Implicit Regularization in Deep Learning: Lessons Learned from Matrix and Tensor Factorization

Nadav Cohen

Tel Aviv University

AI Week by Tel Aviv University

23 February 2021

Sources

Implicit Regularization in Deep Matrix Factorization

Arora + ***C*** + Hu + Luo (alphabetical order)

NeurIPS 2019

Implicit Regularization in Deep Learning May Not Be Explainable by Norms

Razin + ***C***

NeurIPS 2020

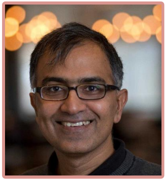
Implicit Regularization in Tensor Factorization

Razin* + Maman* + ***C***

Preprint

*Equal contribution

Collaborators



Sanjeev Arora

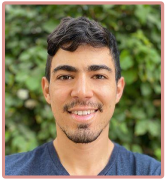
Wei Hu

Yuping Luo

PRINCETON UNIVERSITY

TEL AVIV UNIVERSITY

Noam Razin



Asaf Maman

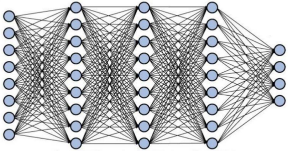
Outline

Generalization in Deep Learning

Generalization in Deep Learning

Deep neural networks (NNs) are typically overparameterized

# of learned weights



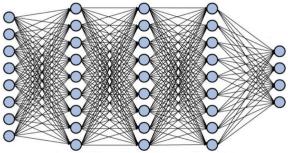
$\gg$

# of training examples

Generalization in Deep Learning

Deep neural networks (NNs) are typically overparameterized

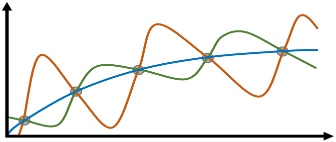
# of learned weights



$\gg$

# of training examples

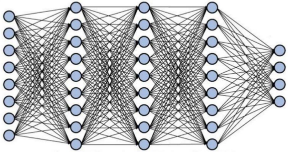
$\Longrightarrow$ many possible solutions (predictors) fit training data



Generalization in Deep Learning

Deep neural networks (NNs) are typically overparameterized

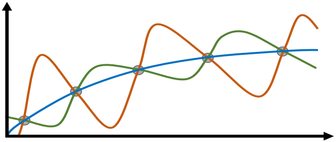
# of learned weights



$\gg$

# of training examples

$\Longrightarrow$ many possible solutions (predictors) fit training data

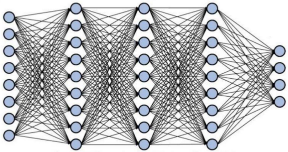


Variants of gradient descent (GD) usually find one of these solutions

Generalization in Deep Learning

Deep neural networks (NNs) are typically overparameterized

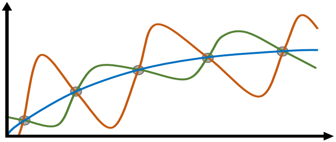
# of learned weights



$\gg$

# of training examples

$\Longrightarrow$ many possible solutions (predictors) fit training data



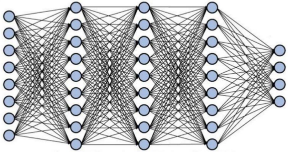
Variants of gradient descent (GD) usually find one of these solutions

With "natural" data solution found often generalizes well

Generalization in Deep Learning

Deep neural networks (NNs) are typically overparameterized

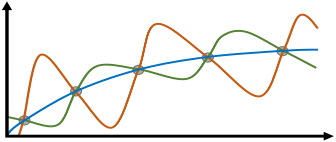
# of learned weights



$\gg$

# of training examples

$\Longrightarrow$ many possible solutions (predictors) fit training data



Variants of gradient descent (GD) usually find one of these solutions

With "natural" data solution found often generalizes well

↑

Even without explicit regularization!

