Implicit Regularization in Deep Learning May Not Be Explainable by Norms

Noam Razin

based on joint work with Nadav Cohen

Tel Aviv University

Outline

Generalization via Bias-Variance Tradeoff

In classical learning theory generalization exhibits the bias-variance tradeoff

Generalization via Bias-Variance Tradeoff

In classical learning theory generalization exhibits the bias-variance tradeoff

Tradeoff can be controlled through regularization:

Generalization via Bias-Variance Tradeoff

In classical learning theory generalization exhibits the bias-variance tradeoff

Tradeoff can be controlled through regularization:

1. Limiting model size

Generalization via Bias-Variance Tradeoff

In classical learning theory generalization exhibits the bias-variance tradeoff

Tradeoff can be controlled through regularization:

1. Limiting model size
2. Adding term to loss (typically a norm)

Generalization in Deep Learning (DL)

DNNs In Practice

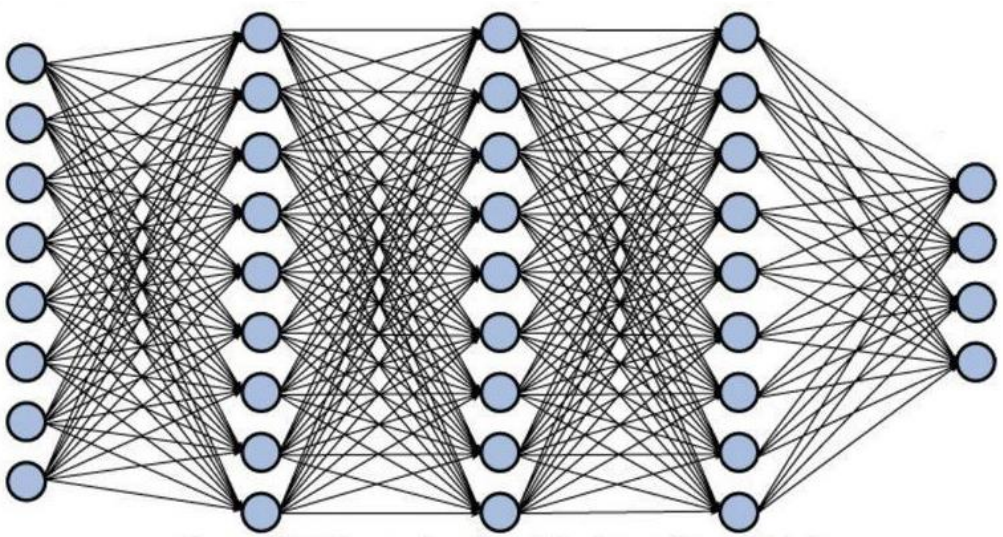
Generalize without explicit regularization:

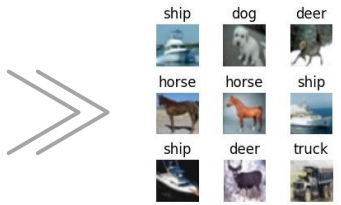
Generalization in Deep Learning (DL)

DNNs In Practice

Generalize without explicit regularization:

1. # of learned weights $\gg$ # of training examples



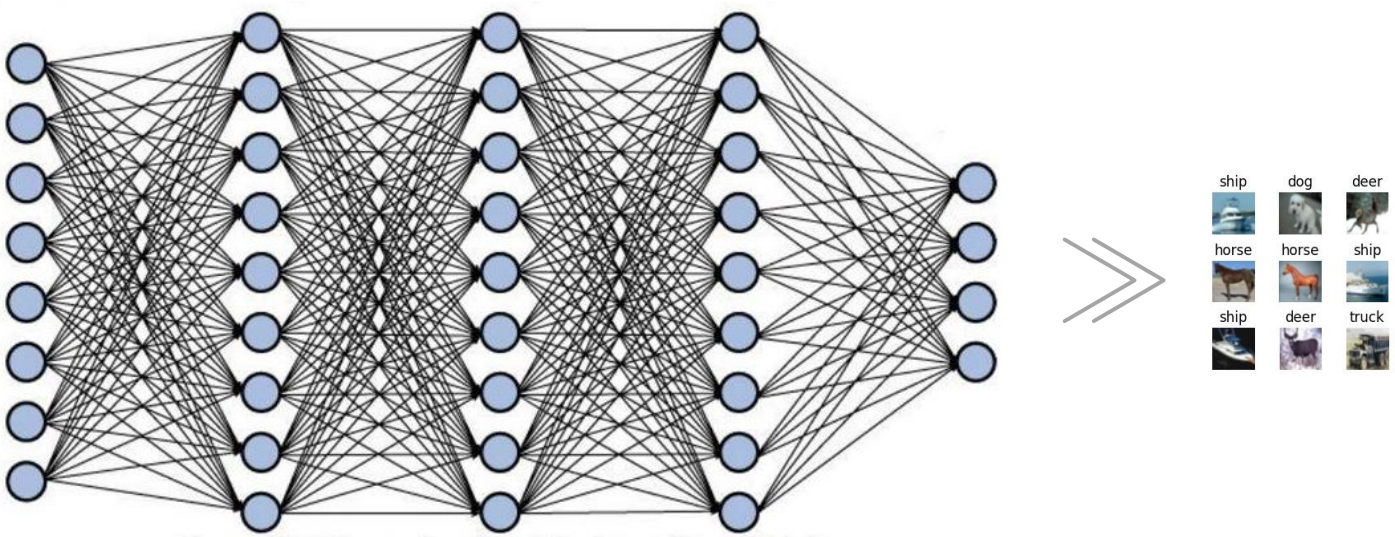


Generalization in Deep Learning (DL)

DNNs In Practice

Generalize without explicit regularization:

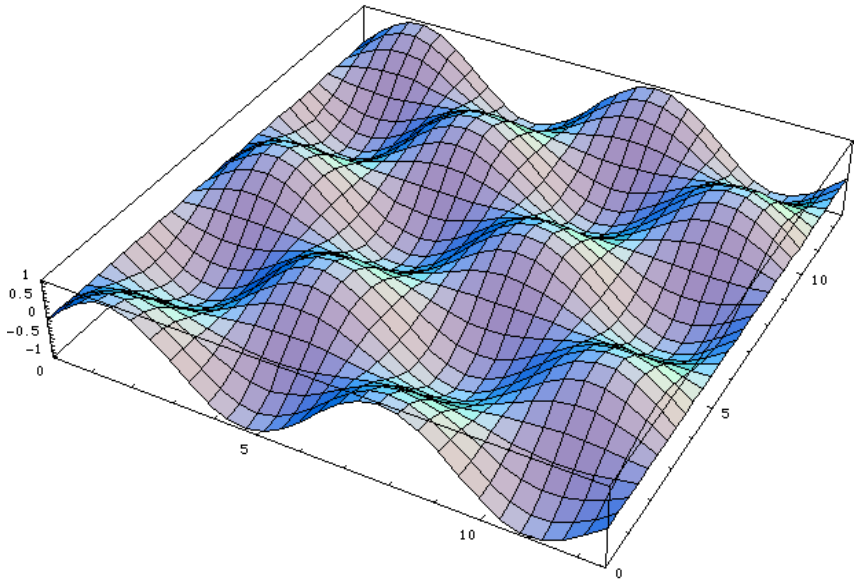
1. # of learned weights $\gg$ # of training examples



2. Loss unchanged (e.g. no weight decay/dropout)

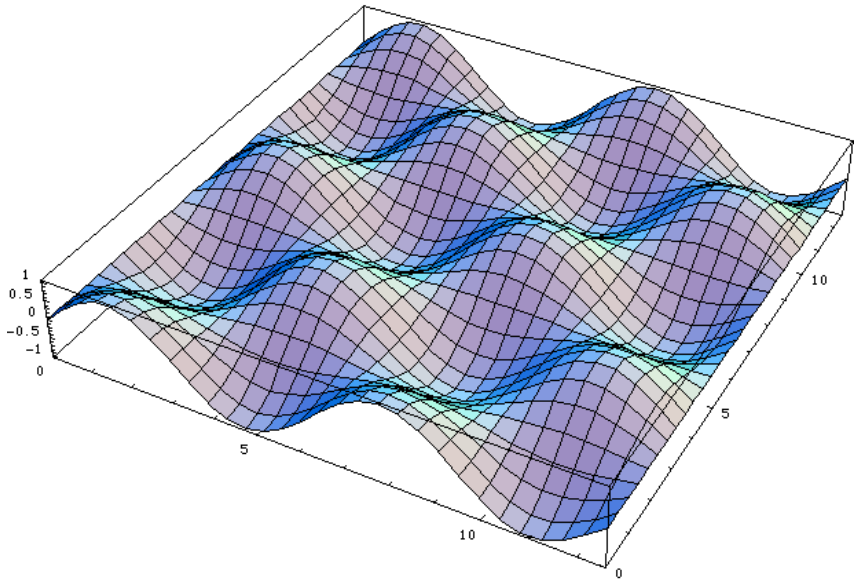
Optimization Induces an Implicit Regularization

Multiple global minima: some generalize well, others don't



Optimization Induces an Implicit Regularization

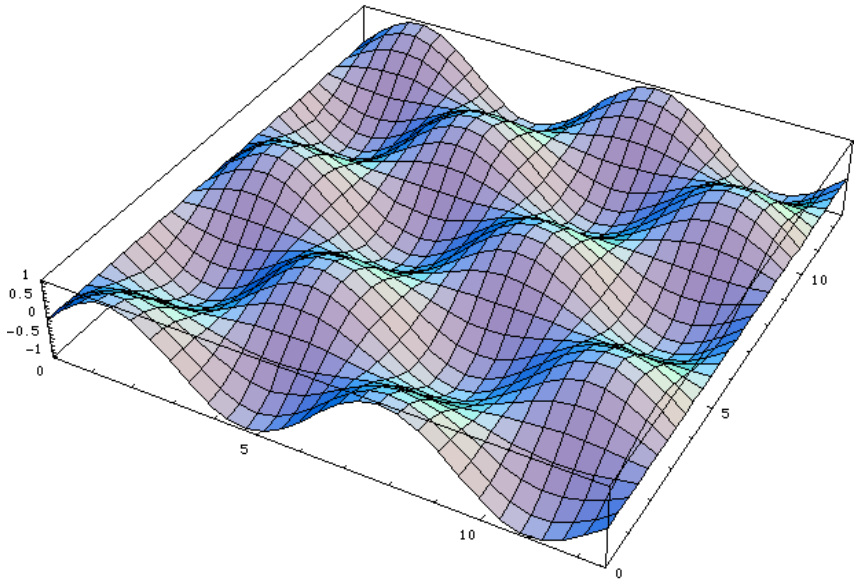
Multiple global minima: some generalize well, others don't



Solution found by Gradient Descent (GD) often generalizes well

Optimization Induces an Implicit Regularization

Multiple global minima: some generalize well, others don't



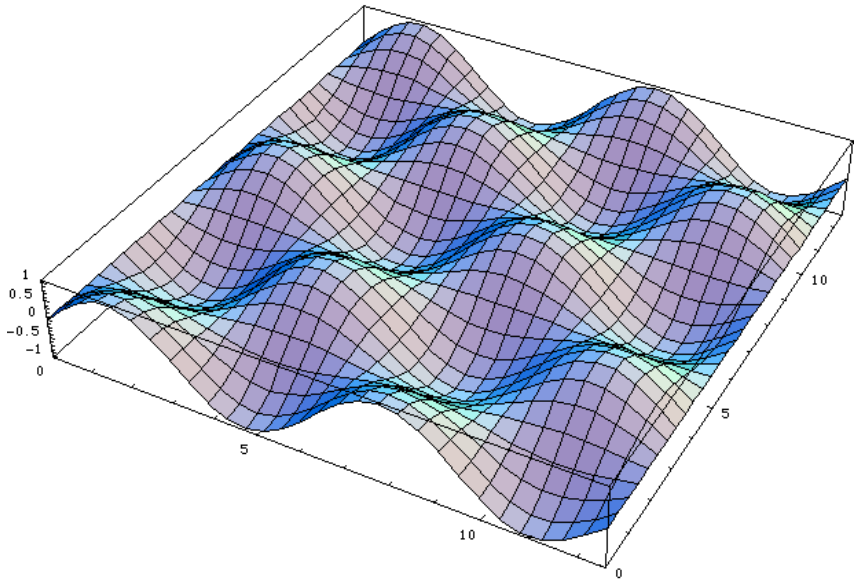
Solution found by Gradient Descent (GD) often generalizes well

Conventional Wisdom

Gradient-based optimization induces an implicit regularization

Optimization Induces an Implicit Regularization

Multiple global minima: some generalize well, others don't



Solution found by Gradient Descent (GD) often generalizes well

Conventional Wisdom

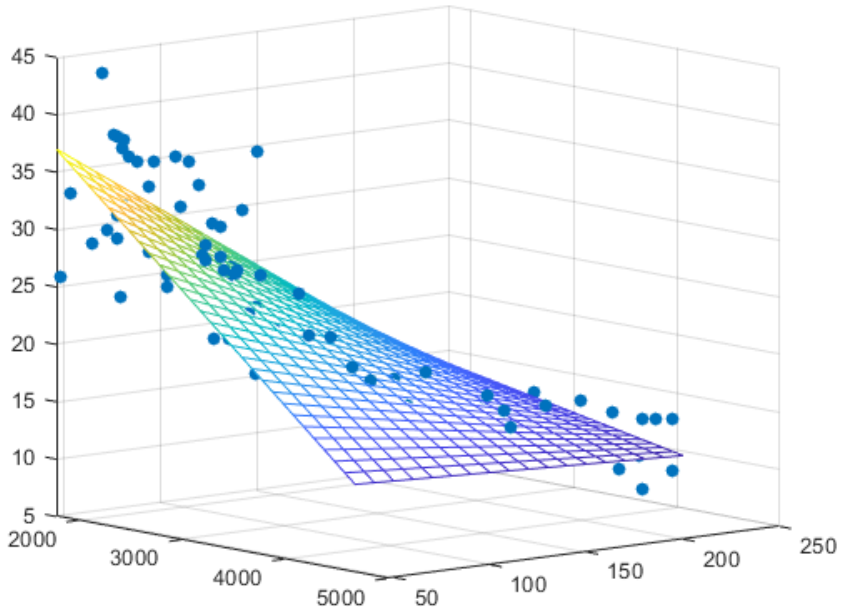
Gradient-based optimization induces an implicit regularization

Question

Can we mathematically understand this effect in concrete settings?

