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

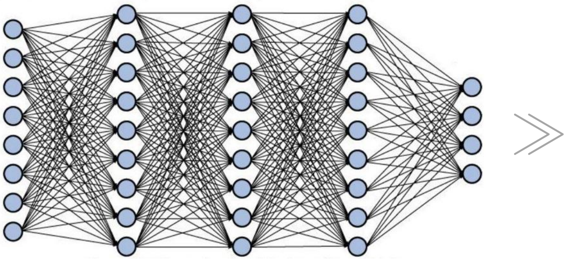
Generalize without explicit regularization:

Generalization in Deep Learning (DL)

DNNs In Practice

Generalize without explicit regularization:

1. # of learned weights   # of training examples

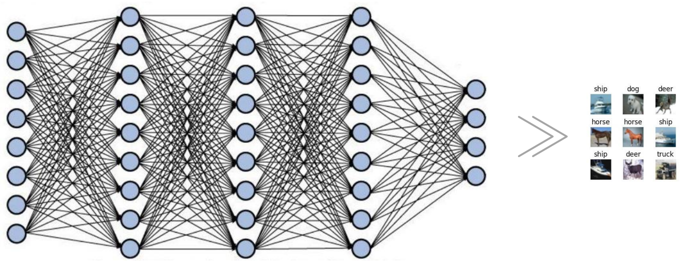


Generalization in Deep Learning (DL)

DNNs In Practice

Generalize without explicit regularization:

1. # of learned weights   # 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

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

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 e ect in concrete settings?

Linear Regression

When # parameters > # training examples:

Linear Regression

When # parameters $>$ # training examples:

GD initialized at 0 converges to min ℓ_2 norm solution

$$\underset{w}{\operatorname{argmin}} \|w\|_2 \text{ s.t. } Xw = y$$

Widespread Hope

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

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

Widespread Hope

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

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

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

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

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

observed entries ! training data

unobserved entries ! test data

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

observed entries $\rightarrow$ training data

unobserved entries $\rightarrow$ test data

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

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

observed entries $\rightarrow$ training data

unobserved entries $\rightarrow$ test data

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

Convex Programming Approach

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

observed entries $\rightarrow$ training data

unobserved entries $\rightarrow$ test data

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

Convex Programming Approach

Find minimal nuclear norm solution:

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

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

observed entries $\rightarrow$ training data

unobserved entries $\rightarrow$ test data

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

<u>Convex Programming Approach</u>

Find minimal nuclear norm solution:

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

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

Deep Learning Approach

Deep Learning Approach

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

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:

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

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?

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 (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 (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 (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 (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 (Arora et al. 2019)

For any normk k, exist observations for which small learning rate and init
can not ensure GD converges to mkink solution.

Conjecture (Arora et al. 2019)

For any normk k, exist observations for which small learning rate and init
can not ensure GD converges to mkink solution.

Arora et al. supported conjecture with:

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

Conjecture (Arora et al. 2019)

For any normk k, exist observations for which small learning rate and init
can not ensure GD converges to mkink 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 spectru

Conjecture (Arora et al. 2019)

For any normk k, exist observations for which small learning rate and ini
can not ensure GD converges to mkink 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 spectru

"

suggests tendency to low rank

Conjecture (Arora et al. 2019)

For any norm $\|\cdot\|$, exist observations for which small learning rate and ini
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 spectru

  ⇓

  suggests tendency to low rank

Open Question

Does the implicit regularization in matrix factorization minimize a norm?

