Implicit Regularization in Tensor Factorization

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

*Equal contribution

Tel Aviv University

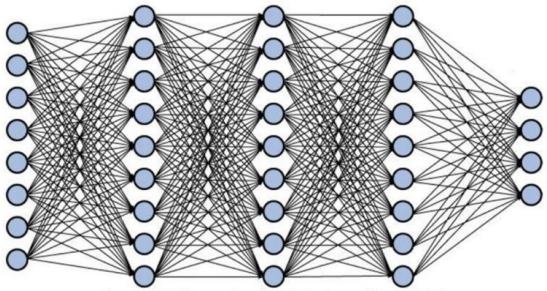
ICMLi 2021

MDLI
Machine & Deep Learning Israel

Implicit Regularization in Deep Learning

Neural networks generalize with no explicit regularization even when:

# of learned weights



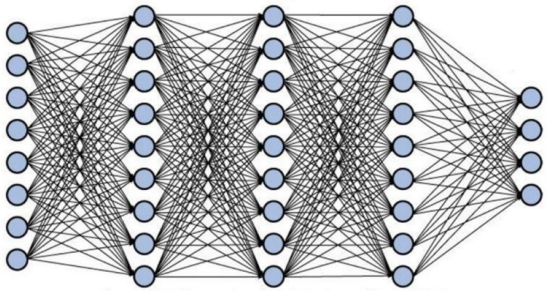
$\gg$

# of training examples

Implicit Regularization in Deep Learning

Neural networks generalize with no explicit regularization even when:

# of learned weights



$\gg$

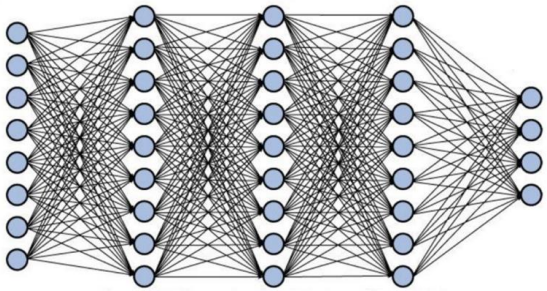
# of training examples

Conventional Wisdom

Implicit Regularization in Deep Learning

Neural networks generalize with no explicit regularization even when:

# of learned weights

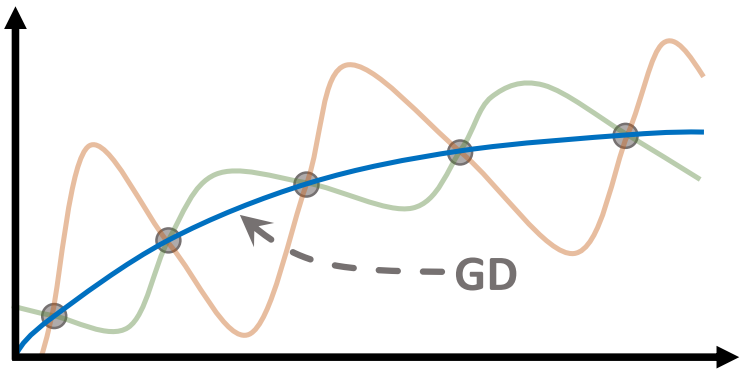


$\gg$

# of training examples

Conventional Wisdom

GD induces implicit regularization towards low "complexity" predictors



Implicit Regularization in Deep Learning

Goal

Mathematically understand this implicit regularization

Implicit Regularization in Deep Learning

Goal

Mathematically understand this implicit regularization

Why?

- Explain why some models generalize while others don't

Implicit Regularization in Deep Learning

Goal

Mathematically understand this implicit regularization

Why?

- Explain why some models generalize while others don't
- Select a model with implicit bias that captures data properties

Implicit Regularization in Deep Learning

Goal

Mathematically understand this implicit regularization

Why?

- Explain why some models generalize while others don't
- Select a model with implicit bias that captures data properties
- Design models with new types of implicit regularization[1]

[1] e.g. Jing, Zbontar and LeCun 2020

Implicit Regularization in Deep Learning

Goal

Mathematically understand this implicit regularization

Why?

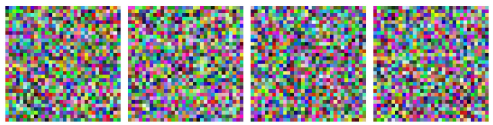
- Explain why some models generalize while others don't
- Select a model with implicit bias that captures data properties
- Design models with new types of implicit regularization[1]

Challenge

Lack complexity measures that capture essence of natural data

✔ **low complexity**

✘ **high complexity**



[1] e.g. Jing, Zbontar and LeCun 2020

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 $\{1, ..., d\} \times \{1, ..., d'\}$ 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 $\{1, ..., d\} \times \{1, ..., d'\}$ 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 $\{1,...,d\} \times \{1,...,d'\}$ 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 $\{1,...,d\} \times \{1,...,d'\}$ 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 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
unobserved entries	$\longleftrightarrow$	test data
matrix	$\longleftrightarrow$	predictor