Warm Up: Linear Models

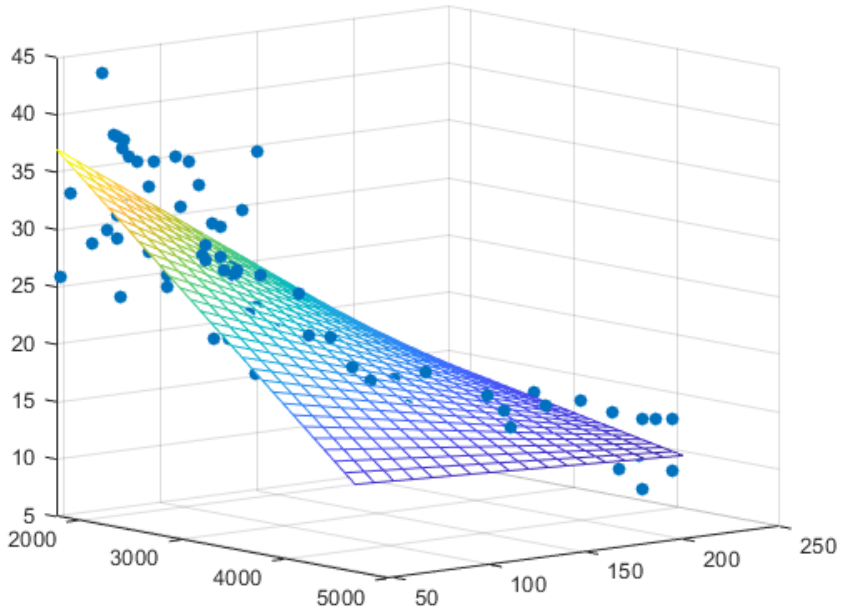
Linear Regression



When # parameters $>$ # training examples:

Warm Up: Linear Models

Linear Regression



When # parameters $>$ # training examples:

GD initialized at 0 converges to min ℓ_2 norm solution

$$\operatorname*{argmin}_{\mathbf{w}} \|\mathbf{w}\|_2 \text{ s.t. } X\mathbf{w} = y$$

Does Implicit Norm Minimization Transfer to DL?

Widespread Hope

GD in DL finds solutions with **minimal norm (or quasi-norm)**

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

Does Implicit Norm Minimization Transfer to DL?

Widespread Hope

GD in DL finds solutions with **minimal norm (or quasi-norm)**

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

Demonstrated in various settings, e.g.:

- Neyshabur et al. 2015
- Gunasekar et al. 2017
- Soudry et al. 2018
- Gunasekar et al. 2018a
- Gunasekar et al. 2018b
- Li et al. 2018
- Jacot et al. 2018
- Mei et al. 2019
- Ji & Telgarsky 2019a
- Ji & Telgarsky 2019b
- Wu et al. 2019
- Oymak & Soltanolkotabi 2019
- Nacson et al. 2019a
- Nacson et al. 2019b
- Woodworth et al. 2020
- Lyu & Li 2020
- Ali et al. 2020
- Chizat & Bach 2020
- Belabbas 2020

Outline

Setting: Matrix Completion

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

Setting: Matrix Completion

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

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

observed entries $\longleftrightarrow$ *training data*

unobserved entries $\longleftrightarrow$ *test data*

Setting: Matrix Completion

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

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

observed entries $\longleftrightarrow$ *training data*

unobserved entries $\longleftrightarrow$ *test data*

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Setting: Matrix Completion

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

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

observed entries $\longleftrightarrow$ *training data*

unobserved entries $\longleftrightarrow$ *test data*

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Setting: Matrix Completion

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

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

observed entries $\longleftrightarrow$ *training data*

unobserved entries $\longleftrightarrow$ *test data*

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Find minimal nuclear norm solution:

$$\min \|W\|_{nuclear} \quad s.t. \quad W_{ij} = b_{ij} \quad \forall (i,j) \in \Omega$$

Setting: Matrix Completion

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

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

observed entries $\longleftrightarrow$ *training data*

unobserved entries $\longleftrightarrow$ *test data*

Denote observations by $\{b_{ij}\}_{(i,j)\in\Omega}$

Convex Programming Approach

Find minimal nuclear norm solution:

$$\min \|W\|_{nuclear} \quad s.t. \quad W_{ij} = b_{ij} \quad \forall (i,j) \in \Omega$$

Perfectly recovers if observations are sufficiently many (Candes & Recht 2008)

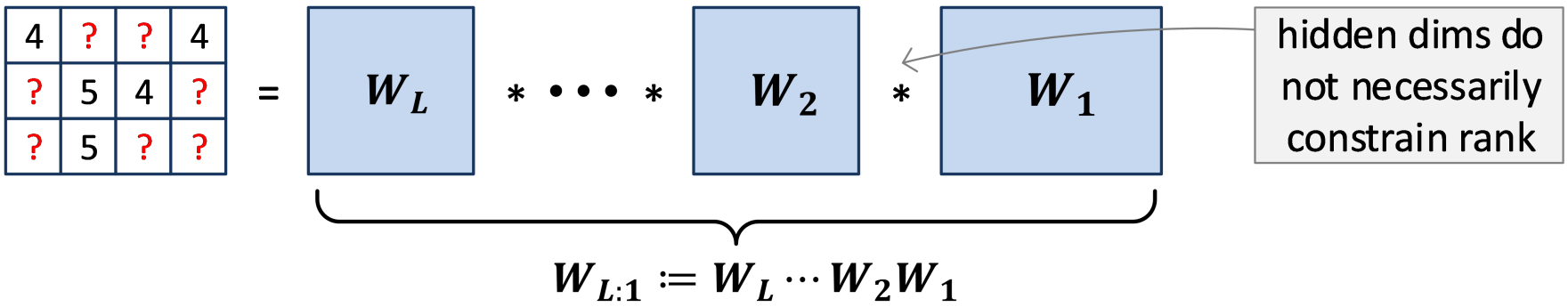
Matrix Factorization $\longleftrightarrow$ Linear Neural Network (LNN)

Deep Learning Approach

Matrix Factorization $\longleftrightarrow$ Linear Neural Network (LNN)

Deep Learning Approach

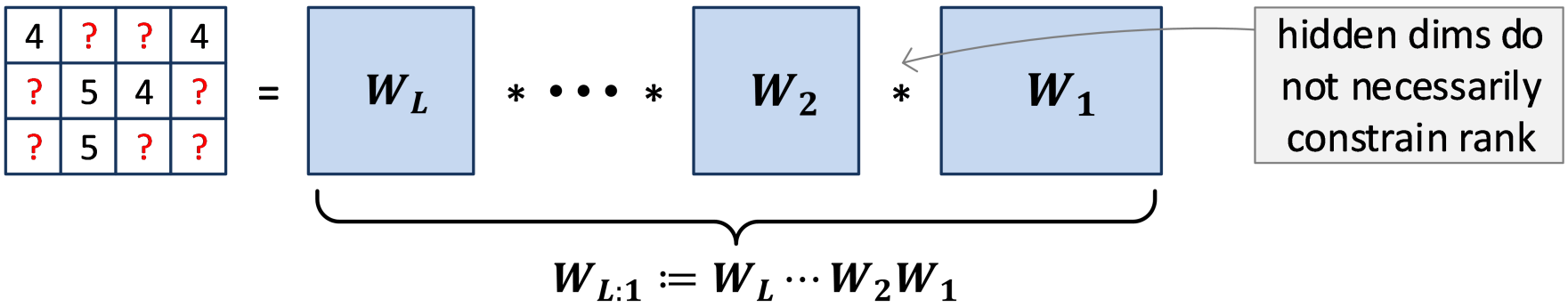
Parameterize solution as LNN and fit observations using GD (over ℓ_2 loss)



Matrix Factorization $\longleftrightarrow$ Linear Neural Network (LNN)

Deep Learning Approach

Parameterize solution as LNN and fit observations using GD (over ℓ_2 loss)



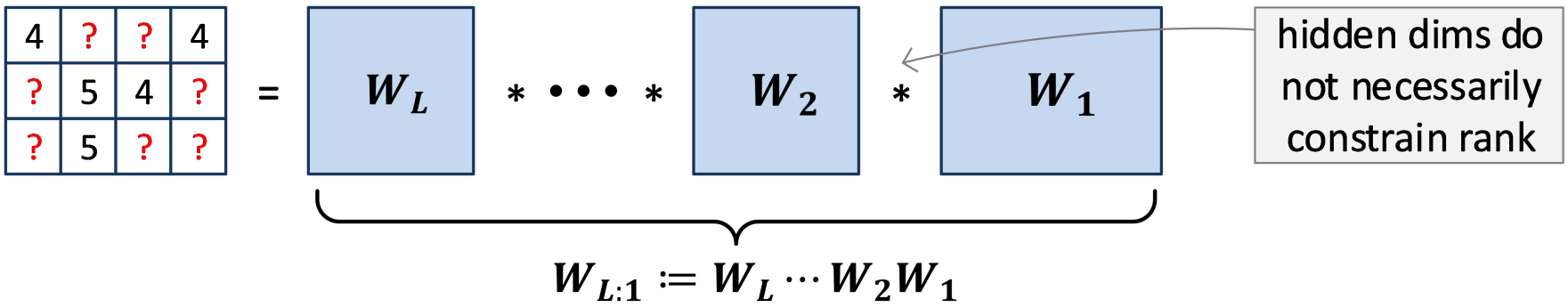
GD is run w.r.t. $W_1, \ldots, W_L$ over:

$$\ell(W_{L:1}) = \frac{1}{2} \sum_{(i,j) \in \Omega} \left((W_{L:1})_{ij} - b_{ij} \right)^2$$

Matrix Factorization $\longleftrightarrow$ Linear Neural Network (LNN)

Deep Learning Approach

Parameterize solution as LNN and fit observations using GD (over ℓ_2 loss)



GD is run w.r.t. $W_1, \ldots, W_L$ over:

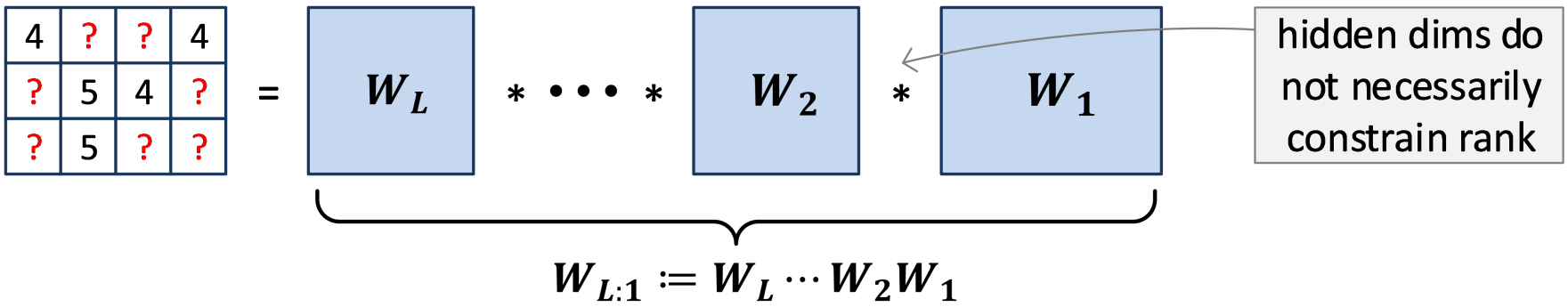
$$\ell(W_{L:1}) = \frac{1}{2} \sum_{(i,j) \in \Omega} \left((W_{L:1})_{ij} - b_{ij} \right)^2$$

To which solutions does **product matrix** $W_{L:1}$ converge?

