Implicit Regularization in Tensor Factorization

Noam Razin* **Asaf Maman*** Nadav Cohen

*Equal contribution

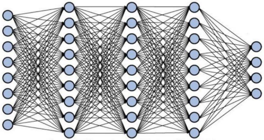
Tel Aviv University

International Conference on Machine Learning (ICML) 2021

Implicit Regularization in Deep Learning

Neural networks generalize with no explicit regularization even when:

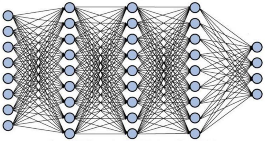
# of learned weights



$\gg$

# of training examples

Neural networks generalize with no explicit regularization even when:

# of learned weights  $\gg$ 

# of training examples

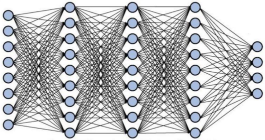
<u>Conventional Wisdom</u>

GD induces implicit regularization towards low “complexity” predictors

Implicit Regularization in Deep Learning

Neural networks generalize with no explicit regularization even when:

# of learned weights



$\gg$

# of training examples

Conventional Wisdom

GD induces implicit regularization towards low "complexity" predictors

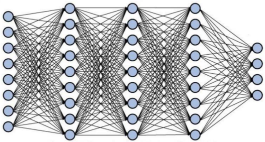
Goal

Mathematically understand this implicit regularization

Implicit Regularization in Deep Learning

Neural networks generalize with no explicit regularization even when:

# of learned weights



$\gg$

# of training examples

Conventional Wisdom

GD induces implicit regularization towards low "complexity" predictors

Goal

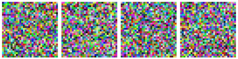
Mathematically understand this implicit regularization

Challenge

Lack complexity measures that capture essence of natural data

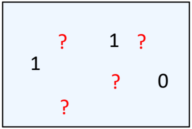
✔ **low complexity**

✘ **high complexity**



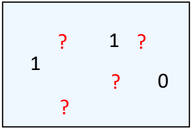
Common Testbed: Matrix Factorization (MF)

Matrix completion: recover unknown matrix given subset of entries



Common Testbed: Matrix Factorization (MF)

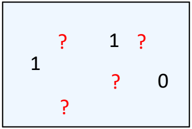
Matrix completion: recover unknown matrix given subset of entries



prediction task over 2 input variables

Common Testbed: Matrix Factorization (MF)

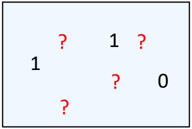
Matrix completion: recover unknown matrix given subset of entries



prediction task over 2 input variables

Natural complexity measure: matrix rank

Matrix completion: recover unknown matrix given subset of entries



$\Leftrightarrow$ prediction task over 2 input variables

Natural complexity measure: matrix rank

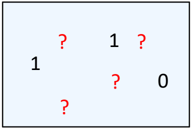
Matrix Factorization

Parameterize solution as product of matrices and fit observations with GD

$$\begin{bmatrix} 4 & ? & ? & 4 \\ ? & 5 & 4 & ? \\ ? & 5 & ? & ? \end{bmatrix} = \boldsymbol{W_L} * \cdots * \boldsymbol{W_2} * \boldsymbol{W_1}$$

Common Testbed: Matrix Factorization (MF)

Matrix completion: recover unknown matrix given subset of entries



⟺ prediction task over 2 input variables

Natural complexity measure: matrix rank

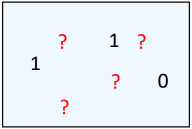
<u>Matrix Factorization</u>

Parameterize solution as product of matrices and fit observations with GD

$$\begin{array}{|c|c|c|c|}\hline 4 & ? & ? & 4 \\ \hline ? & 5 & 4 & ? \\ \hline ? & 5 & ? & ? \\ \hline \end{array} = \boxed{W_L} * \cdots * \boxed{W_2} * \boxed{W_1}$$

MF $\longleftrightarrow$ matrix completion via linear NN

Matrix completion: recover unknown matrix given subset of entries



$\Longleftrightarrow$ prediction task over 2 input variables

Natural complexity measure: matrix rank

Matrix Factorization

Parameterize solution as product of matrices and fit observations with GD

$$\begin{bmatrix} 4 & ? & ? & 4 \\ ? & 5 & 4 & ? \\ ? & 5 & ? & ? \end{bmatrix} = \boxed{W_L} * \cdots * \boxed{W_2} * \boxed{W_1}$$

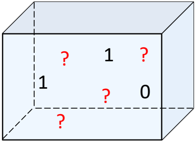
MF $\longleftrightarrow$ matrix completion via linear NN

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

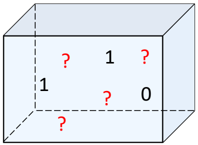
Beyond Matrix Factorization: Tensor Factorization (TF)

Tensor completion: recover unknown tensor given subset of entries



Beyond Matrix Factorization: Tensor Factorization (TF)

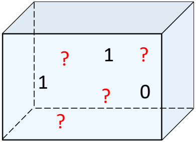
Tensor completion: recover unknown tensor given subset of entries



multi-variable prediction task

Beyond Matrix Factorization: Tensor Factorization (TF)