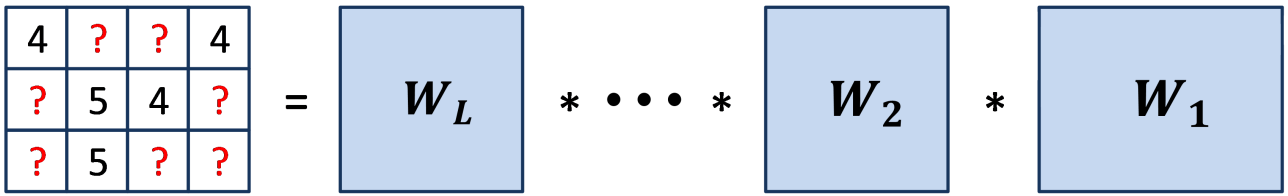
Natural complexity measure: matrix rank

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix factorization (**MF**):

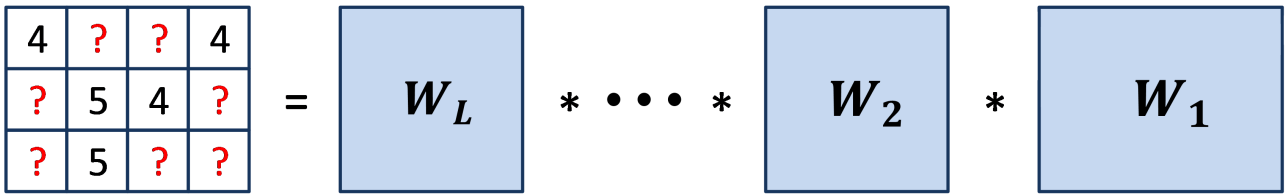
Parameterize solution as product of matrices and fit observations via GD



Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix factorization (**MF**):

Parameterize solution as product of matrices and fit observations via GD

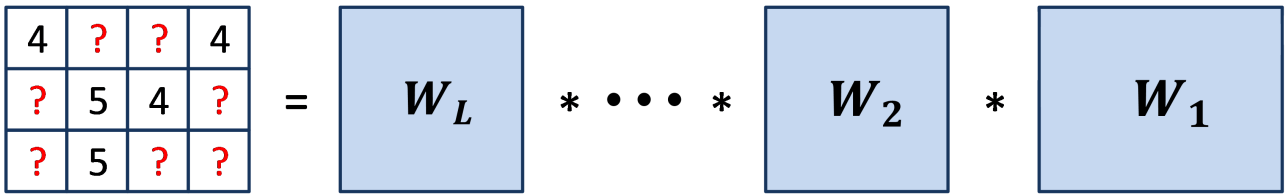


MF $\longleftrightarrow$ matrix completion via linear NN

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix factorization (MF):

Parameterize solution as product of matrices and fit observations via GD



MF $\longleftrightarrow$ matrix completion via linear NN

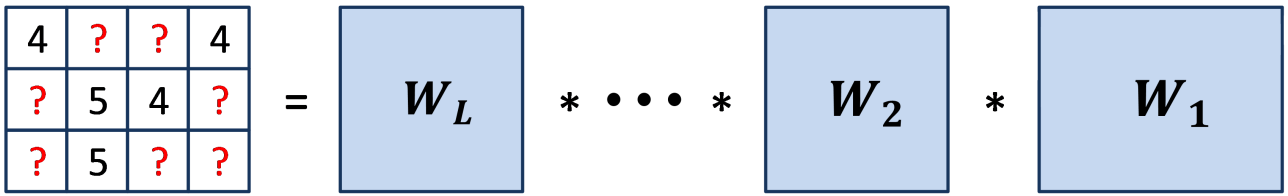
Past Work *(e.g. Arora et al. 2019, Razin & Cohen 2020, Li et al. 2021)*

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

Matrix Factorization $\longleftrightarrow$ Linear Neural Network

Matrix factorization (MF):

Parameterize solution as product of matrices and fit observations via GD



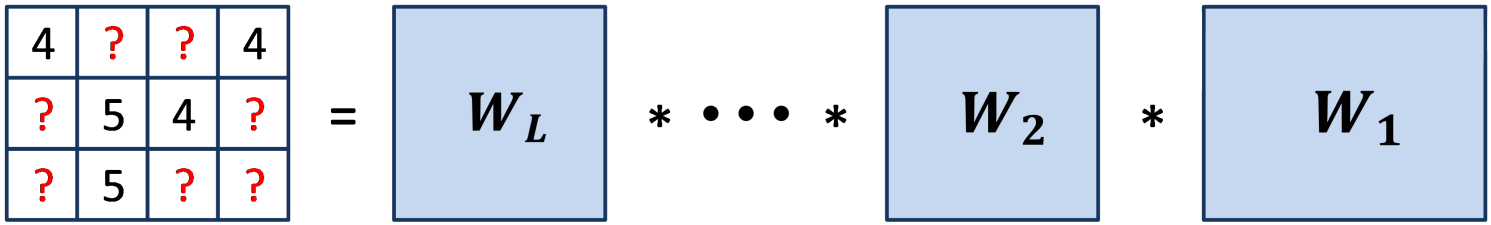
MF $\longleftrightarrow$ matrix completion via linear NN