Theorem (informal)

There exist matrix factorization settings where:

Theorem (informal)

There exist matrix factorization settings where:

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

Theorem (informal)

There exist matrix factorization settings where:

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

Theorem (informal)

There exist matrix factorization settings where:

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

A rms conjecture of Arora et al. 2019

Theorem (informal)

There exist matrix factorization settings where:

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

A rms conjecture of Arora et al. 2019

Result stronger than conjecture:

Theorem (informal)

There exist matrix factorization settings where:

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

A rms conjecture of Arora et al. 2019

Result stronger than conjecture:

Settings jointly disqualify all norms

Norms driven towards 1

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

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

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

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

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

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

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

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

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

Weights $W_1 \ldots W_L$ 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)

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

Gradient flow (GF) is a continuous version of GD (step size → 0):

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

Weights $W_1 \ldots W_L$ 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)

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

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

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

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

$\Uparrow$

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)

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

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

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

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

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

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

!
1
1 0

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

Min nuclear norm() = 0

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

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

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

Min nuclear norm( ) $= 0$

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

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

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

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

Min nuclear norm () = 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 () = 0

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

$$\begin{matrix} & 1 \\ 1 & 0 \end{matrix}$$

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

Min nuclear norm() $= 0$

Generally: Schatten-p (quasi-)norm $:= \sum_i \; {}_i^{p} \;\; {}^{1/p}$ (${}_i$ - singular values)

Proposition

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

Arbitrary norms (or quasi-norms):

Proposition

For minimal k k solutions, is bounded

!
1
1 0

!
1
1 0

(i) All solutions are of rank 2 (ii) j j ! 1) essentially rank ! 1

!
1
1 0

(i) All solutions are of rank 2 (ii) j j ! 1) essentiallyrank ! 1

Continuous measures for rank:

!
1
1 0

(i) All solutions are of rank 2 (ii) j j ! 1) essentially rank ! 1

Continuous measures for rank:

Distance from rank 1 := value of smallest singular value

!
1
1 0

(i) All solutions are of rank 2 (ii) j j ! 1) essentially rank ! 1

Continuous measures for rank:

Distance from rank 1 := value of smallest singular value

E ective rank (erank) entropy of singular values

$$\begin{matrix} & ! \\ & 1 \\ 1 & 0 \end{matrix}$$

(i) All solutions are of rank 2 (ii) j j ! 1) essentially rank ! 1

Continuous measures for rank:

Distance from rank 1 := value of smallest singular value

Effective rank (erank) entropy of singular values

Definition (Rank suboptimality)

rank-subopt := maxf Distance from rank 1; erank inf erankg

!

1

1 0

(i) All solutions are of rank 2 (ii) j j ! 1) essentially rank ! 1

Continuous measures for rank:

Distance from rank 1 := value of smallest singular value

E ective rank (erank) entropy of singular values

De nition (Rank suboptimality)

rank-subopt := max f Distance from rank 1; erank inf erank g

Proposition

rank-subopt is maximized when = 0, and is minimized as j j ! 1

!
1
1 0

(i) All solutions are of rank 2 (ii) j j ! 1) essentially rank ! 1

Continuous measures for rank:

Distance from rank 1 := value of smallest singular value

E ective rank (erank) entropy of singular values

De nition (Rank suboptimality)

rank-subopt := maxf Distance from rank 1; erank inf erankg

Proposition

rank-subopt is maximized when = 0, and is minimized as j j ! 1

Contradiction between norm and rank minimization

Theorem

Theorem

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

Theorem

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

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

Theorem

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

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

Theorem

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

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

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

- All norms (and quasi-norms) driven towards ∞
- Rank is essentially minimized

Theorem

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

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

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

- All norms (and quasi-norms) driven towards ∞
- Rank is essentially minimized

Claim

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

Theorem

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

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

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

- All norms (and quasi-norms) driven towards ∞
- 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!

