On the Convergence and Implicit Bias of Overparametrized Linear Networks

Hancheng Min, Salma Tarmoun, René Vidal and Enrique Mallada

2022 MINDS Retreat
Jan. 18th – 21st
Introduction

• In deep learning, neural networks are typically overparametrized
  (# of parameters) \gg (# of training examples)
  ▪ Highly underdetermined problem, many solutions
  ▪ Variants of gradient descent often finds those with good generalization
• Theoretically understand the nonlinear training dynamics of gradient methods
• Prior works suggest that in this overparametrized regime, specific initialization may:
  ▪ Accelerate convergence (implicit acceleration)
  ▪ Promote generalization (implicit bias)

• Question: Are there general properties of initialization that benefit convergence and implicit bias?
• Our setting: two-layer linear networks, gradient flow, the answer is YES!
Outline

• *(Convergence)* **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence

• *(Implicit Bias)* **Orthogonal initialization** leads to min-norm solution
**Contributions: Convergence**

- Existing analysis for convergence of two-layer linear networks requires **strong assumptions on the initialization (balanced, or spectral)**

<table>
<thead>
<tr>
<th></th>
<th>Spectral</th>
<th>Non-spectral (with sufficient margin)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balanced</td>
<td>[Saxes’14]</td>
<td>[Arora’18]</td>
</tr>
<tr>
<td></td>
<td>[Gidel’19]</td>
<td></td>
</tr>
<tr>
<td>Sufficiently Imbalanced</td>
<td>[Tarmoun’21]</td>
<td><strong>Our work</strong></td>
</tr>
</tbody>
</table>

S Tarmoun, G França, B D Haeffele, and R Vidal. “Understanding the dynamics of gradient flow in overparameterized linear models.” ICML 2021
Contributions: Convergence

• Existing analysis for convergence of two-layer linear networks requires **strong** assumptions on the initialization (balanced, or spectral)

<table>
<thead>
<tr>
<th></th>
<th>Spectral</th>
<th>Non-spectral (with sufficient <strong>margin</strong>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balanced</td>
<td>[Saxes’14]</td>
<td>[Arora’18]</td>
</tr>
<tr>
<td></td>
<td>[Gidel’19]</td>
<td></td>
</tr>
<tr>
<td>Sufficiently Imbalanced</td>
<td>[Tarmoun’21]</td>
<td><strong>Our work</strong></td>
</tr>
</tbody>
</table>

• We show

\[
\text{Rate} \geq \sqrt{(\text{Imbalance})^2 + 4(Margin)^2}
\]

• **Exponential convergence** via **sufficient imbalance** or **sufficient margin**
Problem Setup

- Training data \( X = [x^{(1)} \ldots x^{(P)}]^T \in \mathbb{R}^{P \times n}, Y = [y^{(1)} \ldots y^{(P)}]^T \in \mathbb{R}^{P \times m} \)
- Two-layer linear network, squared loss (Regression task)
  \[
  L(U, V) = \frac{1}{2} \| Y - XUV^T \|_F^2, \quad U \in \mathbb{R}^{n \times h}, V \in \mathbb{R}^{m \times h}
  \]
- Overparametrized model: \( h \geq min\{n, m\} \)
- Gradient flow dynamics
  \[
  \dot{U} = -\frac{\partial L}{\partial U}, \quad \dot{V} = -\frac{\partial L}{\partial V}
  \]
Outline - Convergence

• *Convergence* **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence

  ▪ *(Warm-up) Scalar case*:
    \[ L_s(u, v) = \frac{1}{2} |y - uv|^2 \]

  ▪ **Matrix case**:
    \[ L(U, V) = \frac{1}{2} \|Y - UV^T\|_F^2 \]

  ▪ **Convergence results for regression**:
    \[ L(U, V) = \frac{1}{2} \|Y - XUV^T\|_F^2 \]
Scalar Dynamics: Imbalance

- Gradient flow on $L_s(u, v) = \frac{1}{2} |y - uv|^2$
  \[
  \dot{u} = (y - uv)v \\
  \dot{v} = (y - uv)u
  \]

- Imbalance of the weights
  \[d := u^2 - v^2\]

- Imbalance is time-invariant [Saxes’14]
  \[\dot{d} = 0\]

---

Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u, v) = \frac{1}{2} |y - uv|^2$
  \[
  \dot{u} = (y - uv)v, \quad \dot{v} = (y - uv)u
  \]
- We need a lower bound on the instantaneous rate $-\dot{L}_s / L_s$

Grönwall’s inequality
\[
\dot{L}_s(t) \leq -\alpha L_s(t)
\]
\[
\Rightarrow L_s(t) \leq \exp(-\alpha t) L_s(0)
\]

For **exponential convergence**, show
\[
-\dot{L}_s / L_s \geq \alpha
\]
for some $\alpha > 0$
Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u, v) = \frac{1}{2} |y - uv|^2$
  \[
  \dot{u} = (y - uv)v, \quad \dot{v} = (y - uv)u
  \]
- We need a lower bound on the instantaneous rate $-\dot{L}_s/L_s$
- $-\frac{\dot{L}_s}{L_s} = 2(u^2 + v^2)$
  - $d$ is time-invariant
  - A lower bound on $(uv)^2$ ??