Matrix Factorization $\longleftrightarrow$ Linear Neural Network (LNN)

Deep Learning Approach

Parameterize solution as LNN and fit observations using GD (over ℓ_2 loss)



GD is run w.r.t. $W_1, \ldots, W_L$ over:

$$\ell(W_{L:1}) = \frac{1}{2} \sum_{(i,j) \in \Omega} \left((W_{L:1})_{ij} - b_{ij} \right)^2$$

To which solutions does **product matrix** $W_{L:1}$ converge?

Empirical phenomenon: low-rank matrix often recovered accurately

Conjecture: Nuclear Norm Minimization

Conjecture (Gunasekar et al. 2017)

With small learning rate and init close to the origin, GD over depth 2 matrix factorization converges to min nuclear norm solution.

Conjecture: Nuclear Norm Minimization

Conjecture (Gunasekar et al. 2017)

With small learning rate and init close to the origin, GD over depth 2 matrix factorization converges to min nuclear norm solution.

GD implicitly solves convex programming approach?

Conjecture: Nuclear Norm Minimization

Conjecture (Gunasekar et al. 2017)

With small learning rate and init close to the origin, GD over depth 2 matrix factorization converges to min nuclear norm solution.

GD implicitly solves convex programming approach?

Gunasekar et al. supported conjecture with:

- Experiments
- Proof for certain restricted case

Conjecture: Nuclear Norm Minimization

Conjecture (Gunasekar et al. 2017)

With small learning rate and init close to the origin, GD over depth 2 matrix factorization converges to min nuclear norm solution.

GD implicitly solves convex programming approach?

Gunasekar et al. supported conjecture with:

- Experiments
- Proof for certain restricted case

Conjecture established under other restricted conditions:

- Li et al. 2018
- Belabbas 2020

Conjecture: No Norm is Being Minimized

Conjecture (Arora et al. 2019)

For any norm $\|\cdot\|$, exist observations for which small learning rate and init can not ensure GD converges to min $\|\cdot\|$ solution.

Conjecture: No Norm is Being Minimized

Conjecture (Arora et al. 2019)

For any norm $\|\cdot\|$, exist observations for which small learning rate and init can not ensure GD converges to min $\|\cdot\|$ solution.

Arora et al. supported conjecture with:

- Experiments: nuclear norm not always minimized, bias to low rank

Conjecture: No Norm is Being Minimized

Conjecture (Arora et al. 2019)

For any norm $\|\cdot\|$, exist observations for which small learning rate and init can not ensure GD converges to min $\|\cdot\|$ solution.

Arora et al. supported conjecture with:

- Experiments: nuclear norm not always minimized, bias to low rank
- Dynamical analysis of singular values: GD promotes sparse spectrum

Conjecture: No Norm is Being Minimized

Conjecture (Arora et al. 2019)

For any norm $\|\cdot\|$, exist observations for which small learning rate and init can not ensure GD converges to min $\|\cdot\|$ solution.

Arora et al. supported conjecture with:

- Experiments: nuclear norm not always minimized, bias to low rank
- Dynamical analysis of singular values: GD promotes sparse spectrum

↑

suggests tendency to low rank

Conjecture: No Norm is Being Minimized

Conjecture (Arora et al. 2019)

For any norm $\|\cdot\|$, exist observations for which small learning rate and init can not ensure GD converges to min $\|\cdot\|$ solution.

Arora et al. supported conjecture with:

- Experiments: nuclear norm not always minimized, bias to low rank
- Dynamical analysis of singular values: GD promotes sparse spectrum

  ↑

  suggests tendency to low rank

Open Question

Does the implicit regularization in matrix factorization minimize a norm?

Our Work: Resolving Open Question (Negatively)

Theorem (informal)

There exist matrix factorization settings where:

Our Work: Resolving Open Question (Negatively)

Theorem (informal)

There exist matrix factorization settings where:

1. ***All norms (and quasi-norms) are driven towards*** ∞

Our Work: Resolving Open Question (Negatively)

Theorem (informal)

There exist matrix factorization settings where:

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

Our Work: Resolving Open Question (Negatively)

Theorem (informal)

There exist matrix factorization settings where:

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

Affirms conjecture of Arora et al. 2019

Our Work: Resolving Open Question (Negatively)

Theorem (informal)

There exist matrix factorization settings where:

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

Affirms conjecture of Arora et al. 2019

Result stronger than conjecture:

Our Work: Resolving Open Question (Negatively)

Theorem (informal)

There exist matrix factorization settings where:

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

Affirms conjecture of Arora et al. 2019

Result stronger than conjecture:

- Settings jointly disqualify **all** norms
- Norms driven towards ∞

Outline

Optimization Scheme

Common surrogate for GD with small learning rate and init:[1]

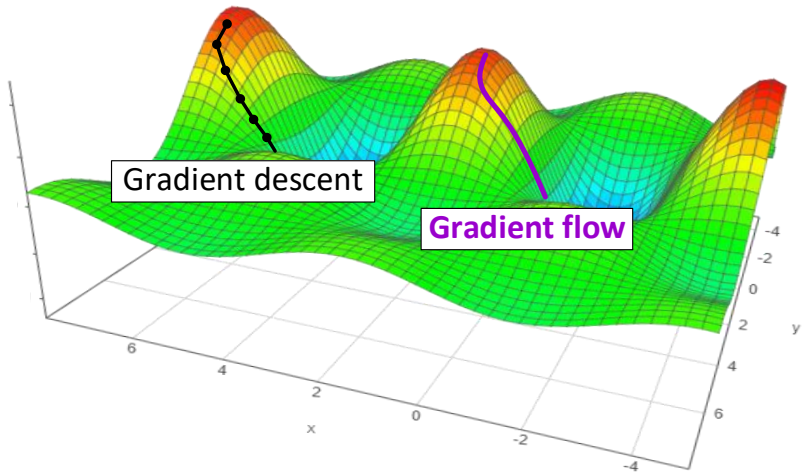
[1] (e.g. used by Gunasekar et al. 2017, Arora et al. 2019)

Optimization Scheme

Common surrogate for GD with small learning rate and init:[1]

Gradient flow (GF) is a continuous version of GD (step size $\to 0$):

$$\frac{d}{dt}\mathbf{w}(t) = -\nabla f(\mathbf{w}(t)) \quad , t \in \mathbb{R}_{\geq 0}$$



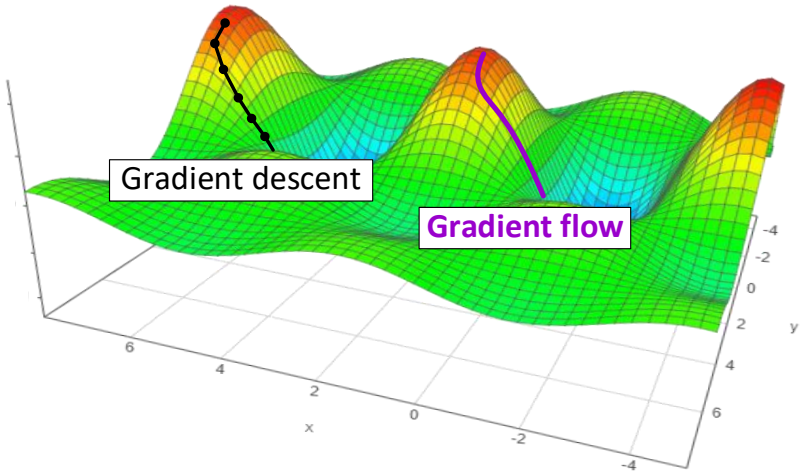
[1] (e.g. used by Gunasekar et al. 2017, Arora et al. 2019)

Optimization Scheme

Common surrogate for GD with small learning rate and init:[1]

Gradient flow (GF) is a continuous version of GD (step size $\to 0$):

$$\frac{d}{dt}\mathbf{w}(t) = -\nabla f(\mathbf{w}(t)) \quad , t \in \mathbb{R}_{\geq 0}$$



Weights $W_1 \dots W_L$ are **balanced** at init: $W_{j+1}^\top W_{j+1} = W_j W_j^\top \ , \forall j$.

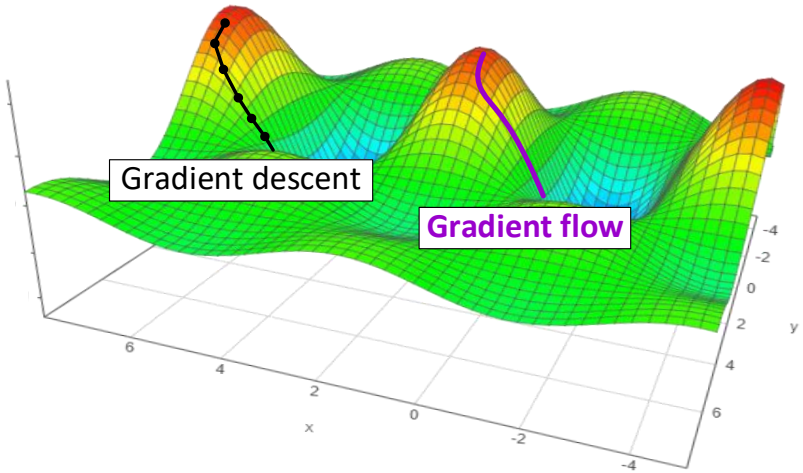
[1] (e.g. used by Gunasekar et al. 2017, Arora et al. 2019)

Optimization Scheme

Common surrogate for GD with small learning rate and init:[1]

Gradient flow (GF) is a continuous version of GD (step size $\to 0$):

$$\frac{d}{dt}\mathbf{w}(t) = -\nabla f(\mathbf{w}(t)) \quad , t \in \mathbb{R}_{\geq 0}$$



Weights $W_1 \dots W_L$ are **balanced** at init: $W_{j+1}^\top W_{j+1} = W_j W_j^\top \ , \forall j.$

↑

Holds approximately under ≈ 0 init, exactly under residual (I_d) init

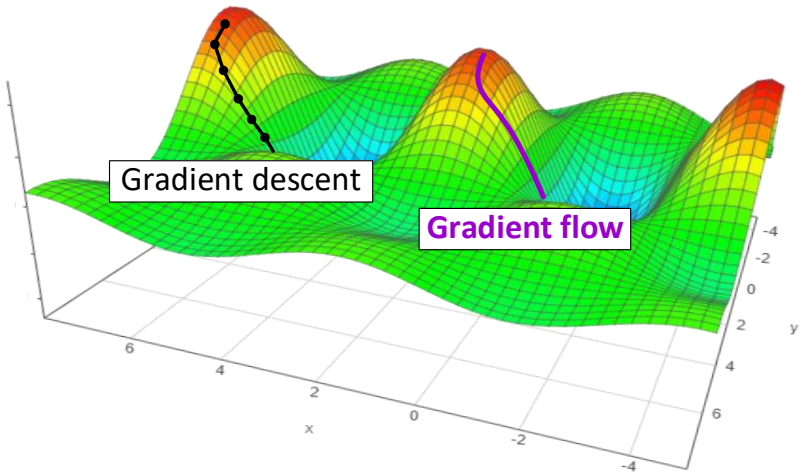
[1] (e.g. used by Gunasekar et al. 2017, Arora et al. 2019)

Optimization Scheme

Common surrogate for GD with small learning rate and init:[1]

Gradient flow (GF) is a continuous version of GD (step size $\to 0$):

$$\frac{d}{dt}\mathbf{w}(t) = -\nabla f(\mathbf{w}(t)) \quad , t \in \mathbb{R}_{\geq 0}$$



Weights $W_1 \dots W_L$ are **balanced** at init: $W_{j+1}^\top W_{j+1} = W_j W_j^\top \ , \forall j.$

↑

Holds approximately under ≈ 0 init, exactly under residual (I_d) init

Closely matches GD in practice for linear neural networks

[1] (e.g. used by Gunasekar et al. 2017, Arora et al. 2019)

A Simple Matrix Completion Problem

Consider the following 2-by-2 matrix completion problem:

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

A Simple Matrix Completion Problem

Consider the following 2-by-2 matrix completion problem:

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

Min Frobenius norm $\iff * = 0$ (depth $L = 1$ solution)

A Simple Matrix Completion Problem

Consider the following 2-by-2 matrix completion problem:

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

Min Frobenius norm $\iff * = 0$ (depth $L = 1$ solution)

Min nuclear norm $\iff * = 0$

A Simple Matrix Completion Problem

Consider the following 2-by-2 matrix completion problem:

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

Min Frobenius norm $\iff * = 0$ (depth $L = 1$ solution)

Min nuclear norm $\iff * = 0$

Generally: Schatten-p (quasi-)norm $:= \left(\sum_i \sigma_i^p\right)^{1/p}$ (σ_i - singular values)

A Simple Matrix Completion Problem

Consider the following 2-by-2 matrix completion problem:

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

Min Frobenius norm $\iff * = 0$ (depth $L = 1$ solution)

Min nuclear norm $\iff * = 0$

Generally: Schatten-p (quasi-)norm $:= \left(\sum_i \sigma_i^p\right)^{1/p}$ (σ_i - singular values)

Proposition

Schatten-p (quasi-)norms are minimized $\iff * = 0$

A Simple Matrix Completion Problem

Consider the following 2-by-2 matrix completion problem:

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

Min Frobenius norm $\iff * = 0$ (depth $L = 1$ solution)