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. $\text{rank-subopt}(W_{L:1}(t)) = \mathcal{O}(\sqrt{\ell(t)})$

Proof Sketch[1]

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

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. $\text{rank-subopt}(W_{L:1}(t)) = O(\sqrt{\ell(t)})$

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

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. $\text{rank-subopt}(W_{L:1}(t)) = O(\sqrt{\ell(t)})$

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

Theorem

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

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

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

Theorem

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

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

Proof Sketch

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

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

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

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

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. $\text{rank-subopt}(W_{L:1}(t)) = O(\sqrt{\ell(t)})$

<u>Proof Sketch</u>

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

Customary

Separating aspects of convergence to global min and implicit regularizati

Customary

Separating aspects of convergence to global min and implicit regularizati

"

commonly observed in practice

<u>Customary</u>

Separating aspects of convergence to global min and implicit regularizati

"

commonly observed in practice

<u>Convergence to Zero Loss (in our setting)</u>

Customary

Separating aspects of convergence to global min and implicit regularizati

"

commonly observed in practice

Convergence to Zero Loss (in our setting)

Experiments: GD consistently nds global min

Customary

Separating aspects of convergence to global min and implicit regularizati

$\Downarrow$

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

Customary

Separating aspects of convergence to global min and implicit regularizati

$\Downarrow$

commonly observed in practice

Convergence to Zero Loss (in our setting)

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

Proposition

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

Customary

Separating aspects of convergence to global min and implicit regularizati

$\Downarrow$

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 $W_{L:1}(0) = \quad I$, for depth $L = 2$ and $0 < \quad 1$, then $\ell(t) \to 0$

Proof Approach

Careful analysis of GF differential equations

What happens when observations are perturbed?

!
1
1 0

What happens when observations are perturbed?

$$\begin{matrix} & 1 \\ 1 & 0 \end{matrix} \implies \begin{matrix} & z \\ z' & \end{matrix} \qquad \text{non-zero } z; z' \text{ arbitrary}$$

What happens when observations are perturbed?

$$\begin{pmatrix} & \overset{\to}{1} \\ 1 & 0 \end{pmatrix} \implies \begin{pmatrix} & \vec{z} \\ z' & \end{pmatrix} \quad \begin{matrix} \text{non-zero } z;\, z' \\ \text{arbitrary} \end{matrix}$$

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. $\text{rank-subopt}(W_{L:1}(t)) = O(\sqrt{\ell(t)})$

What happens when observations are perturbed?

$$\begin{pmatrix} * & 1 \\ 1 & 0 \end{pmatrix} \implies \begin{pmatrix} * & z \\ z' & * \end{pmatrix} \quad \text{non-zero } z, z' \text{ arbitrary}$$

Theorem (original setting)

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

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

What happens when observations are perturbed?

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

Theorem (intermediate)

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

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

What happens when observations are perturbed?

$$\begin{pmatrix} & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} & z \\ z' & \end{pmatrix} \quad \text{non-zero } z, z' \text{ arbitrary}$$

Theorem (intermediate)

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

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

What happens when observations are perturbed?

$$\begin{pmatrix} & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} & z \\ z' & \end{pmatrix} \quad \text{non-zero } z, z' \text{ arbitrary}$$

Theorem

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

1. $\|W_{L:1}(t)\| = \quad \min\{|z|, |z'|\} = (| \; | + \sqrt{2\ell(t)})$
2. $\text{rank-subopt}(W_{L:1}(t)) = O(| \; | + \sqrt{2\ell(t)}) = \min\{|z|, |z'|\}$

What happens when observations are perturbed?

$$\begin{pmatrix} * & \overset{\rightarrow}{1} \\ 1 & 0 \end{pmatrix} \implies \begin{pmatrix} * & \vec{z} \\ z' & \end{pmatrix} \quad \text{non-zero } z, z' \text{ arbitrary}$$

Theorem

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

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

$\epsilon = 0 \implies$ all norms driven to ∞ and rank is minimized

What happens when observations are perturbed?

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

Theorem

If $\mathrm{sign}(\det(W_{L:1}(0))) = -\,\mathrm{sign}(z \cdot z')$ 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. $\text{rank-subopt}(W_{L:1}(t)) = O\left((|\epsilon| + \sqrt{2\ell(t)}) \,/\, \min\{|z|, |z'|\} \right)$

$\epsilon = 0 \Longrightarrow$ all norms driven to ∞ and rank is minimized

Phenomenon gracefully recedes as ϵ perturbed from 0

What happens when observations are perturbed?