Conventional Wisdom: Implicit Regularization

Conventional Wisdom

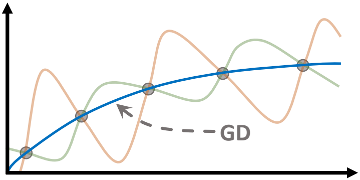
Implicit regularization minimizes "complexity":

Conventional Wisdom: Implicit Regularization

Conventional Wisdom

Implicit regularization minimizes "complexity":

- GD fits training data with predictor of lowest possible complexity

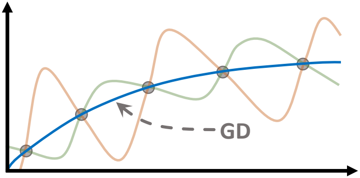


Conventional Wisdom: Implicit Regularization

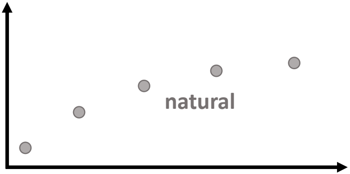
Conventional Wisdom

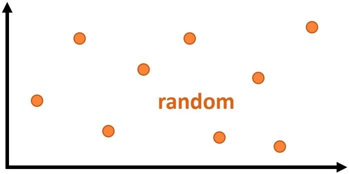
Implicit regularization minimizes "complexity":

- GD fits training data with predictor of lowest possible complexity



- Natural data can be fit with low complexity, other data cannot





Challenge: Formalizing Notion of Complexity

Goal

Mathematically formalize implicit regularization in deep learning (DL)

Challenge: Formalizing Notion of Complexity

Goal

Mathematically formalize implicit regularization in deep learning (DL)

Challenge

We lack definitions for predictor complexity that are:

Challenge: Formalizing Notion of Complexity

Goal

Mathematically formalize implicit regularization in deep learning (DL)

Challenge

We lack definitions for predictor complexity that are:

- quantitative (admit generalization bounds)

$$\text{test error} \leq \text{train error} + \mathcal{O}\Big(\text{complexity} \,/\, (\#\text{ of train examples})\Big)$$

Challenge: Formalizing Notion of Complexity

Goal

Mathematically formalize implicit regularization in deep learning (DL)

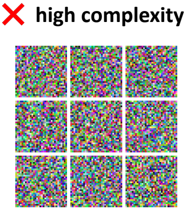
Challenge

We lack definitions for predictor complexity that are:

- quantitative (admit generalization bounds)

$$\text{test error} \leq \text{train error} + \mathcal{O}\Big(\text{complexity} \,/\, (\#\text{ of train examples})\Big)$$

- and capture essence of natural data (allow its fit with low complexity)



Outline

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

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

observations $\{y_{ij}\}_{(i,j)\in\Omega}$

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

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

observations $\{y_{ij}\}_{(i,j)\in\Omega}$

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

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

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

observations $\{y_{ij}\}_{(i,j)\in\Omega}$

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

value of entry (i, j) $\longleftrightarrow$ label of input (i, j)

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

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

observations $\{y_{ij}\}_{(i,j)\in\Omega}$

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

value of entry (i, j) $\longleftrightarrow$ label of input (i, j)

observed entries $\longleftrightarrow$ train data

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

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

observations $\{y_{ij}\}_{(i,j)\in\Omega}$

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

value of entry (i, j) $\longleftrightarrow$ label of input (i, j)

observed entries $\longleftrightarrow$ train data

unobserved entries $\longleftrightarrow$ test data

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

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

observations $\{y_{ij}\}_{(i,j)\in\Omega}$

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

value of entry (i, j) $\longleftrightarrow$ label of input (i, j)

observed entries $\longleftrightarrow$ train data

unobserved entries $\longleftrightarrow$ test data

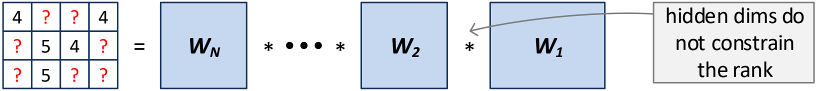
matrix $\longleftrightarrow$ predictor

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix Factorization ⟷ Linear Neural Network