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

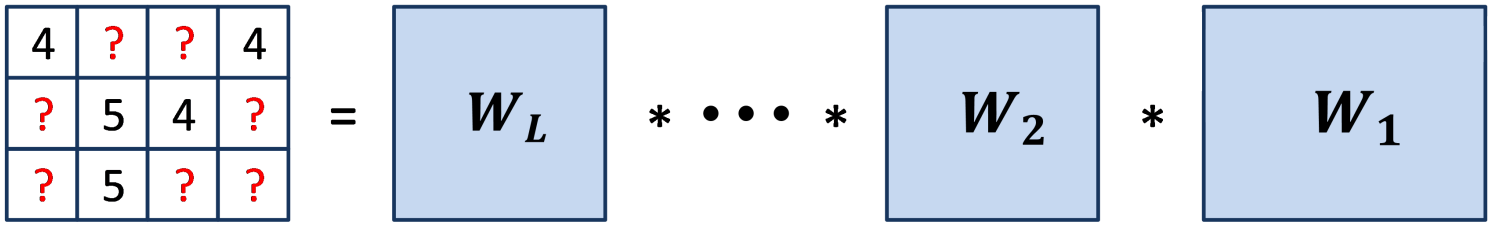
Implicit regularization to low rank + data is low rank
$\Longrightarrow$ generalization

Drawbacks of Studying Matrix Factorization (MF)



As a surrogate for deep learning, MF is inherently limited:

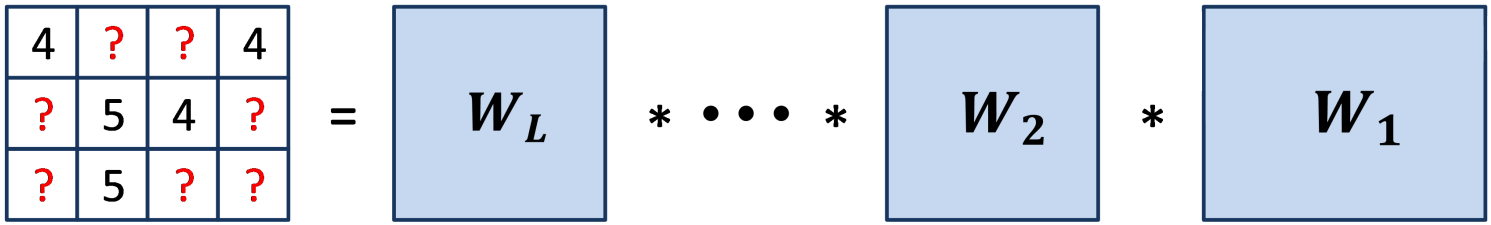
Drawbacks of Studying Matrix Factorization (MF)



As a surrogate for deep learning, MF is inherently limited:

(1) Misses crucial aspect of non-linearity

Drawbacks of Studying Matrix Factorization (MF)

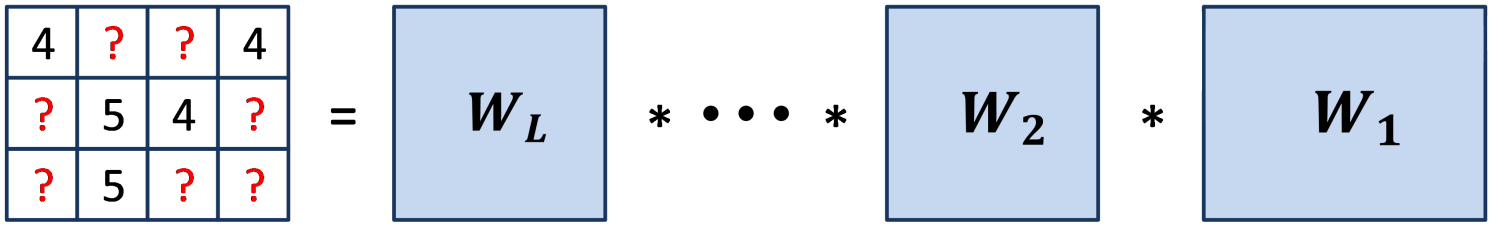


As a surrogate for deep learning, MF is inherently limited:

(1) Misses crucial aspect of non-linearity

(2) Does not capture prediction with more than 2 input variables

Drawbacks of Studying Matrix Factorization (MF)



As a surrogate for deep learning, MF is inherently limited:

(1) Misses crucial aspect of non-linearity

(2) Does not capture prediction with more than 2 input variables

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

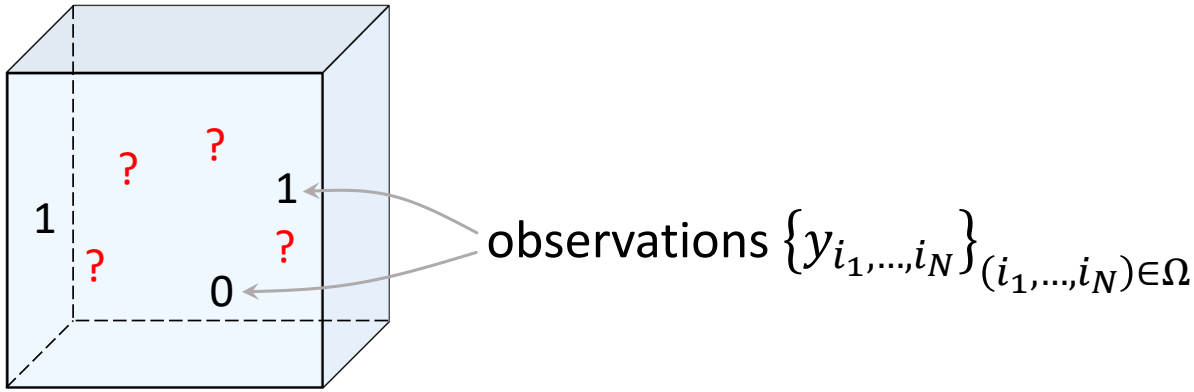
Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array (N = order of tensor)

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array (N = order of tensor)

