi-norms) are driven towards ∞**
- **Rank is essentially minimized**

Implicit regularization in matrix factorization $\neq$ norm minimization

V) Analysis: Implicit Regularization Can Drive All Norms to Infinity

A Simple Matrix Completion Problem



construction is easily generalized to arbitrary dimensions and different configurations of observed entries

What are the min norm solutions?

- Min nuclear/Frobenius/spectral norm solution $\iff * = 0$
- Minimizing an arbitrary norm (or quasi-norm) requires value of $*$ to be bounded

When is rank minimized?

- When $|*| \to \infty$ distance from rank 1 is minimized
- When $* = 0$ distance from rank 1 is maximal

We construct settings where norm and rank minimization are contradictory

Theorem

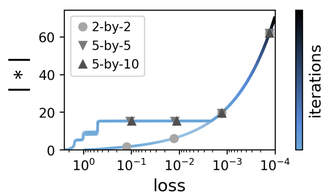
Under gradient flow (GD with learning rate $\to 0$), if $\det(W_{L:1}(0)) > 0$ (holds w.p. 0.5 over standard inits), then $|*| \geq \Omega(1/\sqrt{loss(t)})$. This implies:

- $\|W_{L:1}(t)\| \geq \Omega(1/\sqrt{loss(t)})$, for any $\|\cdot\|$
- Distance from rank 1 of $W_{L:1} \leq \mathcal{O}(\sqrt{loss(t)})$

GD drives all norms to infinity in favor of minimizing rank

Experiment

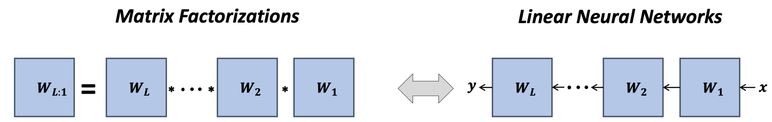
GD over the construction above, generalized to different matrix dimensions:



Theory transfers to practice: $|*| \to \infty$

VI) Implicit Regularization = Rank Minimization?

Matrix factorizations correspond to linear neural networks:

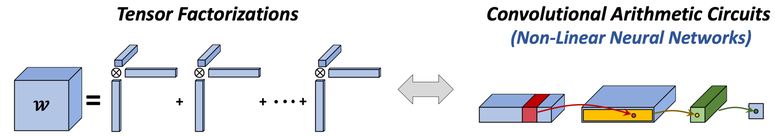


In our construction: implicit regularization provably minimizes matrix rank

Arora et al. 2019: theory and experiments suggest this holds in general

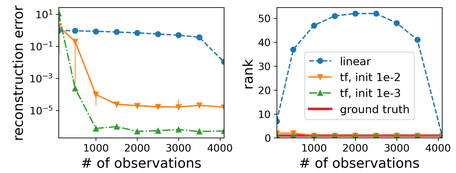
Does rank minimization extend beyond matrix factorization (linear NN)?

Tensor factorizations correspond to non-linear neural networks:



Experiment

GD over tensor factorization for low-rank tensor completion:



linear model $\Rightarrow$ high rank

tensor factorizations (tf) $\Rightarrow$ low rank

In the experiments: implicit regularization minimizes tensor rank

Implicit rank minimization occurs in tensor factorization (non-linear NN)

Hypothesis

Implicit regularization in DL minimizes rank of input-output mapping

Rank minimization may be key to explaining generalization in DL

VII) Takeaway

To understand implicit regularization in deep learning:

- Language of standard norm regularizers might not suffice
- Developing notions of rank for input-output mappings of NNs may be key