Matrix factorization (MF):

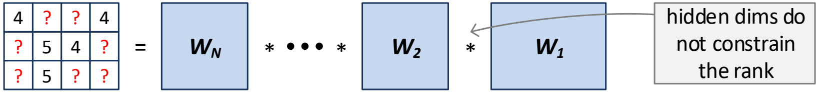
Parameterize solution as product of matrices and fit observations via GD



Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix factorization (MF):

Parameterize solution as product of matrices and fit observations via GD

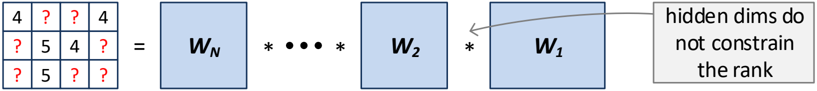


$$\min_{W_1,\ldots,W_N} \sum_{(i,j)\in\Omega} \left([W_N W_{N-1} \cdots W_1]_{ij} - y_{ij} \right)^2$$

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix factorization (MF):

Parameterize solution as product of matrices and fit observations via GD



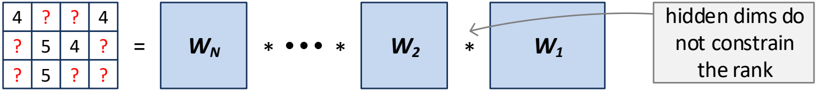
$$\min_{W_1,\ldots,W_N} \sum_{(i,j)\in\Omega} \left([W_N W_{N-1} \cdots W_1]_{ij} - y_{ij} \right)^2$$

MF $\longleftrightarrow$ matrix completion via linear NN (with no explicit regularization!)

Matrix Factorization ⟷ Linear Neural Network

Matrix factorization (MF):

Parameterize solution as product of matrices and fit observations via GD



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

MF ⟷ matrix completion via linear NN (with no explicit regularization!)

Empirical Phenomenon *(Gunasekar et al. 2017)*

MF (with small init and step size) accurately recovers low rank matrices

Implicit Regularization = Norm Minimization?

Implicit Regularization = Norm Minimization?

Classic Result *(Candes & Recht 2008)*

If (i) unknown matrix has low rank; (ii) observations are sufficiently many, then fitting them while minimizing nuclear norm yields accurate recovery

Implicit Regularization = Norm Minimization?

Classic Result *(Candes & Recht 2008)*

If (i) unknown matrix has low rank; (ii) observations are sufficiently many, then fitting them while minimizing nuclear norm yields accurate recovery

Conjecture *(Gunasekar et al. 2017)*

MF of depth 2 (with small init and step size) fits observations while minimizing nuclear norm

Implicit Regularization = Norm Minimization?

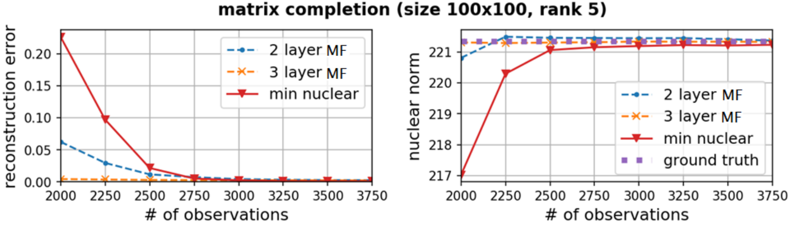
Classic Result *(Candes & Recht 2008)*

If (i) unknown matrix has low rank; (ii) observations are sufficiently many, then fitting them while minimizing nuclear norm yields accurate recovery

Conjecture *(Gunasekar et al. 2017)*

MF of depth 2 (with small init and step size) fits observations while minimizing nuclear norm

Experiment



Implicit Regularization = Norm Minimization?