Min nuclear norm $\iff * = 0$

Generally: Schatten-p (quasi-)norm $:= \left(\sum_i \sigma_i^p\right)^{1/p}$ (σ_i - singular values)

Proposition

Schatten-p (quasi-)norms are minimized $\iff * = 0$

Arbitrary norms (or quasi-norms):

Proposition

For minimal $\|\cdot\|$ solutions, $*$ is bounded

A Simple Matrix Completion Problem (Cont'd)

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

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially *rank* $\to 1$

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially *rank* $\to 1$

Continuous measures for rank:

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially *rank* $\to 1$

Continuous measures for rank:

- *Distance from rank 1* := value of smallest singular value

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially *rank* $\to 1$

Continuous measures for rank:

- *Distance from rank 1* := value of smallest singular value
- *Effective rank (erank)* $\approx$ entropy of singular values

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially $\mathit{rank} \to 1$

Continuous measures for rank:

- *Distance from rank 1* := value of smallest singular value
- *Effective rank (erank)* $\approx$ entropy of singular values

Definition (Rank suboptimality)

$\mathit{rank\text{-}subopt} := \max\{\mathit{Distance\ from\ rank\ 1}\ ,\ \mathit{erank} - \inf_* \mathit{erank}\}$

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially $\mathit{rank} \to 1$

Continuous measures for rank:

- *Distance from rank 1* := value of smallest singular value
- *Effective rank (erank)* $\approx$ entropy of singular values

Definition (Rank suboptimality)

$\mathit{rank\text{-}subopt} := \max\{\mathit{Distance\ from\ rank\ 1}\ ,\ \mathit{erank} - \inf_* \mathit{erank}\}$

Proposition

rank-subopt is maximized when $* = 0$, *and is minimized as* $|*| \to \infty$

A Simple Matrix Completion Problem (Cont'd)

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

(i) All solutions are of rank 2 ***(ii)*** $|*| \to \infty \Rightarrow$ essentially *rank* $\to 1$

Continuous measures for rank:

- *Distance from rank 1* := value of smallest singular value
- *Effective rank (erank)* $\approx$ entropy of singular values

Definition (Rank suboptimality)

rank-subopt $:= \max\{$*Distance from rank 1* , *erank* $- \inf_*$ *erank*$\}$

Proposition

rank-subopt is maximized when $* = 0$, *and is minimized as* $|*| \to \infty$

Contradiction between norm and rank minimization

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

$\ell(t) \to 0$ implies:

1. ***All* norms (and quasi-norms) driven towards ∞**
2. **Rank is essentially minimized**

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

$\ell(t) \to 0$ implies:

1. ***All* norms (and quasi-norms) driven towards ∞**
2. **Rank is essentially minimized**

Claim

Assumption on $\det(W_{L:1}(0))$ *holds with prob* 0.5 *under standard inits*

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

$\ell(t) \to 0$ implies:

1. ***All* norms (and quasi-norms) driven towards ∞**
2. **Rank is essentially minimized**

Claim

Assumption on $\det(W_{L:1}(0))$ *holds with prob* 0.5 *under standard inits*

Implicit regularization $\neq$ norm minimization!

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$ — Proof Sketch

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch

[1] based on singular values differential equations from Arora et al. 2019

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$ — Proof Sketch

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch

GF trajectory analysis: $\det(W_{L:1}(t))$ does not change sign[1]

[1] based on singular values differential equations from Arora et al. 2019

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$ — Proof Sketch

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch

GF trajectory analysis: $\det(W_{L:1}(t))$ does not change sign[1]

$$\det(W_{L:1}(t)) = w_{1,1}(t)w_{2,2}(t) - w_{1,2}(t)w_{2,1}(t) > 0$$

[1] based on singular values differential equations from Arora et al. 2019

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$ — Proof Sketch

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch

GF trajectory analysis: $\det(W_{L:1}(t))$ does not change sign[1]

$$\det(W_{L:1}(t)) = w_{1,1}(t)\underbrace{w_{2,2}(t)}_{\to\ 0} - \underbrace{w_{1,2}(t)w_{2,1}(t)}_{\to\ 1} > 0$$

[1] based on singular values differential equations from Arora et al. 2019

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$ — Proof Sketch

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch

GF trajectory analysis: $\det(W_{L:1}(t))$ does not change sign[1]

$$\det(W_{L:1}(t)) = w_{1,1}(t)\underbrace{w_{2,2}(t)}_{\to\ 0} - \underbrace{w_{1,2}(t)w_{2,1}(t)}_{\to\ 1} > 0$$

$\Longrightarrow |w_{1,1}(t)| \to \infty$ (behaves as $\Omega(1/\sqrt{\ell(t)})$)

[1] based on singular values differential equations from Arora et al. 2019

Loss $\searrow \Rightarrow$ Norms $\nearrow$ Rank $\searrow$ — Proof Sketch

Theorem

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch

GF trajectory analysis: $\det(W_{L:1}(t))$ does not change sign[1]

$$\det(W_{L:1}(t)) = w_{1,1}(t)\underbrace{w_{2,2}(t)}_{\to\ 0} - \underbrace{w_{1,2}(t)w_{2,1}(t)}_{\to\ 1} > 0$$

$\Longrightarrow |w_{1,1}(t)| \to \infty$ (behaves as $\Omega(1/\sqrt{\ell(t)})$)

Bound on $|w_{1,1}(t)|$ implies bounds for norms and rank suboptimality

[1] based on singular values differential equations from Arora et al. 2019

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

$\uparrow$

commonly observed in practice

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

↑

commonly observed in practice

Convergence to Zero Loss (in our setting)

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

$\uparrow$

commonly observed in practice

Convergence to Zero Loss (in our setting)

- Experiments: GD consistently finds global min

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

$\uparrow$

commonly observed in practice

Convergence to Zero Loss (in our setting)

- Experiments: GD consistently finds global min
- Proof for depth 2 with scaled identity init

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

$\uparrow$

commonly observed in practice

Convergence to Zero Loss (in our setting)

- Experiments: GD consistently finds global min
- Proof for depth 2 with scaled identity init

Proposition

If at init $W_{L:1}(0) = \alpha \cdot I$, *for depth* $L = 2$ *and* $0 < \alpha \leq 1$, *then* $\ell(t) \to 0$

Convergence to Global Minimum

Customary

Separating aspects of convergence to global min and implicit regularization

$\uparrow$

commonly observed in practice

Convergence to Zero Loss (in our setting)

- Experiments: GD consistently finds global min
- Proof for depth 2 with scaled identity init

Proposition

If at init $W_{L:1}(0) = \alpha \cdot I$, *for depth* $L = 2$ *and* $0 < \alpha \leq 1$, *then* $\ell(t) \to 0$

Proof Approach

Careful analysis of GF differential equations

Robustness to Perturbations

What happens when observations are perturbed?

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

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix}$$

non-zero z, z'
arbitrary ϵ

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem (original setting)

If $\det(W_{L:1}(0)) > 0$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem (original setting)

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(1 \cdot 1)$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{array}{l} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{array}$$

Theorem (intermediate)

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(z \cdot z')$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega(1/\sqrt{\ell(t)})$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem (intermediate)

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(z \cdot z')$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega\left(\min\{|z|, |z'|\} \;/\; (|\epsilon| + \sqrt{2\ell(t)})\right)$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(z \cdot z')$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega\left(\min\{|z|, |z'|\} \;/\; (|\epsilon| + \sqrt{2\ell(t)})\right)$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}\left((|\epsilon| + \sqrt{2\ell(t)}) \;/\; \min\{|z|, |z'|\}\right)$

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(z \cdot z')$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega\left(\min\{|z|, |z'|\} \;/\; (|\epsilon| + \sqrt{2\ell(t)})\right)$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}\left((|\epsilon| + \sqrt{2\ell(t)}) \;/\; \min\{|z|, |z'|\}\right)$

- $\epsilon = 0 \Rightarrow$ **all norms driven to ∞ and rank is minimized**

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(z \cdot z')$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega\left(\min\{|z|, |z'|\} \;/\; (|\epsilon| + \sqrt{2\ell(t)})\right)$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}\left((|\epsilon| + \sqrt{2\ell(t)}) \;/\; \min\{|z|, |z'|\}\right)$

- $\epsilon = 0 \Rightarrow$ **all norms driven to ∞ and rank is minimized**
- Phenomenon gracefully recedes as ϵ perturbed from 0

Robustness to Perturbations

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} * & z \\ z' & \epsilon \end{pmatrix} \qquad \begin{matrix} \text{non-zero } z, z' \\ \text{arbitrary } \epsilon \end{matrix}$$

Theorem

If $\operatorname{sign}(\det(W_{L:1}(0))) = \operatorname{sign}(z \cdot z')$ *at init, then for any (quasi-)norm* $\|\cdot\|$*:*

1. $\|W_{L:1}(t)\| = \Omega\left(\min\{|z|, |z'|\} \;/\; (|\epsilon| + \sqrt{2\ell(t)})\right)$
2. *rank-subopt*$(W_{L:1}(t)) = \mathcal{O}\left((|\epsilon| + \sqrt{2\ell(t)}) \;/\; \min\{|z|, |z'|\}\right)$

- $\epsilon = 0 \Rightarrow$ **all norms driven to ∞ and rank is minimized**
- Phenomenon gracefully recedes as ϵ perturbed from 0