Tensor completion: recover unknown tensor given subset of entries



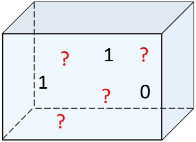
multi-variable prediction task

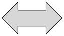
Tensor Factorization

Parameterize solution as sum of outer products and fit observations via GD

$$\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N$$

Tensor completion: recover unknown tensor given subset of entries



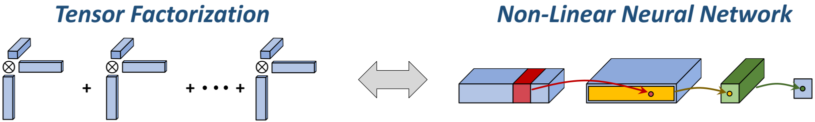


multi-variable prediction task

Tensor Factorization

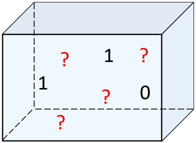
Parameterize solution as sum of outer products and fit observations via GD

$$\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N$$



Beyond Matrix Factorization: Tensor Factorization (TF)

Tensor completion: recover unknown tensor given subset of entries



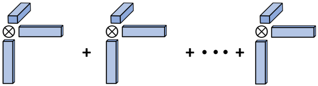
⟺ multi-variable prediction task

Tensor Factorization

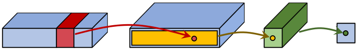
Parameterize solution as sum of outer products and fit observations via GD

$$\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N$$

Tensor Factorization



Non-Linear Neural Network



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

Dynamical Analysis of Implicit Regularization

Dynamical Analysis of Implicit Regularization

Theorem

In training TF (with small init and step size): $\frac{d}{dt}\|\otimes_{n=1}^{N}\mathbf{w}_r^n\| \propto \|\otimes_{n=1}^{N}\mathbf{w}_r^n\|^{2-\frac{2}{N}}$

Dynamical Analysis of Implicit Regularization

Theorem

In training TF (with small init and step size): $\frac{d}{dt}\left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\| \propto \left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Dynamical Analysis of Implicit Regularization

Theorem

In training TF (with small init and step size): $\frac{d}{dt}\left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\| \propto \left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init

Dynamical Analysis of Implicit Regularization

Theorem

In training TF (with small init and step size): $\frac{d}{dt}\|\otimes_{n=1}^{N}\mathbf{w}_r^n\| \propto \|\otimes_{n=1}^{N}\mathbf{w}_r^n\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\implies$ Incremental growth of components

Dynamical Analysis of Implicit Regularization

Theorem

In training TF (with small init and step size): $\frac{d}{dt}\|\otimes_{n=1}^{N}\mathbf{w}_r^n\| \propto \|\otimes_{n=1}^{N}\mathbf{w}_r^n\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\implies$ Incremental growth of components $\implies$ low tensor rank

Dynamical Analysis of Implicit Regularization

Theorem

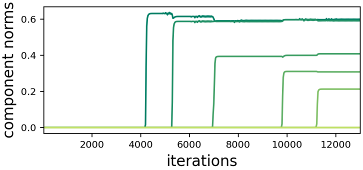
In training TF (with small init and step size): $\frac{d}{dt}\|\otimes_{n=1}^{N}\mathbf{w}_r^n\| \propto \|\otimes_{n=1}^{N}\mathbf{w}_r^n\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\implies$ Incremental growth of components $\implies$ low tensor rank

Experiment

Completion of low rank tensor via TF



Dynamical Analysis of Implicit Regularization

Theorem

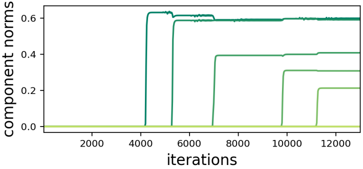
In training TF (with small init and step size): $\frac{d}{dt}\left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\| \propto \left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\implies$ Incremental growth of components $\implies$ low tensor rank

Experiment

Completion of low rank tensor via TF



GD over TF leads to low tensor rank!

Dynamical Analysis of Implicit Regularization

Theorem

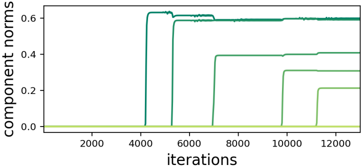
In training TF (with small init and step size): $\frac{d}{dt}\left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\| \propto \left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\implies$ Incremental growth of components $\implies$ low tensor rank

Experiment

Completion of low rank tensor via TF



GD over TF leads to low tensor rank!

Proposition (under technical conditions)

If tensor completion has rank 1 solution, then TF will reach it

Tensor Rank as Measure of Complexity

Recall

Goal: understanding implicit regularization in DL

Tensor Rank as Measure of Complexity

Recall

Goal: understanding implicit regularization in DL

Challenge: lack measures of complexity that capture natural data

Tensor Rank as Measure of Complexity

Recall

Goal: understanding implicit regularization in DL

Challenge: lack measures of complexity that capture natural data

Our Analysis

Tensor rank captures the implicit regularization of a non-linear NN

Tensor Rank as Measure of Complexity

Recall

Goal: understanding implicit regularization in DL

Challenge: lack measures of complexity that capture natural data

Our Analysis

Tensor rank captures the implicit regularization of a non-linear NN

Can tensor rank serve as a measure of complexity?

Tensor Rank as Measure of Complexity

Recall

Goal: understanding implicit regularization in DL

Challenge: lack measures of complexity that capture natural data

Our Analysis

Tensor rank captures the implicit regularization of a non-linear NN

Can tensor rank serve as a measure of complexity?

Experiments

Standard datasets can be fit with predictors of low tensor rank

Tensor Rank as Measure of Complexity

Recall

Goal: understanding implicit regularization in DL

Challenge: lack measures of complexity that capture natural data

Our Analysis

Tensor rank captures the implicit regularization of a non-linear NN

Can tensor rank serve as a measure of complexity?

Experiments

Standard datasets can be fit with predictors of low tensor rank

Tensor rank may shed light on both implicit regularization of NNs and properties of real-world data translating it to generalization