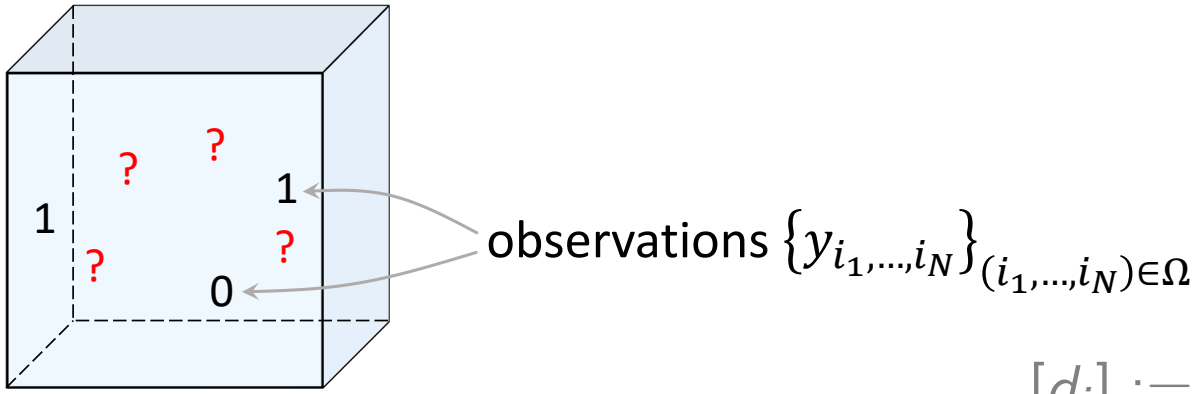
Tensor completion: recover unknown tensor given subset of entries



Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor completion: recover unknown tensor given subset of entries



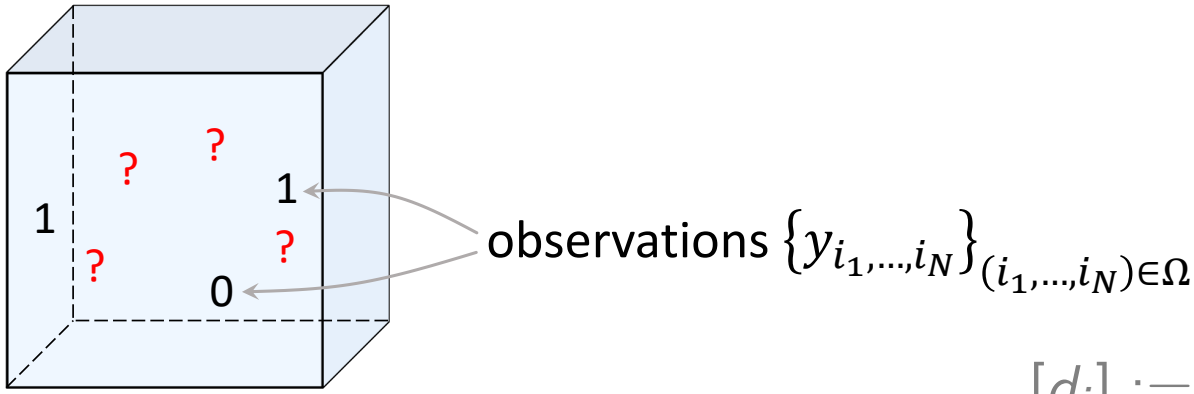
$$[d_j] := \{1, \ldots, d_j\}$$

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

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

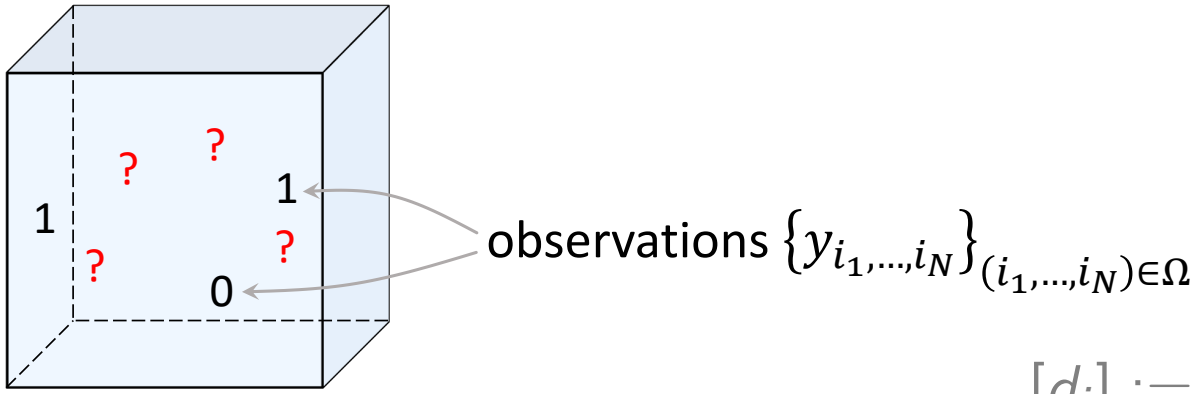
$d_1 \times \cdots \times d_N$ tensor completion $\longleftrightarrow$ prediction from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