Same results hold when changing unobserved entry location

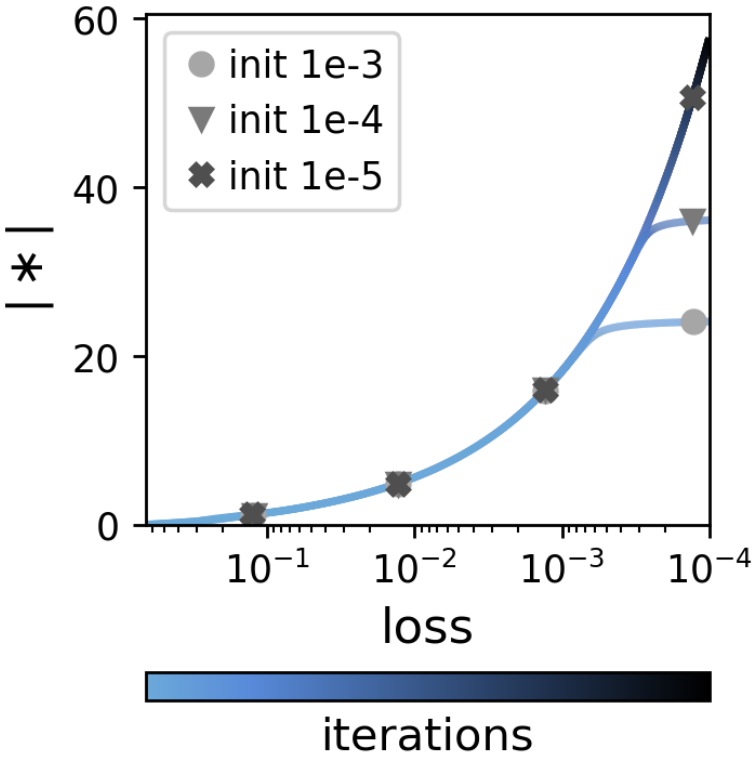
Experiments: Unobserved Entry ↗

Experiments: Unobserved Entry ↗

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

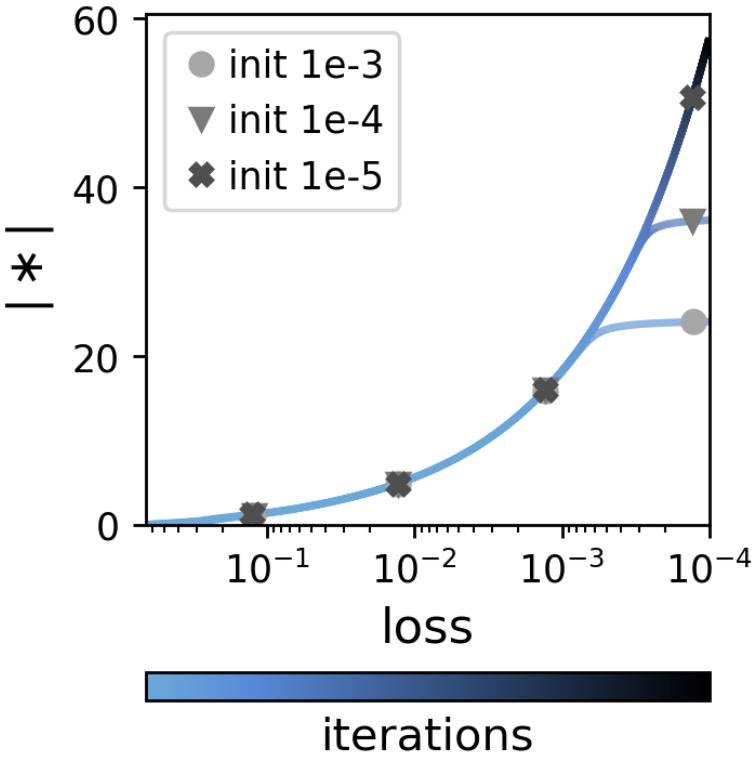
Experiments: Unobserved Entry ↗

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



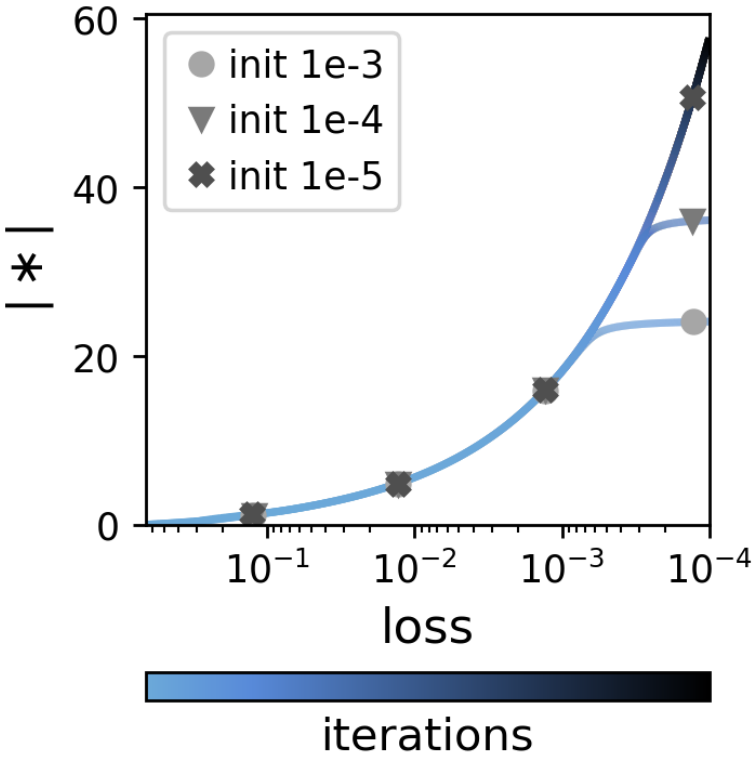
Experiments: Unobserved Entry ↗

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} * & -2 \\ 0.5 & \epsilon \end{pmatrix}$$

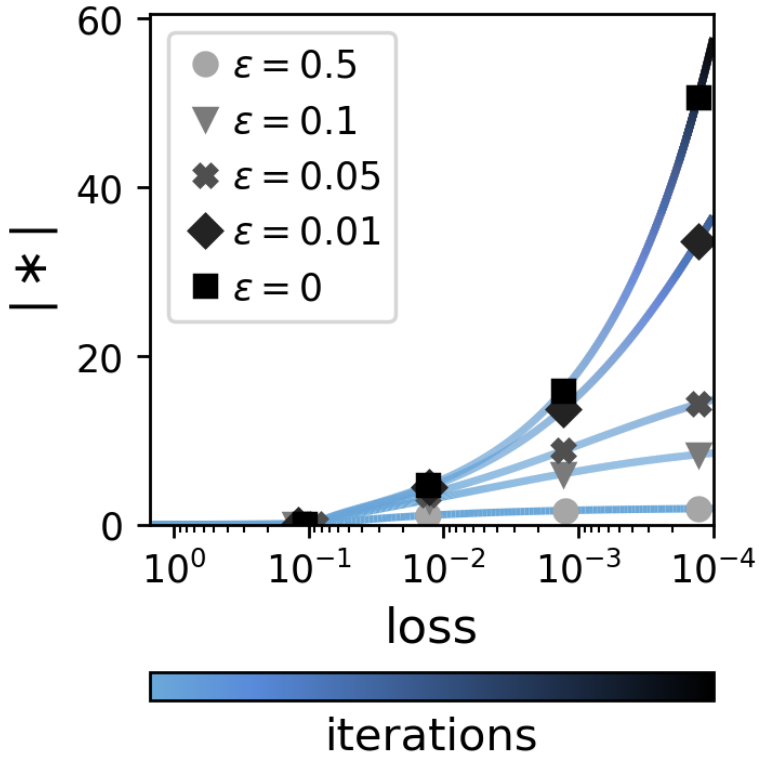


Experiments: Unobserved Entry ↗

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

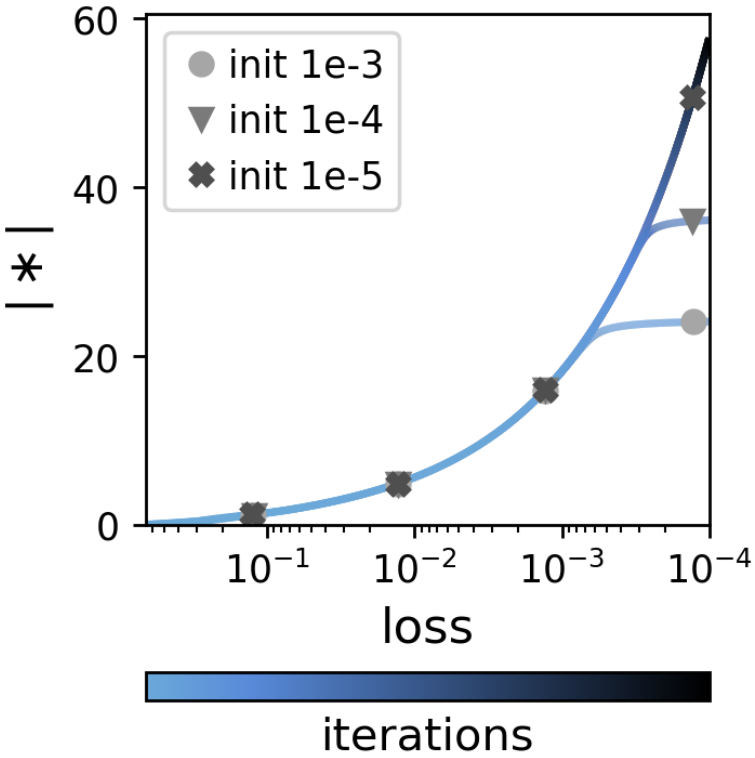


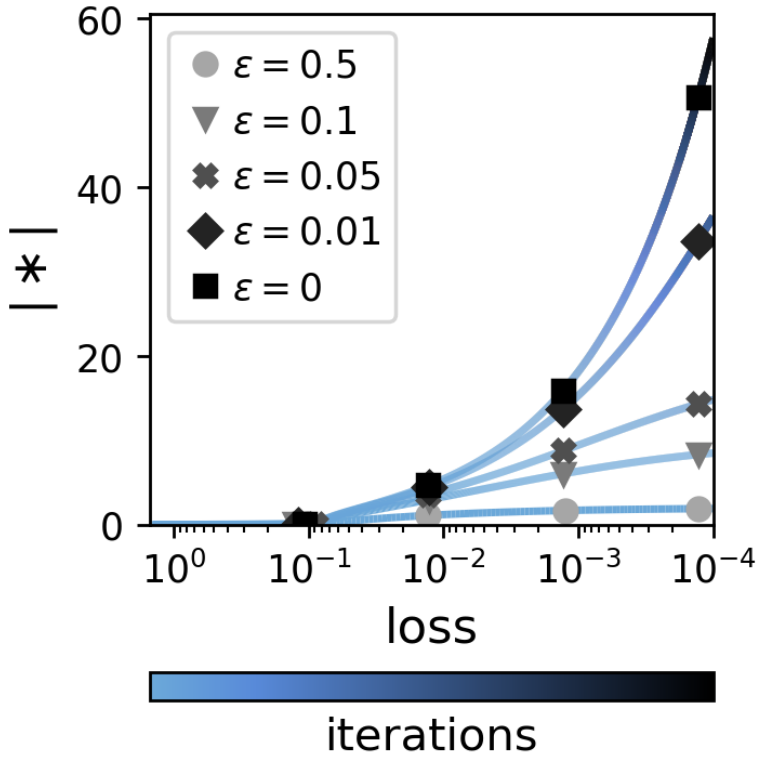
$$\begin{pmatrix} * & -2 \\ 0.5 & \epsilon \end{pmatrix}$$



Experiments: Unobserved Entry ↗

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} * & -2 \\ 0.5 & \epsilon \end{pmatrix}$$





Theory transfers to practice: unobserved entry $\to \infty$

Outline

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms to ∞ to minimize rank

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms to ∞ to minimize rank

Past Work

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms to ∞ to minimize rank

Past Work

- Empirical evidence: low-rank tendency in matrix factorization

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms to ∞ to minimize rank

Past Work

- Empirical evidence: low-rank tendency in matrix factorization
- Theoretical analysis: sparsity in singular values (Arora et al. 2019)

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms to ∞ to minimize rank

Past Work

- Empirical evidence: low-rank tendency in matrix factorization
- Theoretical analysis: sparsity in singular values (Arora et al. 2019)

Better interpretation — rank minimization?

Implicit Regularization = Rank Minimization?

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms to ∞ to minimize rank

Past Work

- Empirical evidence: low-rank tendency in matrix factorization
- Theoretical analysis: sparsity in singular values (Arora et al. 2019)

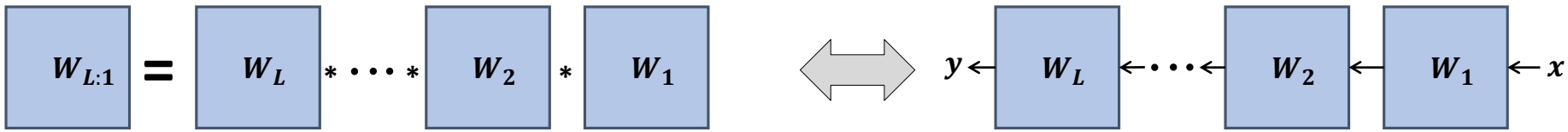
Better interpretation — rank minimization?

Does this interpretation extend beyond matrix factorization?

Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

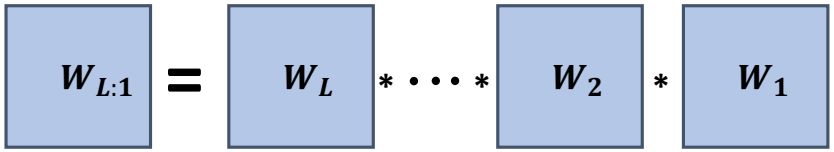
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Matrix Factorizations ***Linear Neural Networks***



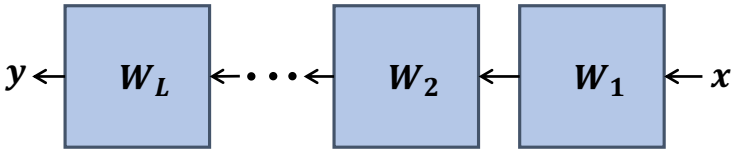
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Matrix Factorizations

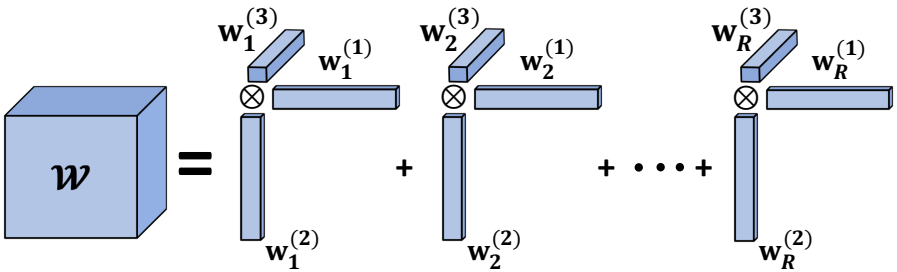


$\Longleftrightarrow$

Linear Neural Networks



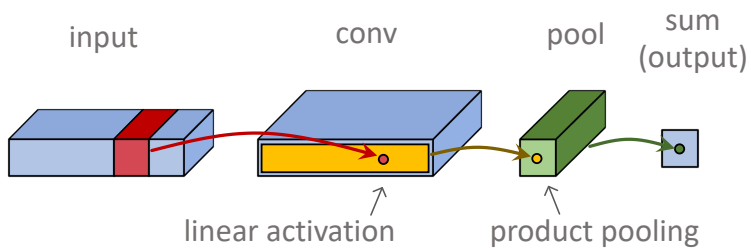
Tensor Factorizations



$\Longleftrightarrow$

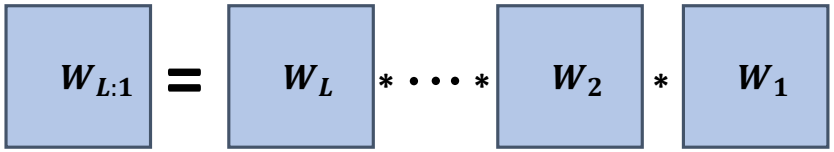
Convolutional Arithmetic Circuits

(Non-Linear Neural Networks)