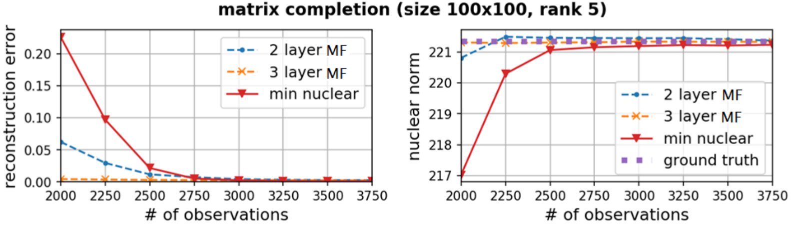
Classic Result *(Candes & Recht 2008)*

If (i) unknown matrix has low rank; (ii) observations are sufficiently many, then fitting them while minimizing nuclear norm yields accurate recovery

Conjecture *(Gunasekar et al. 2017)*

MF of depth 2 (with small init and step size) fits observations while minimizing nuclear norm

Experiment



MF gives up min nuclear norm for low rank (more so with depth)!

Dynamical Analysis of Implicit Regularization

Dynamical Analysis of Implicit Regularization

Denote:

$W_e := W_N \cdots W_1$ — end matrix of MF

Dynamical Analysis of Implicit Regularization

Denote:

$W_e := W_N \cdots W_1$ — end matrix of MF $\{\sigma_r\}_r$ — singular vals of W_e

Dynamical Analysis of Implicit Regularization

Denote:

$W_e := W_N \cdots W_1$ — end matrix of MF $\quad\quad \{\sigma_r\}_r$ — singular vals of W_e

Theorem

In training MF of depth N (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/N}$

Dynamical Analysis of Implicit Regularization

Denote:

$W_e := W_N \cdots W_1$ — end matrix of MF $\{\sigma_r\}_r$ — singular vals of W_e

Theorem

In training MF of depth N (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/N}$

Depth speeds up (slows down) large (small) singular vals!

Dynamical Analysis of Implicit Regularization

Denote:

$W_e := W_N \cdots W_1$ — end matrix of MF $\{\sigma_r\}_r$ — singular vals of W_e

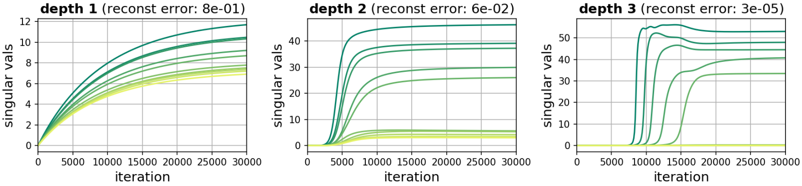
Theorem

In training MF of depth N (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/N}$

Depth speeds up (slows down) large (small) singular vals!

Experiment

Completion of low rank matrix via MF



Dynamical Analysis of Implicit Regularization

Denote:

$W_e := W_N \cdots W_1$ — end matrix of MF $\quad \{\sigma_r\}_r$ — singular vals of W_e

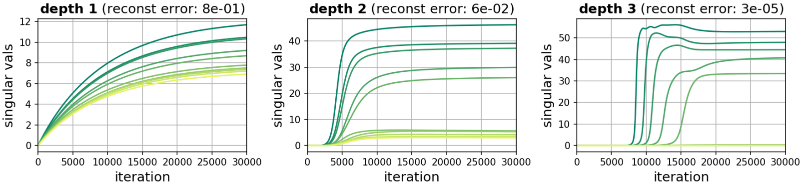
Theorem

In training MF of depth N (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/N}$

Depth speeds up (slows down) large (small) singular vals!

Experiment

Completion of low rank matrix via MF



MF depth leads to larger gaps between singular vals (lower rank)!