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

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

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

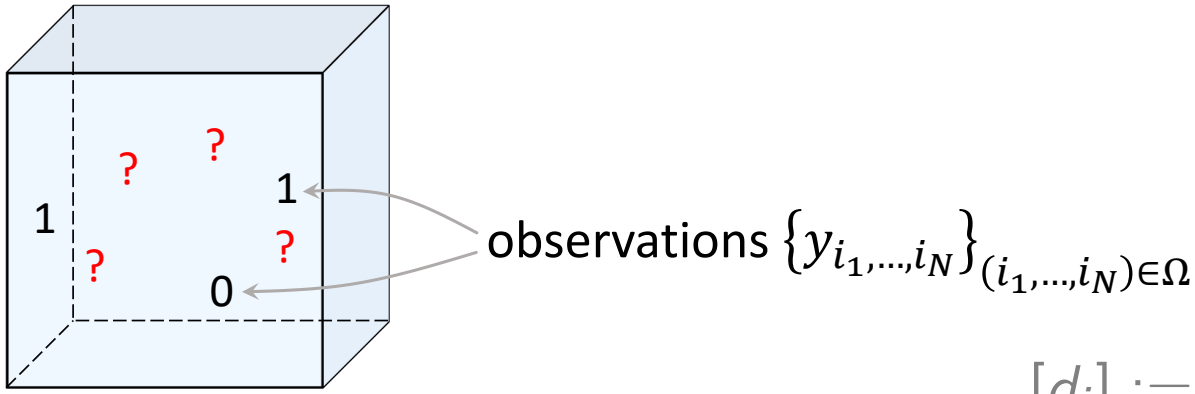
value of entry $(i_1, \ldots, i_N)$ $\longleftrightarrow$ label of input $(i_1, \ldots, i_N)$

observed entries $\longleftrightarrow$ train data

Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

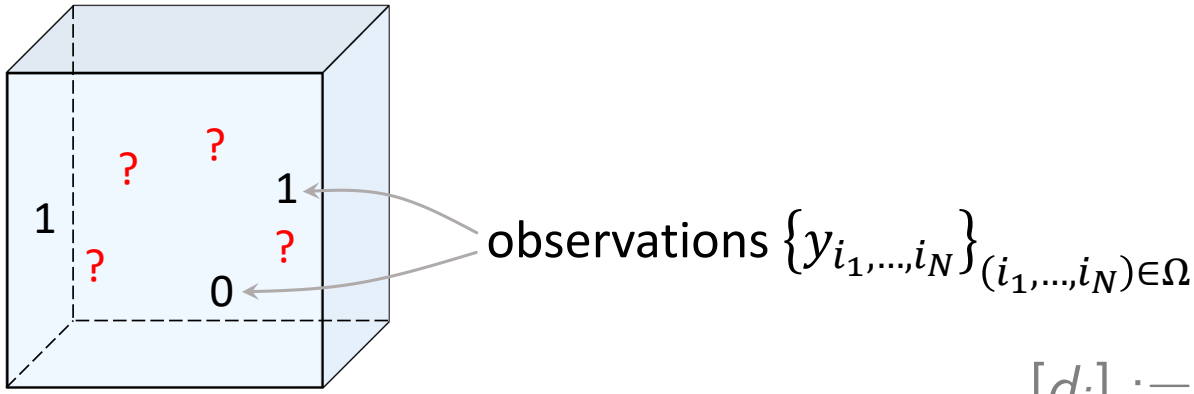
$d_1 \times \cdots \times d_N$ tensor completion $\longleftrightarrow$ prediction from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

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 $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

$d_1 \times \cdots \times d_N$ 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	$\longleftrightarrow$	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

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

Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

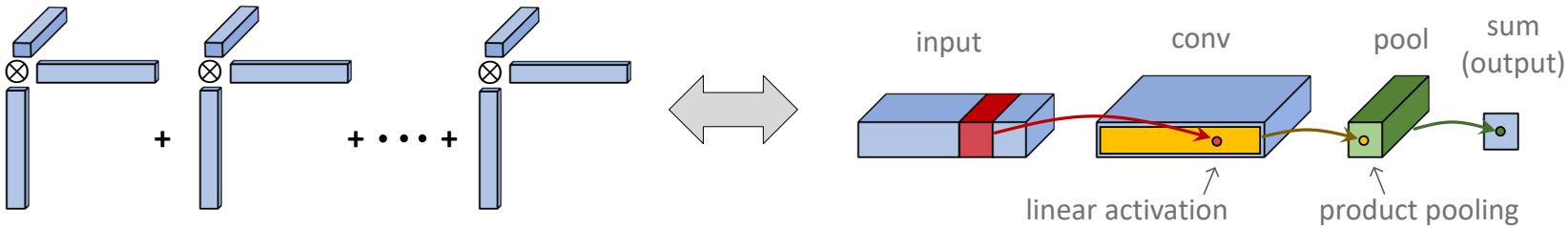
Tensor factorization (**TF**):

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

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

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

Tensor Factorization ***Non-Linear Neural Network***



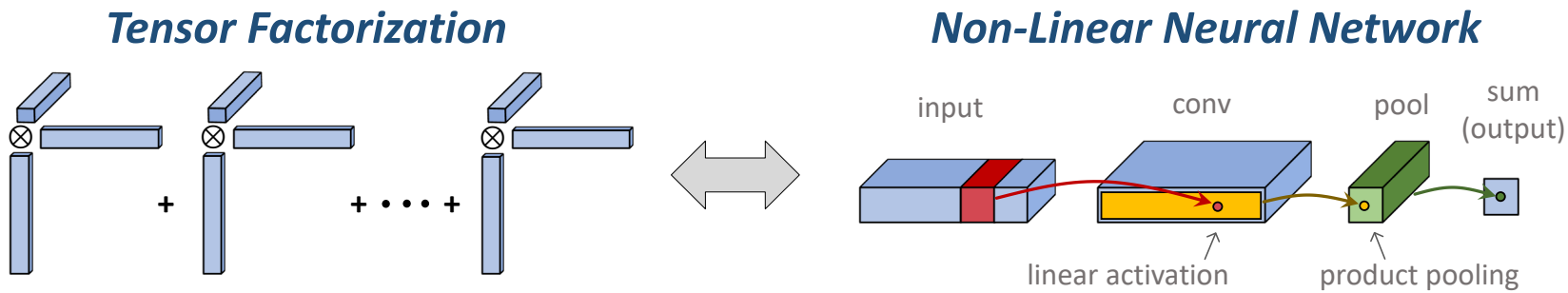
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor factorization (**TF**):