Express $u^2, v^2$ by imbalance $d$ and product $uv$

\[
\begin{align*}
  u^2 &= \frac{d + \sqrt{d^2 + 4(uv)^2}}{2} \\
  v^2 &= \frac{-d + \sqrt{d^2 + 4(uv)^2}}{2}
\end{align*}
\]
Scalar Dynamics: Exponential Convergence

• Gradient flow on $L_s(u, v) = \frac{1}{2}|y - uv|^2$
  \[
  \dot{u} = (y - uv)v, \quad \dot{v} = (y - uv)u
  \]

• We need a lower bound on the instantaneous rate $-\dot{L}_s/L_s$
  \[
  -\frac{\dot{L}_s}{L_s} = 2(u^2 + v^2) = 2\sqrt{d^2 + 4(uv)^2}
  \]
  - $d$ is time-invariant ✔
  - $(uv)^2 \geq (\text{Margin})^2$ ✔

$|uv|$ stays above the margin

\[
|u(t)v(t)| \geq |y| - |y - u(t)v(t)| \\
\geq |y| - |y - u(0)v(0)| \\
:= \text{Margin}
\]
Scalar Dynamics: Exponential Convergence

- Gradient flow on $L_s(u, v) = \frac{1}{2} |y - uv|^2$
  $$\dot{u} = (y - uv)v, \quad \dot{v} = (y - uv)u$$
- We need a lower bound on the instantaneous rate $-\dot{L}_s/L_s$
- $$-\frac{\dot{L}_s}{L_s} = 2(u^2 + v^2) = 2\sqrt{d^2 + 4(uv)^2}$$
  - $d$ is time-invariant
  - $(uv)^2 \geq (Margin)^2$
- $$-\frac{\dot{L}_s(t)}{L_s(t)} = 2 \sqrt{d^2 + 4(u(t)v(t))^2} \geq 2\sqrt{d^2 + 4(max\{|y| - |y - u(0)v(0)|, 0\})^2}$$

Rate $\geq 2\sqrt{(Imbalance)^2 + 4(Margin)^2}$
From Scalar Case to Matrix Case

Scalar Case

• \( L_s(u, v) = \frac{1}{2} |y - uv|^2 \)

• Imbalance \( d = u^2 - v^2 \)

• Rate depends on imbalance \( d \) and product \( uv \)

• \( L_s \) converges exponentially via
  • Sufficient imbalance
  • Sufficient margin

Matrix Case

• \( L(U, V) = \frac{1}{2} \| Y - UV^T \|_F^2 \)
  \((U \in \mathbb{R}^{n \times h}, V \in \mathbb{R}^{h \times m}, h \geq \min\{n, m\})\)

• Imbalance \( D = U^T U - V^T V \)

• Rate depends on imbalance quantities \( \Delta, \Delta_+, \Delta_- \) and product \( UV^T \)

• \( L \) converges exponentially via
  • Sufficient level of imbalance \( \Delta \)
  • Sufficient margin
Imbalance quantities

- \( L(U,V) = \frac{1}{2} \| Y - UV^T \|_F^2 \) \( (U \in \mathbb{R}^{n \times h}, V \in \mathbb{R}^{h \times m}) \)
- Imbalance \( D = U^T U - V^T V \)
Imbalance quantities

- \( L(U, V) = \frac{1}{2} \| Y - UV^T \|_F^2 \) (\( U \in \mathbb{R}^{n \times h}, V \in \mathbb{R}^{h \times m} \))
- Imbalance \( D = U^T U - V^T V \)
Main Results: Instantaneous Rate

- For the scalar case
  \[ \text{Rate} = 2\sqrt{(\text{Imbalance})^2+4(\text{Product})^2} \]

- For the matrix case
  \[ \text{Rate} \geq -\text{Spread} + \sqrt{(\text{lvl. of imbalance} + \text{Spread})^2+4\sigma^2(\text{Product})} \]

\textbf{Proposition 1.} (Lower bound on instantaneous rate) Define \( D = U^TU - V^TV \).
Consider the gradient flow on \( L(U, V) = \frac{1}{2} \|Y - UV^T\|_F^2 \). Then we have
\[ -\frac{\dot{L}}{L} \geq -\Delta_+ + \sqrt{(\Delta_+ + \Delta)^2 + 4\sigma_m^2(UV^T) - \Delta_- + \sqrt{(\Delta_- + \Delta)^2 + 4\sigma_n^2(UV^T)},} \]
Main Results: Exponential Convergence

• We have a lower bound on the instantaneous rate

\[ \text{Rate} \geq -\text{Spread} + \sqrt{(\text{lvl. of imbalance} + \text{Spread})^2 + 4\sigma^2(\text{Product})} \]