Dynamical Analysis of Implicit Regularization (2)

Practical Application

Implicit Rank-Minimizing Autoencoder

Li Jing
Facebook AI Research
New York

Jure Zbontar
Facebook AI Research
New York

Yann LeCun
Facebook AI Research
New York

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

> *"rank ... is implicitly minimized by relying on the fact that gradient descent ... in multi-layer linear networks leads to minimum-rank ..."*

Implicit Regularization $\neq$ Norm Minimization

Implicit Regularization $\neq$ Norm Minimization

Theorem

In training MF of depth N (with small init and step size): $\frac{d}{dt}\sigma_r \propto \sigma_r^{2-2/N}$

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

quantity	***minimizer***

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ does not change sign

Consider the matrix completion problem:

$$\begin{pmatrix} {\color{red}?} & 1 \\ 1 & 0 \end{pmatrix}$$

quantity	***minimizer***
Schatten-p norm	

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ does not change sign

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	? = 0

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ does not change sign

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	? = 0
arbitrary norm	

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	? = 0
arbitrary norm	|?| < ∞

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	? = 0
arbitrary norm	|?| < ∞
rank	

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ does not change sign

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:

- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	$? = 0$
arbitrary norm	$\lvert ? \rvert < \infty$
rank	$\lvert ? \rvert \to \infty$

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ does not change sign

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	$? \ = \ 0$
arbitrary norm	$\lvert ? \rvert < \infty$
rank	$\lvert ? \rvert \to \infty$

Norm minimization contradicts rank minimization!

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	? = 0
arbitrary norm	|?| < ∞
rank	|?| → ∞

Norm minimization contradicts rank minimization!

By corollary, if $\det(W_e) > 0$ at init: fitting observations $\implies |?| \to \infty$

Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ *does not change sign*

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

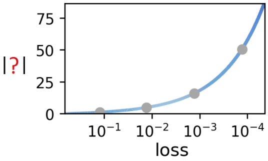
Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	$? = 0$
arbitrary norm	$\lvert ? \rvert < \infty$
rank	$\lvert ? \rvert \to \infty$

Norm minimization contradicts rank minimization!

By corollary, if $\det(W_e) > 0$ at init: fitting observations $\implies$ $\lvert ? \rvert \to \infty$

Experiment



Implicit Regularization $\neq$ Norm Minimization

Corollary

In training MF of depth $N \geq 2$, $\det(W_e)$ does not change sign

Consider the matrix completion problem:

$$\begin{pmatrix} ? & 1 \\ 1 & 0 \end{pmatrix}$$

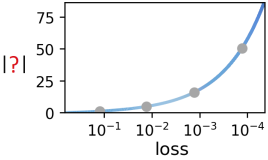
Special cases:
- nuclear norm
- Frobenius norm
- spectral norm

quantity	***minimizer***
Schatten-p norm	$? = 0$
arbitrary norm	$\lvert ? \rvert < \infty$
rank	$\lvert ? \rvert \to \infty$

Norm minimization contradicts rank minimization!

By corollary, if $\det(W_e) > 0$ at init: fitting observations $\implies \lvert ? \rvert \to \infty$

Experiment



There are settings where implicit regularization of MF drives all norms to ∞ while minimizing rank!

Outline

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

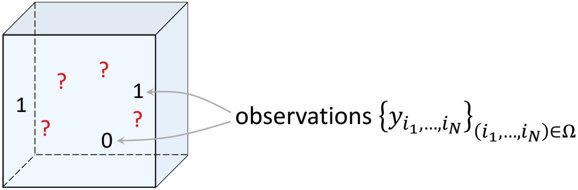
Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: multi-dim array

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: multi-dim array

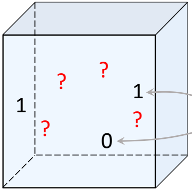
Tensor completion: recover unknown tensor given subset of entries



Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: multi-dim array

Tensor completion: recover unknown tensor given subset of entries



observations $\{y_{i_1,\ldots,i_N}\}_{(i_1,\ldots,i_N)\in\Omega}$

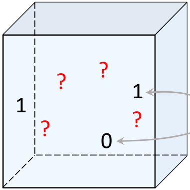
$[d_j] := \{1, \ldots, d_j\}$

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

Tensor Completion ⟷ Multi-Dimensional Prediction

Tensor: multi-dim array

Tensor completion: recover unknown tensor given subset of entries



observations $\{y_{i_1,\ldots,i_N}\}_{(i_1,\ldots,i_N)\in\Omega}$

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

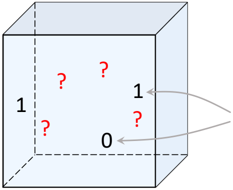
$d_1 \times \cdots \times d_N$ tensor completion ⟷ prediction from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

value of entry $(i_1, \ldots, i_N)$ ⟷ label of input $(i_1, \ldots, i_N)$

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: multi-dim array

Tensor completion: recover unknown tensor given subset of entries



observations $\{y_{i_1,\ldots,i_N}\}_{(i_1,\ldots,i_N)\in\Omega}$

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

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

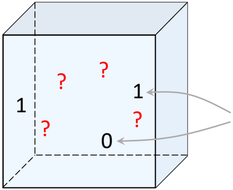
value of entry $(i_1,\ldots,i_N)$ $\longleftrightarrow$ label of input $(i_1,\ldots,i_N)$

observed entries $\longleftrightarrow$ train data

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: multi-dim array

Tensor completion: recover unknown tensor given subset of entries



observations $\{y_{i_1,\ldots,i_N}\}_{(i_1,\ldots,i_N)\in\Omega}$

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

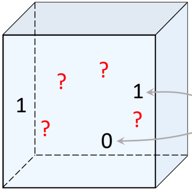
$d_1 \times \cdots \times d_N$ tensor completion $\longleftrightarrow$ prediction from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

value of entry $(i_1, \ldots, i_N)$	$\longleftrightarrow$	label of input $(i_1, \ldots, i_N)$
observed entries	$\longleftrightarrow$	train data
unobserved entries	$\longleftrightarrow$	test data

Tensor Completion ⟷ Multi-Dimensional Prediction

Tensor: multi-dim array

Tensor completion: recover unknown tensor given subset of entries



observations $\{y_{i_1,\ldots,i_N}\}_{(i_1,\ldots,i_N)\in\Omega}$

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

$d_1 \times \cdots \times d_N$ tensor completion ⟷ prediction from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

value of entry $(i_1, \ldots, i_N)$	⟷	label of input $(i_1, \ldots, i_N)$
observed entries	⟷	train data
unobserved entries	⟷	test data
tensor	⟷	predictor

Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor factorization (TF):

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

Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor factorization (TF):

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

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum\nolimits_{(i_1,\ldots,i_N) \in \Omega} \left(\left[\sum\nolimits_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \right]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \right)^2$$

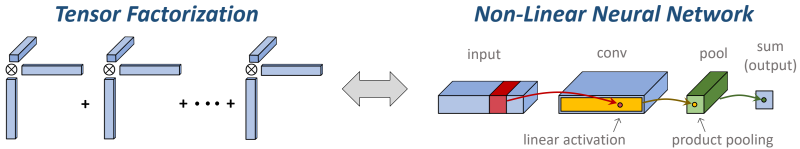
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor factorization (TF):

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

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum\nolimits_{(i_1,\ldots,i_N)\in\Omega} \left(\left[\sum\nolimits_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \right]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \right)^2$$

TF $\longleftrightarrow$ tensor completion via NN with multiplicative non-linearity



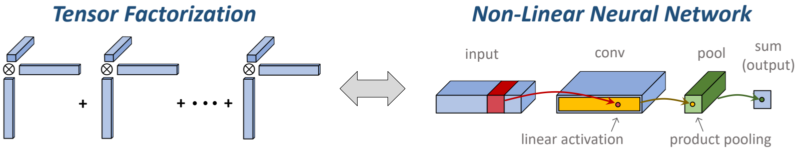
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor factorization (TF):

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

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum_{(i_1,\ldots,i_N)\in\Omega} \left(\left[\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \right]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \right)^2$$

TF $\longleftrightarrow$ tensor completion via NN with multiplicative non-linearity



Experiment

TF (with small init and step size) accurately recovers low rank tensors

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): $\frac{d}{dt}\|\otimes_{n=1}^{N}\mathbf{w}_r^n\| \propto \|\otimes_{n=1}^{N}\mathbf{w}_r^n\|^{2-\frac{2}{N}}$

Component norms accelerate (decelerate) when large (small)!