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

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

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



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

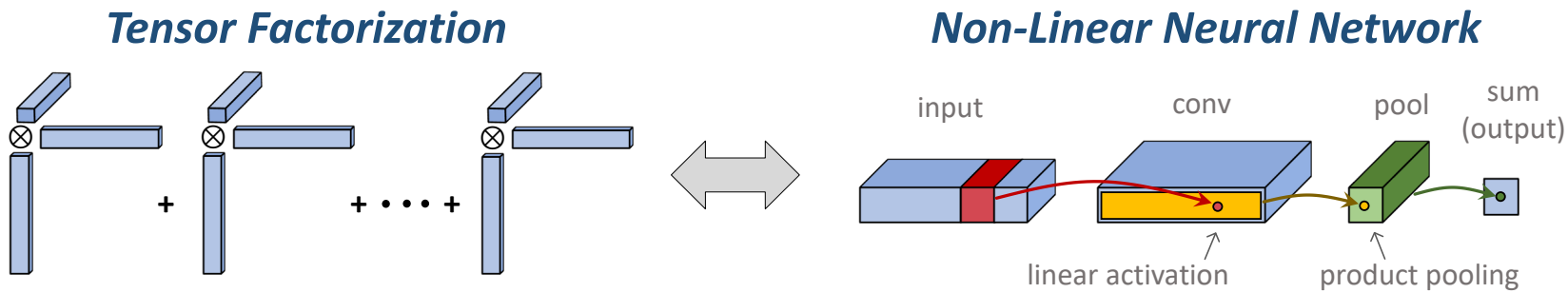
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Tensor factorization (TF):

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

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

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



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

Empirical Phenomenon *(Razin & Cohen 2020)*

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

Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

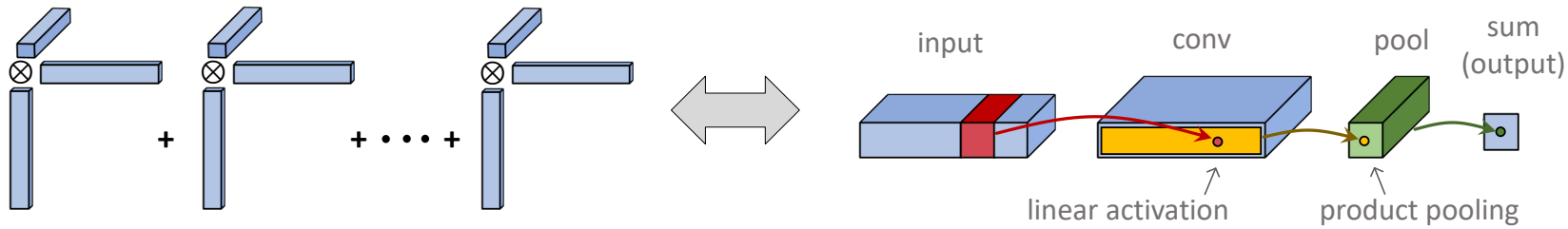
Tensor factorization (**TF**):

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

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

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

Tensor Factorization ***Non-Linear Neural Network***



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

Empirical Phenomenon *(Razin & Cohen 2020)*

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

We provide the first theoretical analysis for that phenomenon

Dynamical Analysis of Implicit Regularization

Dynamical Analysis of Implicit Regularization

Theorem

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

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

Components move slower when small and faster when large!

Small init $\implies$ incremental growth of components

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 $\Longrightarrow$ incremental growth of components $\Longrightarrow$ low tensor rank

Dynamical Analysis of Implicit Regularization

Theorem

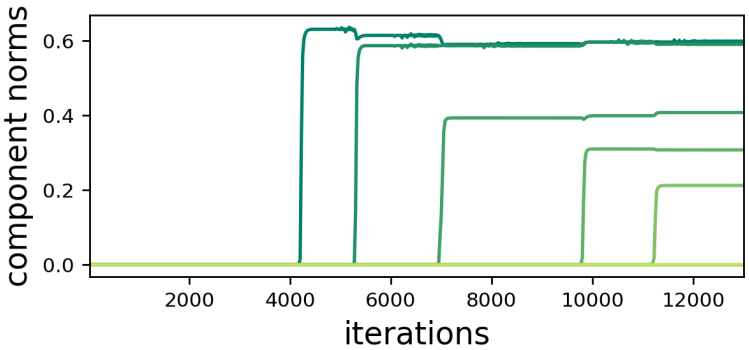
In training TF (with small init and step size): $\frac{d}{dt}\left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\| \propto \left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\Longrightarrow$ incremental growth of components $\Longrightarrow$ low tensor rank