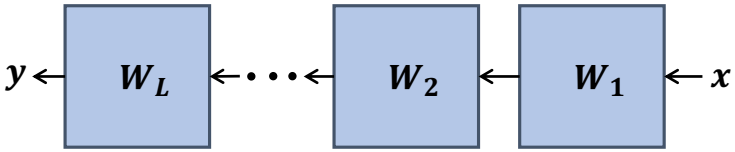


Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

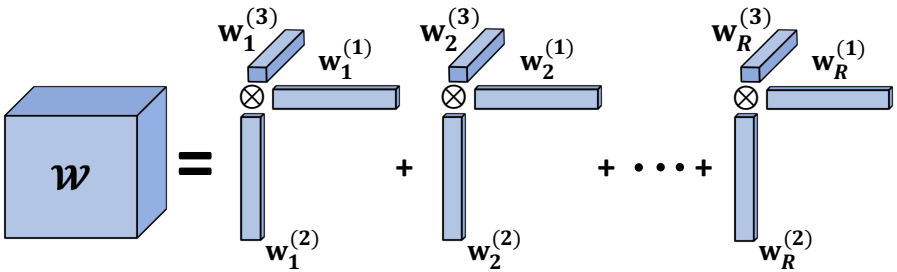
Matrix Factorizations



Linear Neural Networks

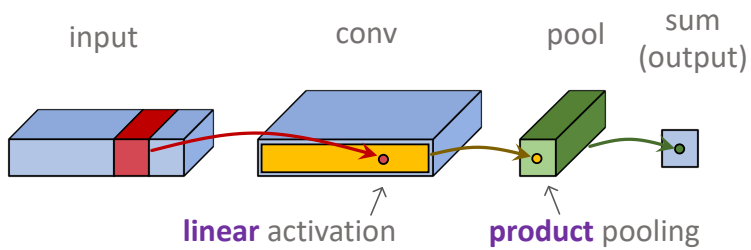


Tensor Factorizations



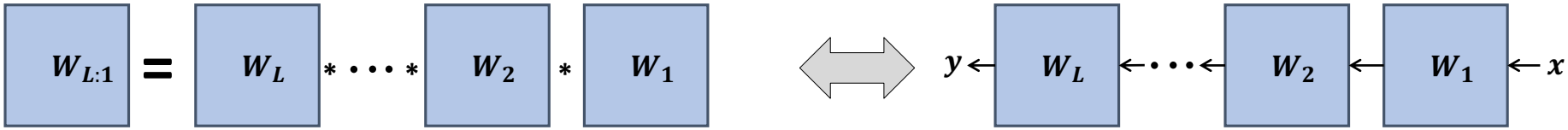
Convolutional Arithmetic Circuits

(Non-Linear Neural Networks)



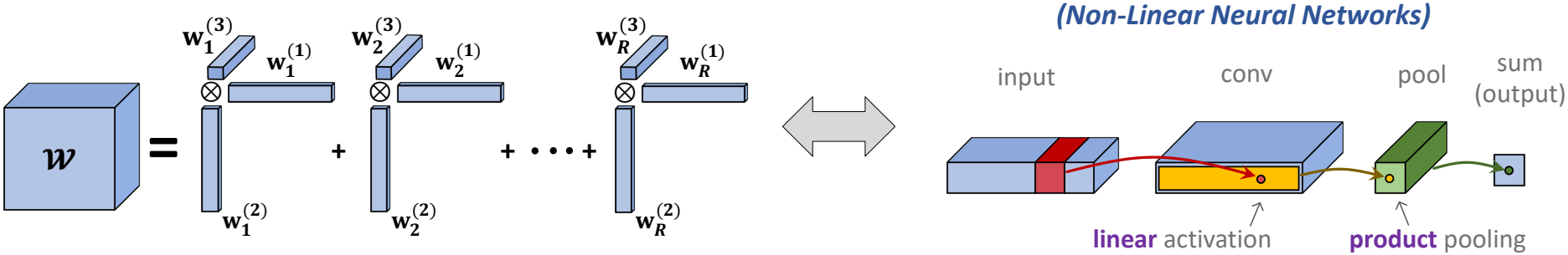
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Matrix Factorizations $\longleftrightarrow$ ***Linear Neural Networks***



Tensor Factorizations $\longleftrightarrow$ ***Convolutional Arithmetic Circuits***

(Non-Linear Neural Networks)

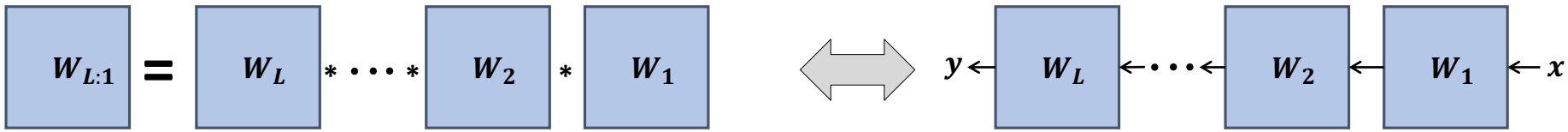


ConvACs are competitive in practice, and admit algebraic structure

Extensively studied (e.g. Cohen et al. 2016, Cohen & Shashua 2016, Cohen & Shashua 2017)

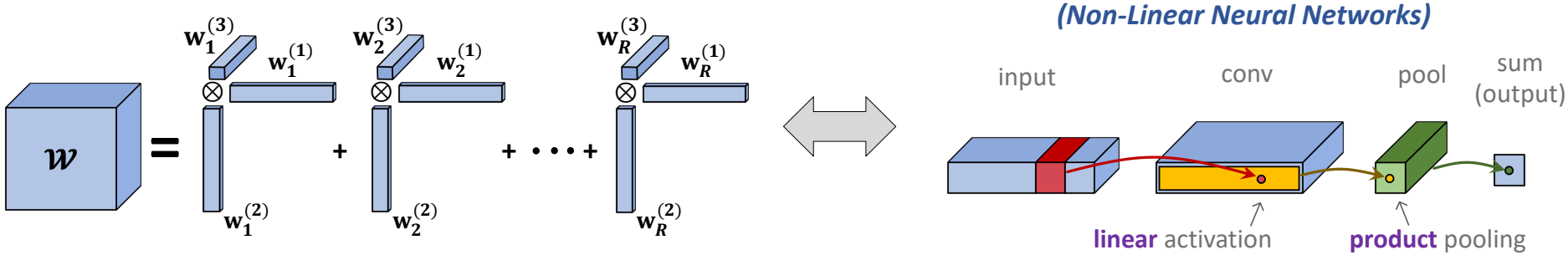
Tensor Factorization $\longleftrightarrow$ Non-Linear Neural Network

Matrix Factorizations $\longleftrightarrow$ ***Linear Neural Networks***



Tensor Factorizations $\longleftrightarrow$ ***Convolutional Arithmetic Circuits***

(Non-Linear Neural Networks)



ConvACs are competitive in practice, and admit algebraic structure

Extensively studied (e.g. Cohen et al. 2016, Cohen & Shashua 2016, Cohen & Shashua 2017)

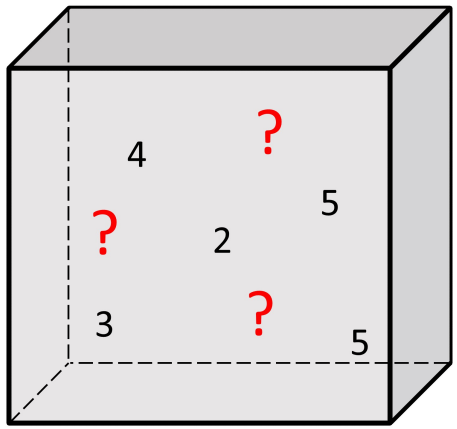
Tensor factorizations correspond to non-linear NN

Tensor Completion

Tensor completion: recover low-rank tensor given subset of entries

Tensor Completion

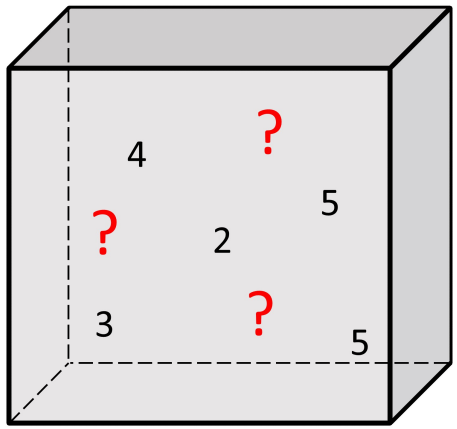
Tensor completion: recover low-rank tensor given subset of entries



Natural extension of matrix completion

Tensor Completion

Tensor completion: recover low-rank tensor given subset of entries

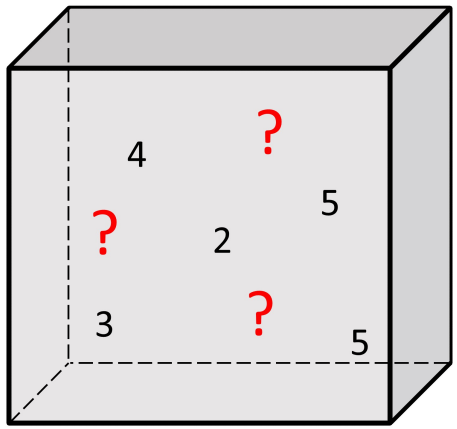


Natural extension of matrix completion

Tensor Basics

Tensor Completion

Tensor completion: recover low-rank tensor given subset of entries



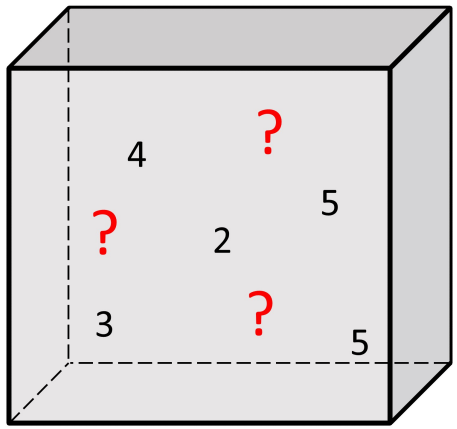
Natural extension of matrix completion

Tensor Basics

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

Tensor Completion

Tensor completion: recover low-rank tensor given subset of entries



Natural extension of matrix completion

Tensor Basics

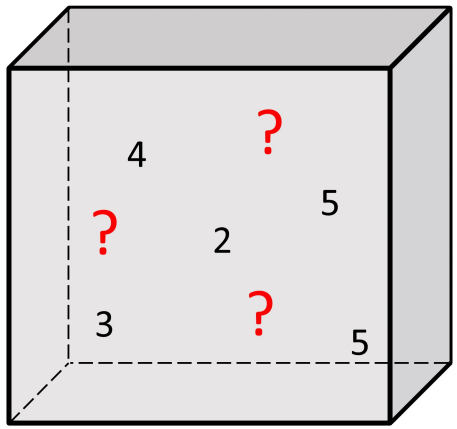
Tensor — N-dimensional array (N = order of tensor)

Tensor rank — minimal R s.t. $\mathcal{W} = \sum_{r=1}^{R} \mathbf{w}_r^{(1)} \otimes \cdots \otimes \mathbf{w}_r^{(N)}$

$\otimes :=$ outer product , $\mathbf{w}_r^{(i)} \in \mathbb{R}^{d_i}$

Tensor Completion

Tensor completion: recover low-rank tensor given subset of entries



Natural extension of matrix completion

Tensor Basics

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

Tensor rank — minimal R s.t. $\mathcal{W} = \sum_{r=1}^{R} \mathbf{w}_r^{(1)} \otimes \cdots \otimes \mathbf{w}_r^{(N)}$

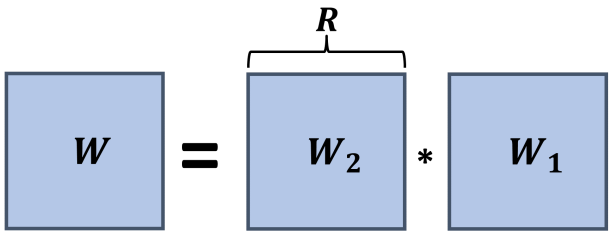
$\otimes :=$ outer product , $\mathbf{w}_r^{(i)} \in \mathbb{R}^{d_i}$