Dynamical Analysis of Implicit Regularization

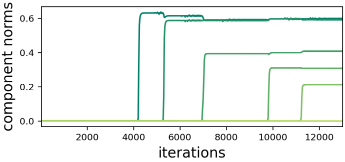
Theorem

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

Component norms accelerate (decelerate) when large (small)!

Experiment

Completion of low rank tensor via TF



Dynamical Analysis of Implicit Regularization

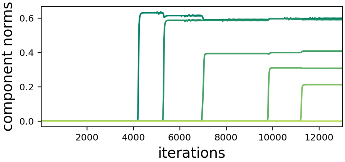
Theorem

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

Component norms accelerate (decelerate) when large (small)!

Experiment

Completion of low rank tensor via TF



Training TF leads to gaps between component norms (low tensor rank)!

Dynamical Analysis of Implicit Regularization

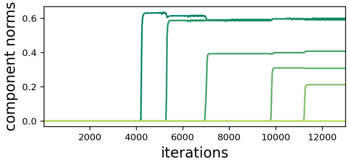
Theorem

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

Component norms accelerate (decelerate) when large (small)!

Experiment

Completion of low rank tensor via TF



Training TF leads to gaps between component norms (low tensor rank)!

Theorem leads to:

Proposition

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

Outline

Challenge: Formalizing Notion of Complexity

Goal

Mathematically formalize implicit regularization in deep learning (DL)

Challenge

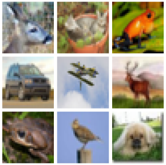
We lack definitions for predictor complexity that are:

- quantitative (admit generalization bounds)

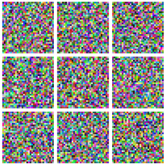
$$\text{test error} \leq \text{train error} + \mathcal{O}\Big(\text{complexity} \,/\, (\text{\# of train examples})\Big)$$

- and capture essence of natural data (allow its fit with low complexity)

✓ **low complexity**



✗ **high complexity**



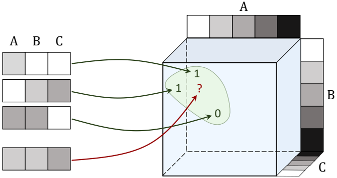
Tensor Rank Captures Non-Linear Neural Network

We saw:

Tensor Rank Captures Non-Linear Neural Network

We saw:

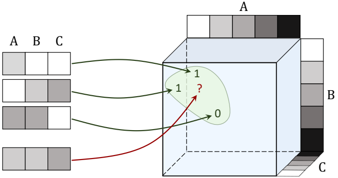
- Tensor completion $\longleftrightarrow$ multi-dim prediction



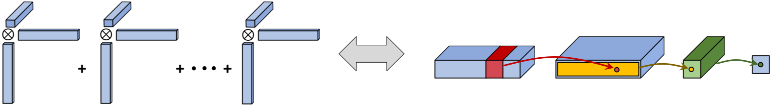
Tensor Rank Captures Non-Linear Neural Network

We saw:

- Tensor completion $\longleftrightarrow$ multi-dim prediction



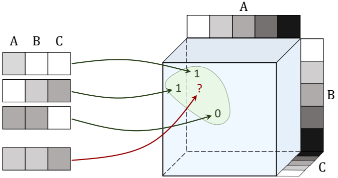
- Tensor factorization $\longleftrightarrow$ non-linear NN