Experiment

Completion of low rank tensor via TF



Dynamical Analysis of Implicit Regularization

Theorem

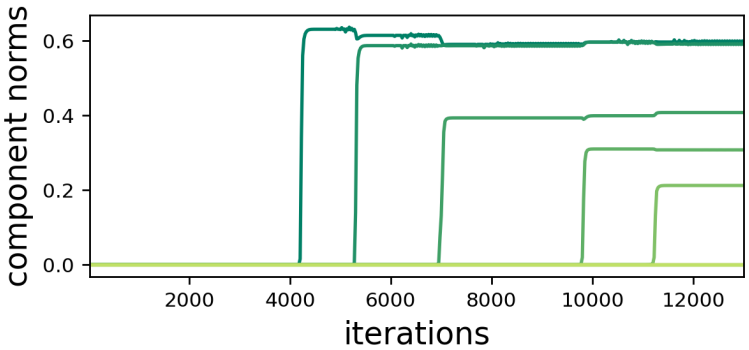
In training TF (with small init and step size): $\frac{d}{dt}\left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\| \propto \left\|\otimes_{n=1}^{N}\mathbf{w}_r^n\right\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\Longrightarrow$ incremental growth of components $\Longrightarrow$ 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

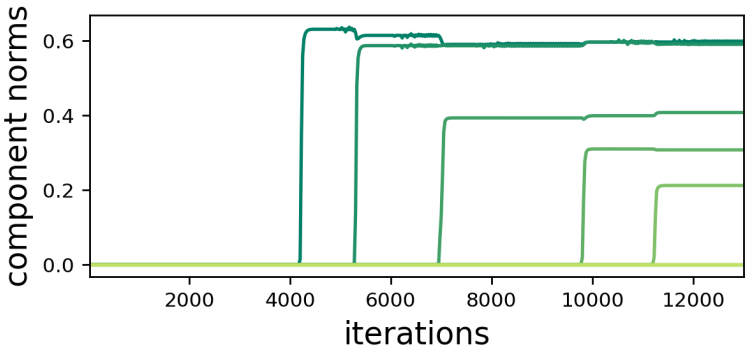
In training TF (with small init and step size): $\frac{d}{dt}\|\otimes_{n=1}^{N}\mathbf{w}_r^n\| \propto \|\otimes_{n=1}^{N}\mathbf{w}_r^n\|^{2-\frac{2}{N}}$

Components move slower when small and faster when large!

Small init $\Longrightarrow$ incremental growth of components $\Longrightarrow$ low tensor rank

Experiment

Completion of low rank tensor via TF



GD over TF leads to low tensor rank!

Proposition

If tensor completion has a rank 1 solution, then under certain technical conditions, tensor factorization will reach it.

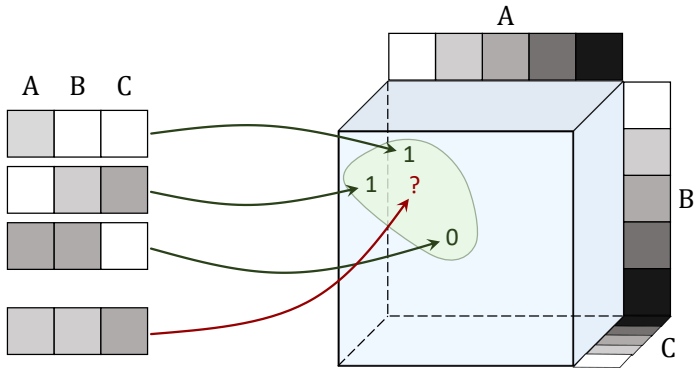
Tensor Rank as a Measure of Complexity

We saw:

Tensor Rank as a Measure of Complexity

We saw:

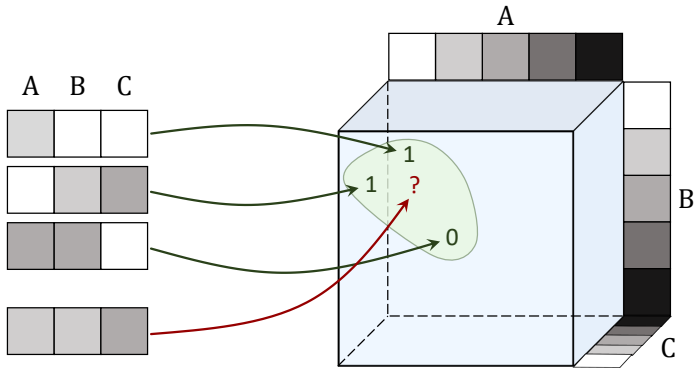
- Tensor completion $\longleftrightarrow$ multi-dimensional prediction



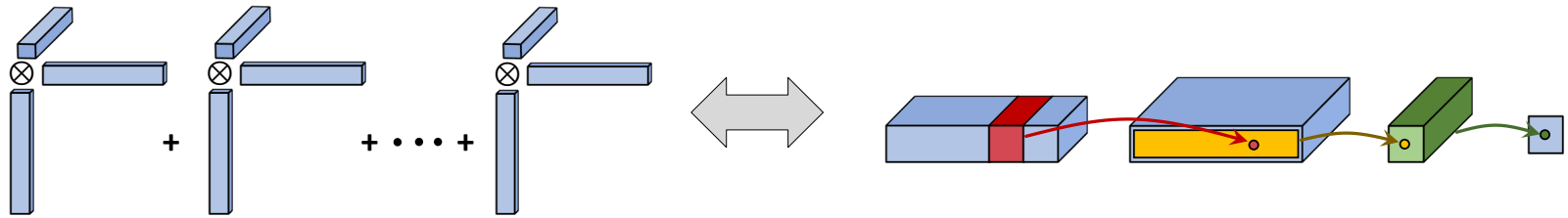
Tensor Rank as a Measure of Complexity

We saw:

- Tensor completion $\longleftrightarrow$ multi-dimensional prediction



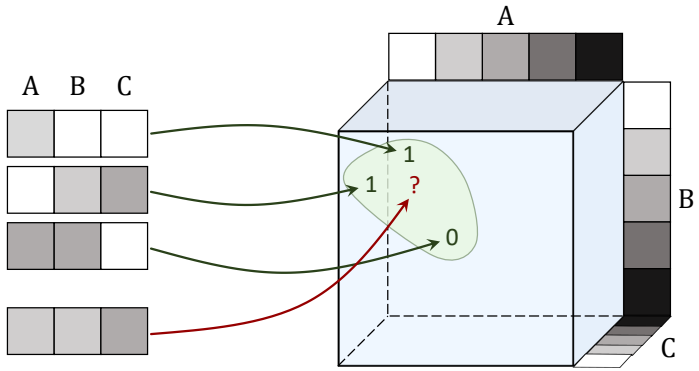
- Tensor factorization $\longleftrightarrow$ non-linear NN



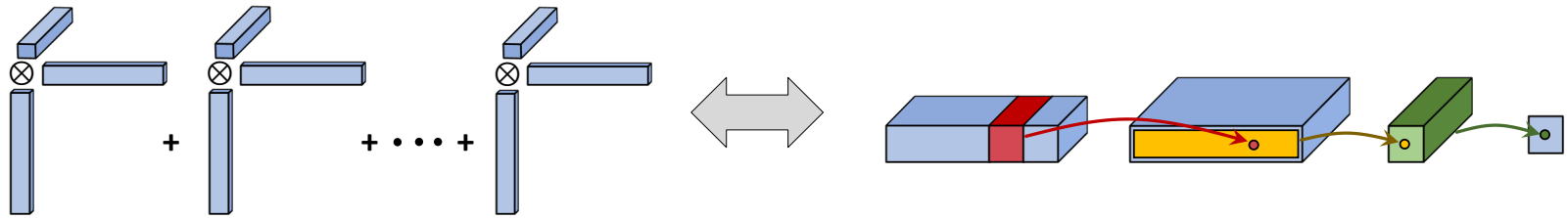
Tensor Rank as a Measure of Complexity

We saw:

- Tensor completion $\longleftrightarrow$ multi-dimensional prediction



- Tensor factorization $\longleftrightarrow$ non-linear NN

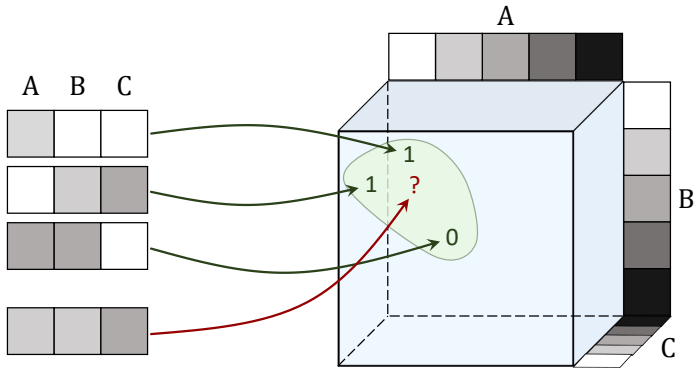


- Implicit regularization favors tensors (predictors) of low tensor rank

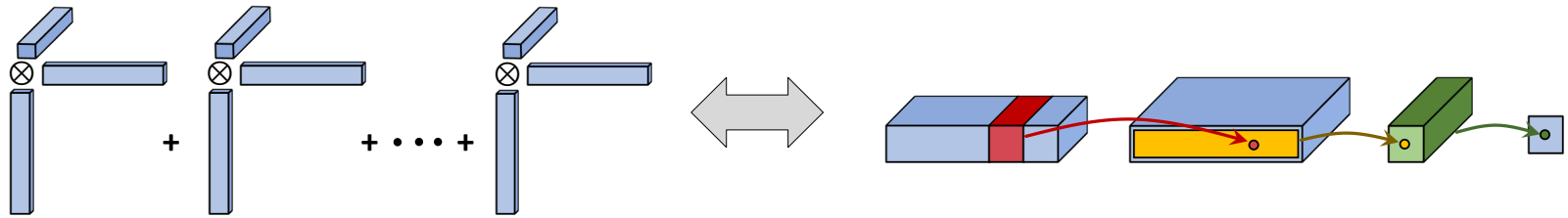
Tensor Rank as a Measure of Complexity

We saw:

- Tensor completion $\longleftrightarrow$ multi-dimensional prediction



- Tensor factorization $\longleftrightarrow$ non-linear NN



- Implicit regularization favors tensors (predictors) of low tensor rank

Can tensor rank serve as a measure of complexity?

Experiment: Fitting Data with Low Tensor Rank

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting standard datasets with predictors of low tensor rank

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting standard datasets with predictors of low tensor rank

Datasets:

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

Experiment: Fitting Data with Low Tensor Rank

Experiment

Fitting standard datasets 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 standard datasets 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 standard datasets 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 standard datasets 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

Thank You

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