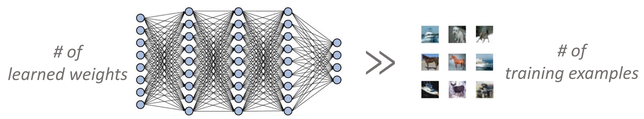
Implicit Regularization in Tensor Factorization

Noam Razin* **Asaf Maman*** **Nadav Cohen**
Tel Aviv University *Equal contribution

International Conference on Machine Learning (ICML) 2021

I) Implicit Regularization in Deep Learning (DL)

Deep neural networks (NNs) are typically overparameterized



With "natural data" predictors found by gradient descent (GD) generalize well

Conventional wisdom: implicit regularization towards low "complexity" predictors

Goal: mathematically understand this implicit regularization

Challenge: lack complexity measures that capture essence of natural data

✔ **low complexity** ✘ **high complexity**

Can we characterize the implicit regularization in concrete settings?

II) Common Testbed: Matrix Factorization (MF)

Matrix Completion: recover unknown matrix given subset of entries

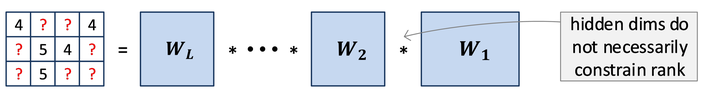
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

$d \times d'$ matrix completion $\longleftrightarrow$ prediction task from $\{1,...,d\} \times \{1,...,d'\}$ to $\mathbb{R}$

In many real-world scenarios matrices of interest have low rank

Matrix Factorization

Parameterize solution as product of matrices and fit observations with GD



MF $\longleftrightarrow$ solving matrix completion via linear NN (w/o explicit regularization!)

Past Work *(e.g. Arora et al. 2019, Razin & Cohen 2020, Li et al. 2021)*

In MF (with small init and step size) implicit regularization minimizes rank

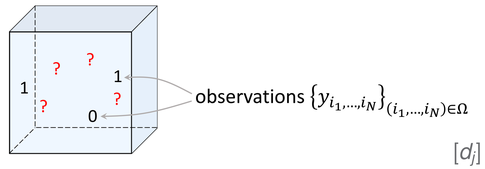
Limitations of Matrix Factorization

(1) Captures prediction over only 2 input variables **(2)** No non-linearity

We study tensor factorization — accounts for both (1) and (2)

III) Beyond Matrix Factorization: Tensor Factorization (TF)

Tensor Completion: recover unknown tensor given subset of entries



$$[d_j] := \{1, \ldots, d_j\}$$

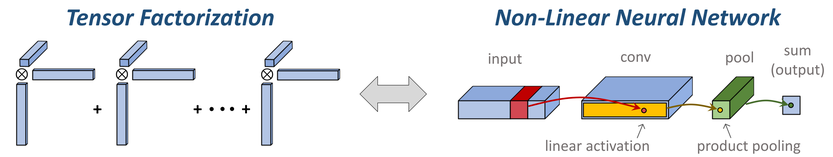
(1) $d_1 \times \cdots \times d_N$ tensor completion $\longleftrightarrow$ prediction task from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

Tensor Factorization

Parameterize solution as sum of outer products and fit observations with GD:

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum_{(i_1,\ldots,i_N)\in\Omega} \ell\left(\left[\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N\right]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N}\right)$$

(2) TF $\longleftrightarrow$ solving tensor completion via NN with multiplicative non-linearity



Tensor rank: min # of components required to express a tensor

Razin & Cohen 2020: GD empirically minimizes tensor rank even when R is large

Question: can this empirical phenomenon be supported theoretically?

IV) Dynamics of Learning: Theoretical Analysis

Theorem (Dynamical Characterization of Component Norms)

GD (with small init and step size) over TF leads component norms to evolve by:

$$\frac{d}{dt}\left\|\otimes_{n=1}^{N} \mathbf{w}_r^n(t)\right\| \propto \left\|\otimes_{n=1}^{N} \mathbf{w}_r^n(t)\right\|^{2-2/N}$$

Interpretation

- $\left\|\otimes_{n=1}^{N} \mathbf{w}_r^n(t)\right\|$ evolves at a rate proportional to its size exponentiated by $2 - 2/N > 1$
- Momentum-like effect: components move slower when small and faster when large
- Small init $\implies$ incremental learning of components $\implies$ low tensor rank

Theorem above leads to:

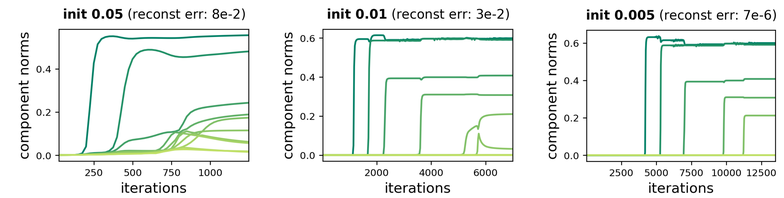
Theorem (Rank 1 Trajectory)

If tensor completion has a rank 1 solution, then under certain technical conditions and a sufficiently small init TF will reach it

V) Dynamics of Learning: Experiments

Experiment (Rank 5 Order 4 Tensor Completion)

Component norms during GD over TF with different init scales:



As init $\to 0$ fewer components depart from zero

Incremental learning of components leads to low tensor rank!

VI) Tensor Rank as Measure of Complexity

Our analysis: tensor rank captures the implicit regularization of a non-linear NN

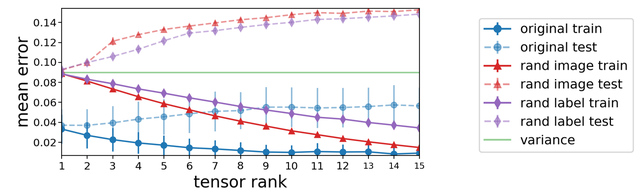
Can tensor rank serve as a measure of complexity for predictors?

Experiment (Fitting Standard Datasets With Predictors of Low Tensor Rank)

Dataset: Fashion-MNIST  (similar results for MNIST)

Compared against two randomized variants:

(i) random images (same labels) (ii) random labels (same images)



Original data fit far more accurately than random (leading to low test err)!

Standard datasets can be fit with predictors of low tensor rank!

VII) Takeaway

Tensor rank may pave way to understanding:

- Implicit regularization of neural networks
- Properties of real-world data translating it to generalization