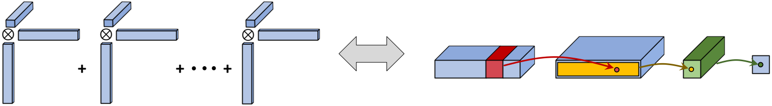
Tensor Rank Captures Non-Linear Neural Network

We saw:

- Tensor completion $\longleftrightarrow$ multi-dim prediction



- Tensor factorization $\longleftrightarrow$ non-linear NN

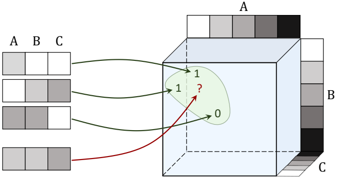


- Implicit regularization favors tensors (predictors) of low rank

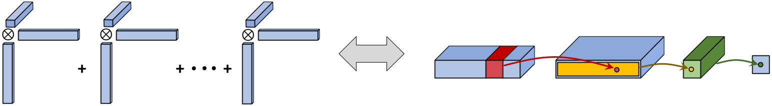
Tensor Rank Captures Non-Linear Neural Network

We saw:

- Tensor completion $\longleftrightarrow$ multi-dim prediction



- Tensor factorization $\longleftrightarrow$ non-linear NN



- Implicit regularization favors tensors (predictors) of low rank

Question

Can tensor rank serve as measure of complexity for predictors?

Experiment: Fitting Data with Low Tensor Rank

Experiment: Fitting Data with Low Tensor Rank

Experiment

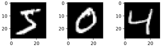
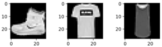
Fitting data with predictors of low tensor rank

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting data with predictors of low tensor rank

Datasets:

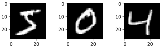
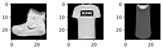
- MNIST  and Fashion-MNIST  (one-vs-all)

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting data with predictors of low tensor rank

Datasets:

- MNIST  and Fashion-MNIST  (one-vs-all)
- Each compared against:

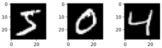
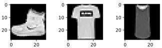
  (i) random images (same labels)   (ii) random labels (same images)

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting data with predictors of low tensor rank

Datasets:

- MNIST  and Fashion-MNIST  (one-vs-all)
- Each compared against:

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

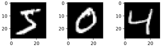
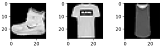
Original data fit far more accurately than random (leading to low test err)!

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting data with predictors of low tensor rank

Datasets:

- MNIST  and Fashion-MNIST  (one-vs-all)
- Each compared against:

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

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

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

Outline

Recap

Recap

Understanding implicit regularization in DL:

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN
- Conjecture: implicit regularization minimizes norm

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN
- Conjecture: implicit regularization minimizes norm
- Dynamical analysis: implicit regularization minimizes rank (not norm)

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN
- Conjecture: implicit regularization minimizes norm
- Dynamical analysis: implicit regularization minimizes rank (not norm)

Tensor factorization:

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN
- Conjecture: implicit regularization minimizes norm
- Dynamical analysis: implicit regularization minimizes rank (not norm)

Tensor factorization:

- Equivalent to multi-dim prediction via non-linear NN

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN
- Conjecture: implicit regularization minimizes norm
- Dynamical analysis: implicit regularization minimizes rank (not norm)

Tensor factorization:

- Equivalent to multi-dim prediction via non-linear NN
- Dynamical analysis: implicit regularization minimizes tensor rank

Recap

Understanding implicit regularization in DL:

- Challenge: lack measures of complexity that capture natural data

Matrix factorization:

- Equivalent to two-dim prediction via linear NN
- Conjecture: implicit regularization minimizes norm
- Dynamical analysis: implicit regularization minimizes rank (not norm)

Tensor factorization:

- Equivalent to multi-dim prediction via non-linear NN
- Dynamical analysis: implicit regularization minimizes tensor rank

Tensor rank as measure of complexity may capture natural data!

Outline

Thank You

Work supported by: Amnon and Anat Shashua, Len Blavatnik and the Blavatnik Family Foundation, Yandex Initiative in Machine Learning, Google Research Gift