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

Theorem

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

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

$\epsilon = 0 \implies$ all norms driven to ∞ and rank is minimized

Phenomenon gracefully recedes as ϵ perturbed from 0

Same results hold when changing unobserved entry location

!
1
1 0

!
1
1 0

!
1
1 0

!
2
0:5

!
1
1 0

!
2
0:5

!
1
1 0

!
2
0:5

Theory transfers to practice: unobserved entry ! 1

Analyzed Setting (our work)

Analyzed Setting (our work)

Contrast between norm and rank minimization

Analyzed Setting (our work)

Contrast between norm and rank minimization

Implicit regularization drives norms to 1 to minimize rank

Analyzed Setting (our work)

Contrast between norm and rank minimization

Implicit regularization drives norms to1 to minimize rank

Past Work

Analyzed Setting (our work)

Contrast between norm and rank minimization

Implicit regularization drives norms to 1 to minimize rank

Past Work

Empirical evidence: low-rank tendency in matrix factorization

Analyzed Setting (our work)

- Contrast between norm and rank minimization
- Implicit regularization drives norms t01 to minimize rank

Past Work

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

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 $\neq$ rank minimization?

Analyzed Setting (our work)

Contrast between norm and rank minimization

Implicit regularization drives norms to1 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?

ConvACs are competitive in practice, and admit algebraic structure

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

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 : recover low-rank tensor given subset of entries

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

Natural extension of matrix completion

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

Natural extension of matrix completion

Tensor Basics

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 : recover low-rank tensor given subset of entries

Natural extension of matrix completion

Tensor Basics

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

Tensor rank $\rightarrow$ minimal R s.t. $W = \sum_{r=1}^{R} w_r^{(1)} \otimes \cdots \otimes w_r^{(N)}$

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

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. $W = \sum_{r=1}^{R} w_r^{(1)} \qquad w_r^{(N)}$

:= outer product , $w_r^{(i)} \in \mathbb{R}^{d_i}$

For $N = 2$ this is exactly matrix rank

Parameterize solution as tensor factorization :

$$W = \sum_{r=1}^{R} w_r^{(1)} \qquad\qquad w_r^{(N)}$$

Parameterize solution as tensor factorization :

$$W = \sum_{r=1}^{R} w_r^{(1)} \qquad\qquad w_r^{(N)}$$

R taken large enough to not constrain rank

Parameterize solution as **tensor factorization**:

$$W = \sum_{r=1}^{R} w_r^{(1)} \otimes \cdots \otimes w_r^{(N)}$$

R taken large enough to not constrain rank

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

Order 4 Rank 1 Tensor Completion

Order 4 Rank 1 Tensor Completion

"linear" baseline exactly ts observations, 0 elsewhere

Order 4 Rank 1 Tensor Completion

"linear" baseline exactly ts observations, 0 elsewhere

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

<u>Theory & Experiments</u>: implicit regularization minimizes matrix rank

Theory & Experiments: implicit regularization minimizes matrix rank

Theory & Experiments: implicit regularization minimizes matrix rank

Experiments: implicit regularization minimizes tensor rank

Theory & Experiments: implicit regularization minimizes matrix rank

Experiments: implicit regularization minimizes tensor rank

Hypothesis
Implicit regularization in DL minimizes rank of input-output mapping

Theory & Experiments: implicit regularization minimizes matrix rank

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

Implicit Regularization $\neq$ Norm Minimization

Matrix factorization: exist cases where all norms go to ∞

Implicit Regularization $\neq$ Norm Minimization

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

Implicit Regularization $\neq$ Norm Minimization

- Matrix factorization: exist cases where all norms go to 1
- Unlikely implicit regularization in DL = norm minimization

Better Interpretation: Bias to Low Rank?

Conclusion

Implicit Regularization = 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 = 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 = 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 = 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