For $N = 2$ this is exactly matrix rank

From Matrix to Tensor Factorization

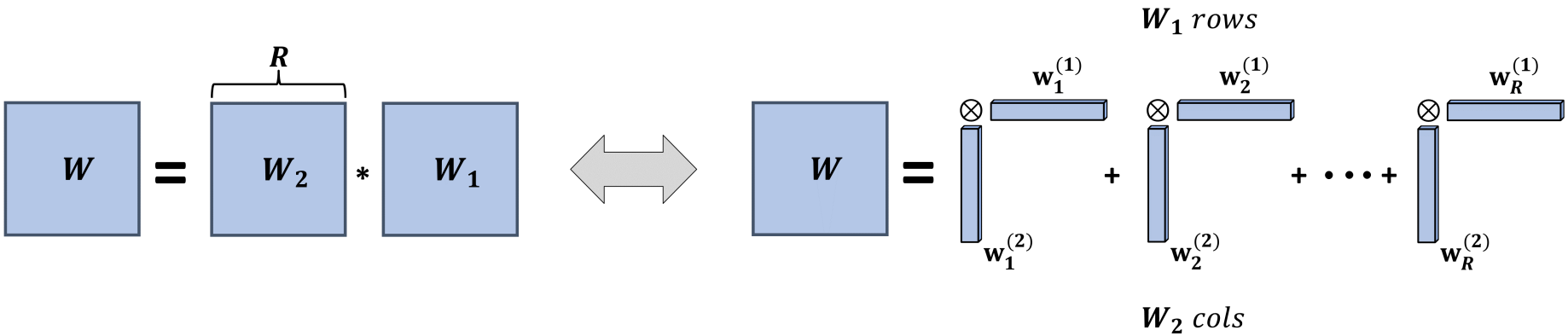
From Matrix to Tensor Factorization

Matrix Factorization



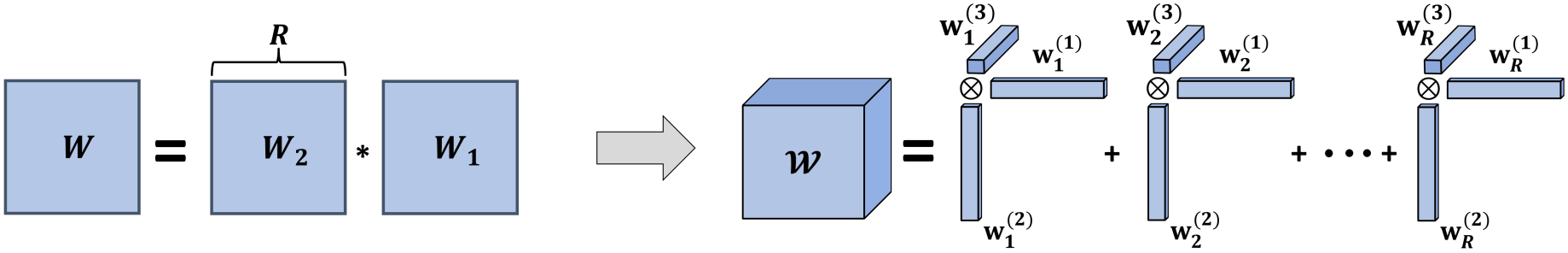
From Matrix to Tensor Factorization

Matrix Factorization



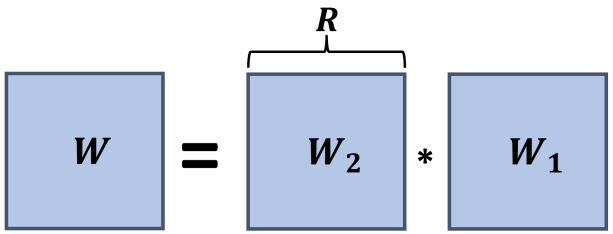
From Matrix to Tensor Factorization

Matrix Factorization

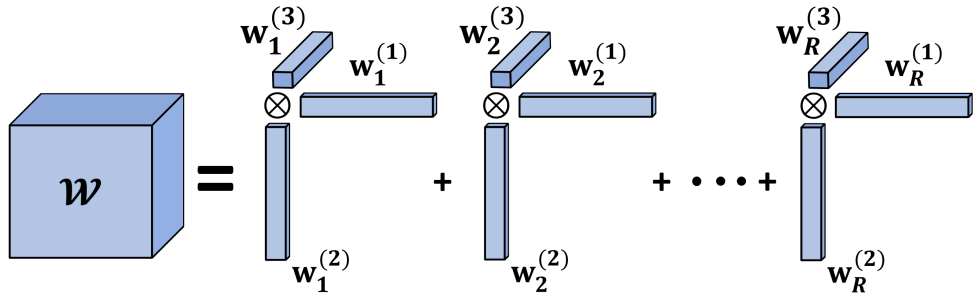


From Matrix to Tensor Factorization

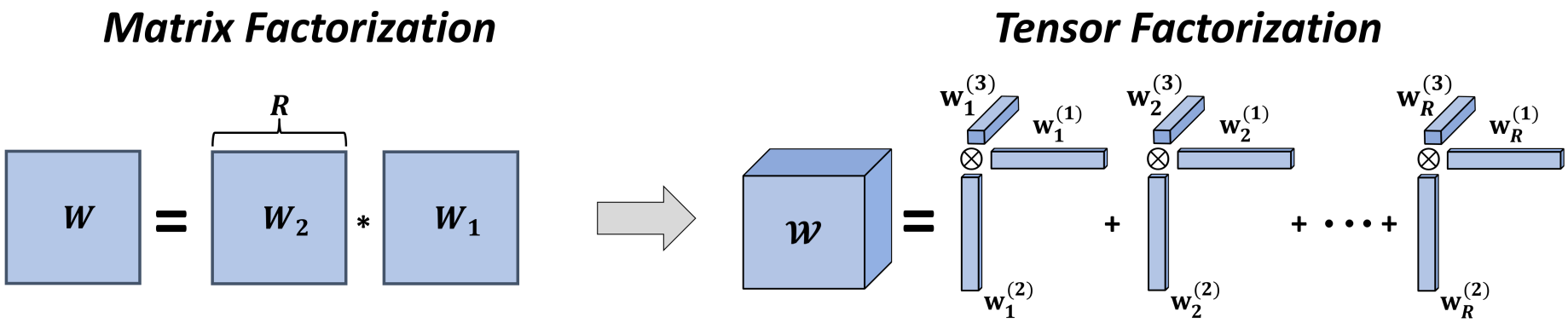
Matrix Factorization



Tensor Factorization



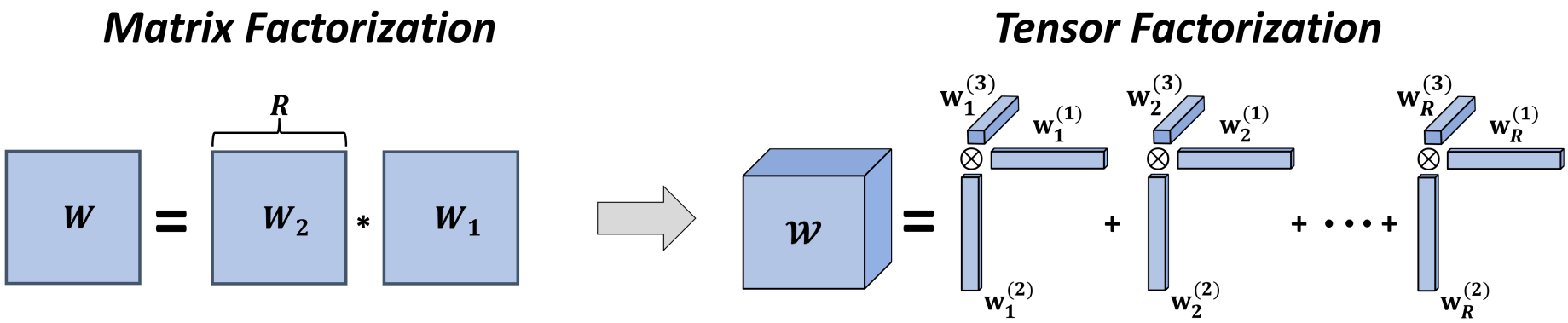
From Matrix to Tensor Factorization



Parameterize solution as **tensor factorization**:

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

From Matrix to Tensor Factorization

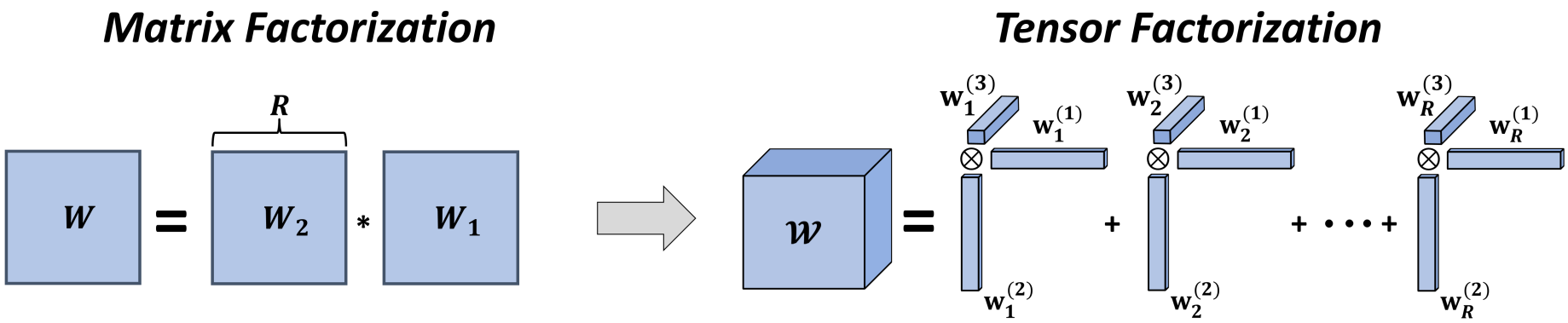


Parameterize solution as **tensor factorization**:

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

R taken large enough to not constrain rank

From Matrix to Tensor Factorization



Parameterize solution as **tensor factorization**:

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

R taken large enough to not constrain rank

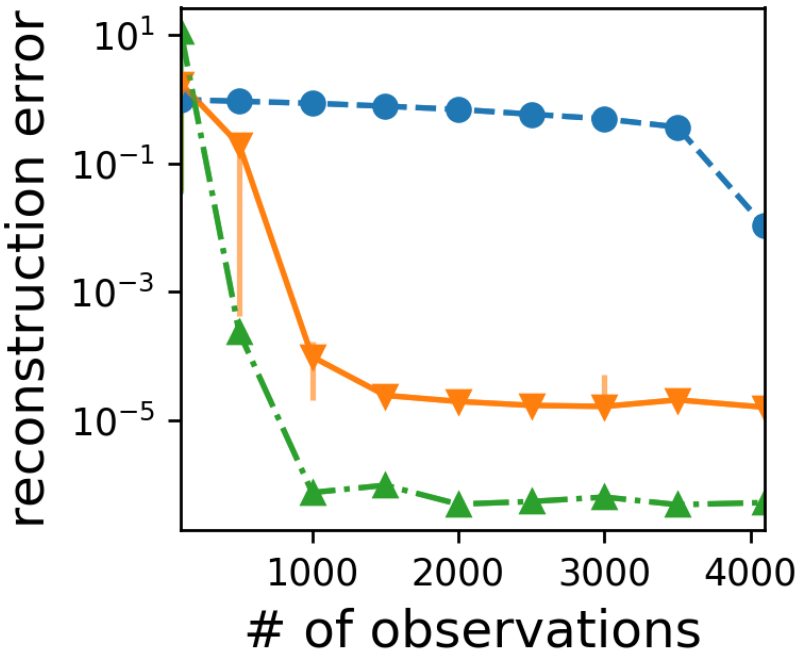
Does $\mathcal{W}$ converge to low-rank tensor when running GD w.r.t. $\left\{\mathbf{w}_r^{(n)}\right\}_{r,n}$?