- Imbalance is time-invariant ✔
- \(\sigma^2(\text{Product}) \geq (\text{Margin})^2\) ✔

• \(L\) converges exponentially via
  
  - **Sufficient level of imbalance**
  
    \[ \Delta > 0 \]

  - **Sufficient margin**

    \[ \sigma_{\text{min}}(Y) - \|Y - UV^T\|_F > 0 \]
Exponential Convergence Guarantees: Summary

**Non-spectral** initializations for the gradient flow on $\frac{1}{2} \| Y - UV^T \|_F^2$

<table>
<thead>
<tr>
<th>Balanced initialization</th>
<th>$D: = U^T U - V^T V = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Margin + approx. balanced [Arora'18]</td>
<td>$\sigma_{\min}(Y) - | Y - UV^T |_F &gt; \delta$</td>
</tr>
<tr>
<td>$| D |_F \leq C \delta^2$</td>
<td></td>
</tr>
<tr>
<td>Homogeneous imbalance [Tarmoun'21]</td>
<td>$D = \lambda_0 I_h$, $</td>
</tr>
<tr>
<td>Sufficient level of imbalance [Min'21]</td>
<td>$\Delta &gt; 0$</td>
</tr>
<tr>
<td>Sufficient margin</td>
<td>$\sigma_{\min}(Y) - | Y - UV^T |_F &gt; 0$</td>
</tr>
</tbody>
</table>
Convergence Result for Linear Regression

• For matrix factorization $L(U, V) = \frac{1}{2} \|Y - UV^T\|_F^2$, we have

$$Rate \geq \sqrt{(Imbalance)^2 + 4(Margin)^2}$$

• For linear regression $\tilde{L}(U, V) = \frac{1}{2} \|Y - XUV^T\|_F^2$, we have ($\Sigma_x = X^TX$)

$$Rate \geq \lambda_{\text{min}}(\Sigma_x) \sqrt{(Imbalance)^2 + 4(Margin)^2} / \lambda_{\text{max}}(\Sigma_x)$$
Outline

• *(Convergence)* **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence

• *(Implicit Bias)* **Orthogonal initialization** leads to min-norm solution
Implicit Bias to Min-norm Solution

• Suppose \( X \in \mathbb{R}^{P \times n} \) \textbf{DOES NOT} have full row rank, \( r = \text{rank}(X) < n \)

• (Underdetermined) Infinitely many solutions to \( \min_\Theta \|Y - X \Theta\|_F \)

• The \textit{minimum-norm solution}
  \[
  \hat{\Theta} = \arg\min_\Theta \{ \|\Theta\|_F : \|Y - X \Theta\|_F = \min_\Theta \|Y - X \Theta\|_F \}
  \]

• We decompose the weight \( U \) using the SVD of \( X \)
  \[
  U = \Phi_1 \Phi_1^T U + \Phi_2 \Phi_2^T U, \quad X = W \left[ \Sigma_x^{1/2} \ 0 \right] \left[ \Phi_1^T \Phi_2^T \right]
  \]

• “orthogonality” among \( U_1, U_2, V \) \( \Rightarrow \) exact minimum-norm solution
Main Results: Implicit Bias to Min-norm Solution

**Proposition 2. (Orthogonal Initialization)** If one have
\[
V(0)U_2^T(0) = 0, \quad U_1(0)U_2^T(0) = 0,
\]
and that the loss converges to a global minimum, then \(U(t)V^T(t)\) converges to exactly the minimum-norm solution \(\hat{\Theta}\)

- Orthogonal initialization may not converge (e.g., zero initialization).
- Sufficient imbalance or margin can provide convergence guarantee.
Random Initialization + Large Width

Random initialization
\[[U(0)]_{ij}, [V(0)]_{ij} \sim \mathcal{N}(0, h^{-1})\]
Large hidden layer width $h$

**exact minimum-norm solution**

- *(Sufficient level of imbalance)*
  \[\Delta(0) > 0\]
- *(Orthogonality)*
  \[
  \left\| \begin{bmatrix} V(0)U_T(0) \\ U_1(0)U_T(0) \end{bmatrix} \right\|_F = 0
  \]

**Non-kernel-regime conditions**

**approximate minimum-norm solution**

- *(Sufficient level of imbalance) w.h.p.*
  \[\Delta(0) > 1\]
- *(Approximate Orthogonality) w.h.p.*
  \[
  \left\| \begin{bmatrix} V(0)U_T(0) \\ U_1(0)U_T(0) \end{bmatrix} \right\|_F = O(h^{-1/2})
  \]

**Initialization in the kernel regime**
Conclusion

We study the gradient flow on two-layer linear networks:

• **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence
• **Orthogonal Initialization** leads to min-norm solution

Future work:

• Convergence analysis extends to other losses
• Deep linear networks
• Imbalance in nonlinear networks (ReLU net, etc.)

Thank you!


S Tarmoun, G França, B D Haeffele, and R Vidal. “Understanding the dynamics of gradient flow in overparameterized linear models.” ICML 2021