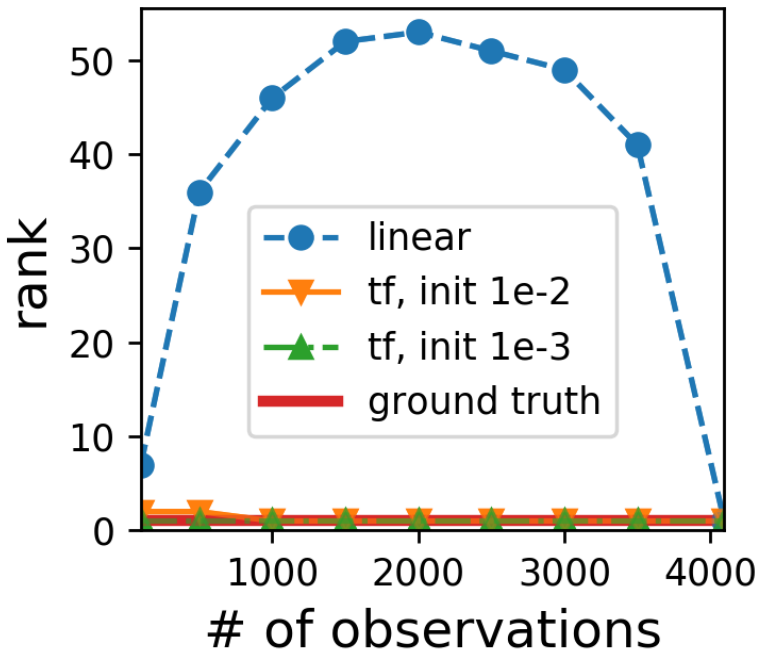
Tensor Completion Experiments

Order 4 Rank 1 Tensor Completion

Tensor Completion Experiments

Order 4 Rank 1 Tensor Completion

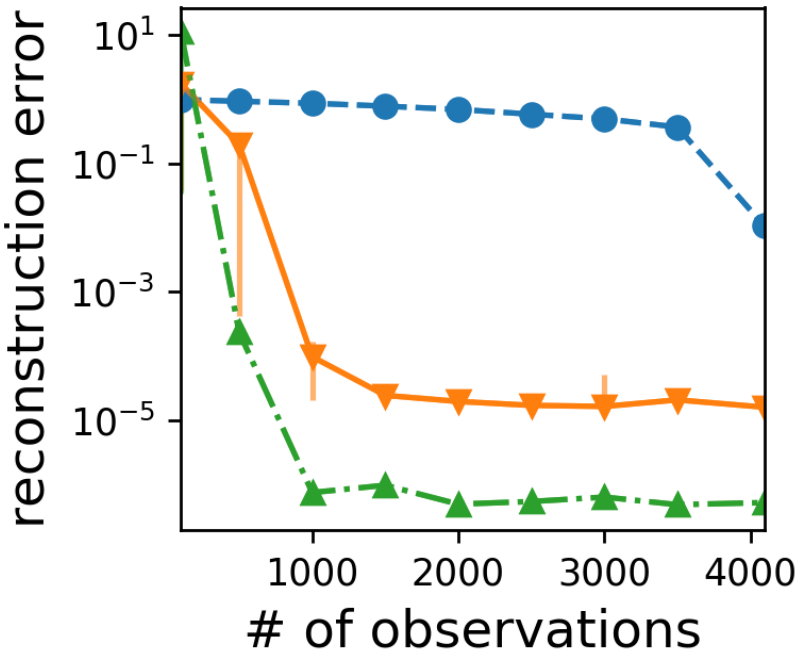


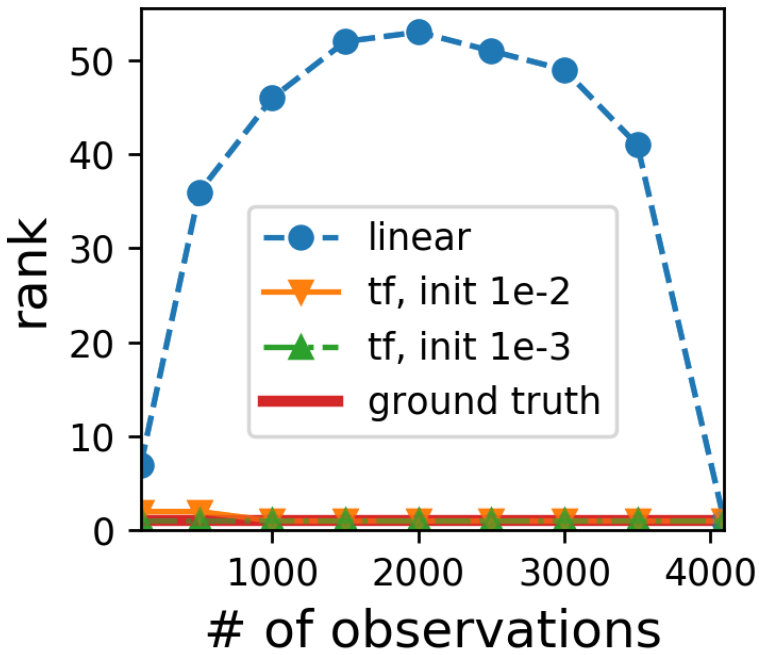


"linear" baseline — exactly fits observations, 0 elsewhere

Tensor Completion Experiments

Order 4 Rank 1 Tensor Completion





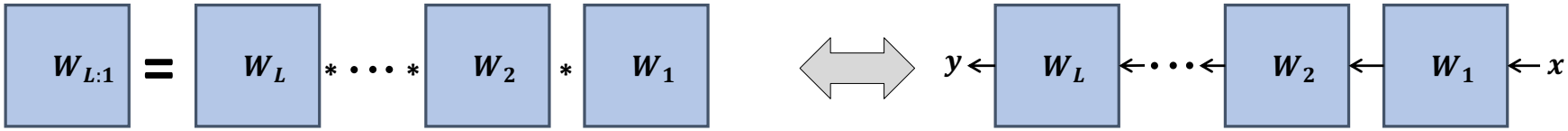
"linear" baseline — exactly fits observations, 0 elsewhere

GD drives rank of a non-linear NN towards minimum!

Implicit Rank Minimization in Deep Learning?

Implicit Rank Minimization in Deep Learning?

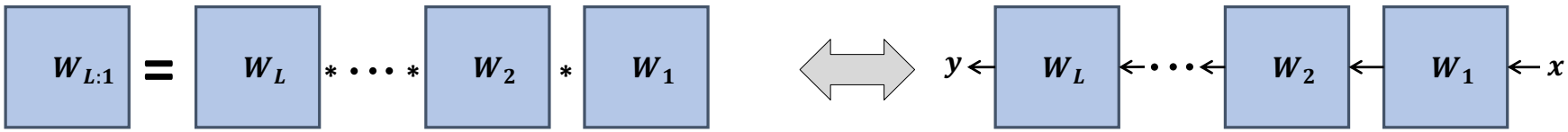
Matrix Factorizations ***Linear Neural Networks***



Theory & Experiments: implicit regularization minimizes matrix rank

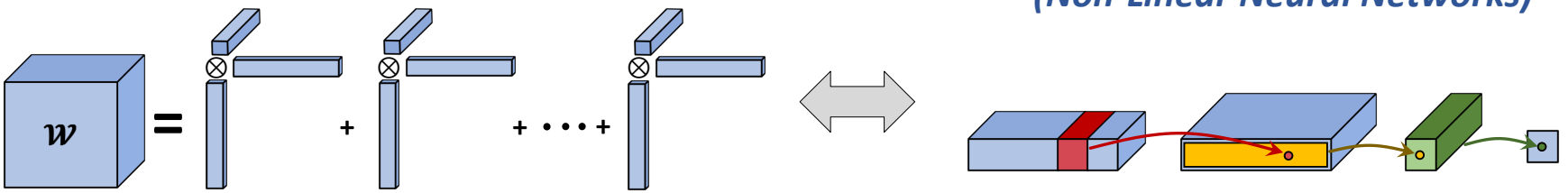
Implicit Rank Minimization in Deep Learning?

Matrix Factorizations ***Linear Neural Networks***



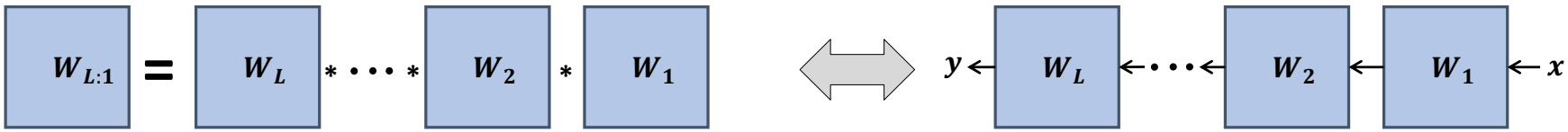
Theory & Experiments: implicit regularization minimizes matrix rank

Tensor Factorizations ***Convolutional Arithmetic Circuits***
(Non-Linear Neural Networks)



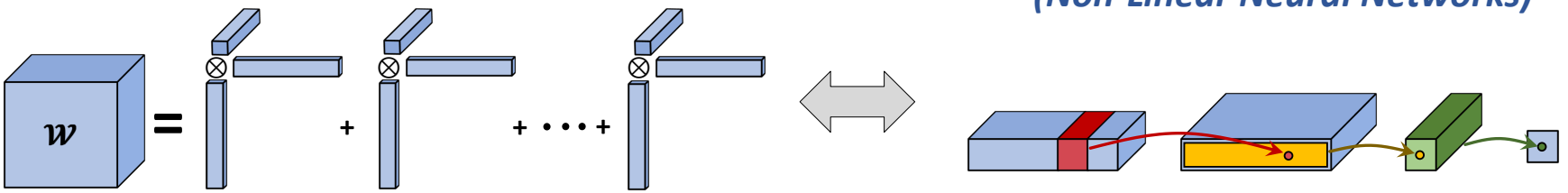
Implicit Rank Minimization in Deep Learning?

Matrix Factorizations ***Linear Neural Networks***



Theory & Experiments: implicit regularization minimizes matrix rank

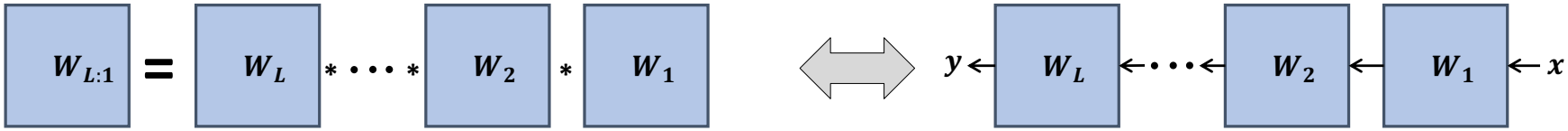
Tensor Factorizations ***Convolutional Arithmetic Circuits***
(Non-Linear Neural Networks)



Experiments: implicit regularization minimizes tensor rank

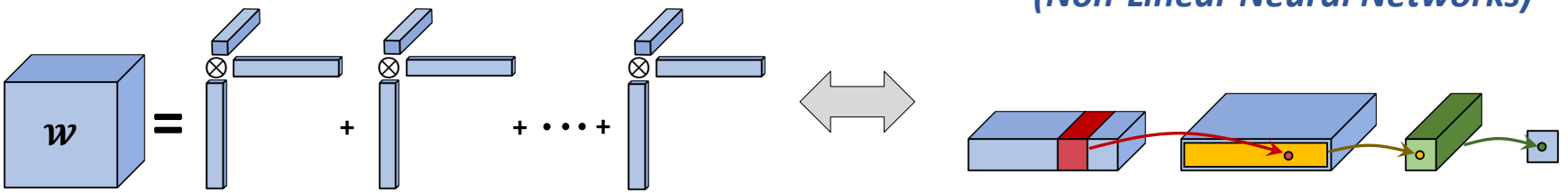
Implicit Rank Minimization in Deep Learning?

Matrix Factorizations ***Linear Neural Networks***



Theory & Experiments: implicit regularization minimizes matrix rank

Tensor Factorizations ***Convolutional Arithmetic Circuits***
(Non-Linear Neural Networks)



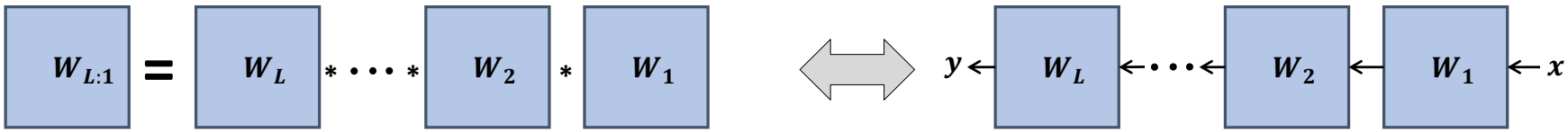
Experiments: implicit regularization minimizes tensor rank

Hypothesis

Implicit regularization in DL minimizes rank of input-output mapping

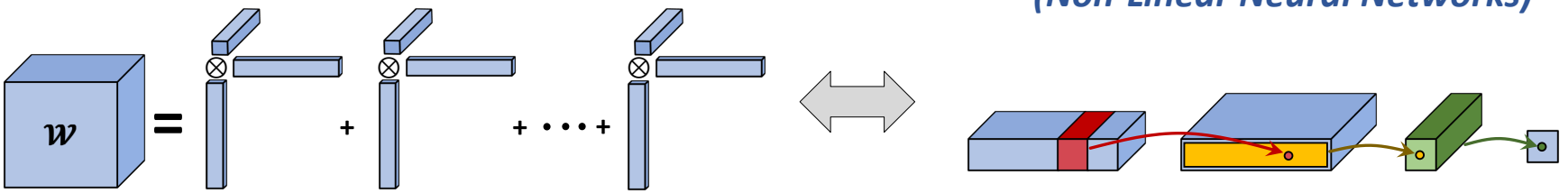
Implicit Rank Minimization in Deep Learning?

Matrix Factorizations ***Linear Neural Networks***



Theory & Experiments: implicit regularization minimizes matrix rank

Tensor Factorizations ***Convolutional Arithmetic Circuits***
(Non-Linear Neural Networks)



Experiments: implicit regularization minimizes tensor rank

Hypothesis

Implicit regularization in DL minimizes rank of input-output mapping

If true, may be key to explaining generalization

Outline

Conclusion

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**
- Unlikely implicit regularization in DL = norm minimization

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**
- Unlikely implicit regularization in DL = norm minimization

Better Interpretation: Bias to Low Rank?

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**
- Unlikely implicit regularization in DL = norm minimization

Better Interpretation: Bias to Low Rank?

- Matrix factorization: growing empirical and theoretical evidence

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**
- Unlikely implicit regularization in DL = norm minimization

Better Interpretation: Bias to Low Rank?

- Matrix factorization: growing empirical and theoretical evidence
- Extends to certain type of **non-linear NN**

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**
- Unlikely implicit regularization in DL = norm minimization

Better Interpretation: Bias to Low Rank?

- Matrix factorization: growing empirical and theoretical evidence
- Extends to certain type of **non-linear NN**

Looking Forward

Developing notions of rank for input-output mappings of NNs may be key

Conclusion

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where **all norms go to ∞**
- Unlikely implicit regularization in DL = norm minimization

Better Interpretation: Bias to Low Rank?

- Matrix factorization: growing empirical and theoretical evidence
- Extends to certain type of **non-linear NN**

Looking Forward

Developing notions of rank for input-output mappings of NNs may be key

Thank You