On the Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks

Hancheng Min, Salma Tarmoun, René Vidal and Enrique Mallada

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Problem Setup

- Training data $X = [x^{(1)} \quad \cdots \quad x^{(n)}]^T \in \mathbb{R}^{n\times D}, Y = [y^{(1)} \quad \cdots \quad y^{(n)}]^T \in \mathbb{R}^{n\times m}$
- Single-hidden layer linear network, squared loss

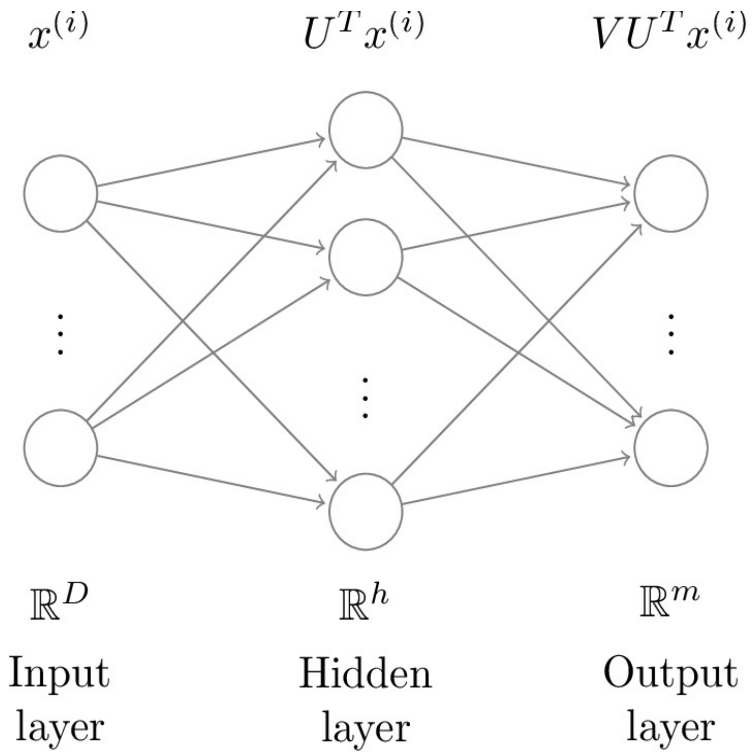
$$L(U,V) = \frac{1}{2}\|Y - XUV^T\|_F^2, \qquad U \in \mathbb{R}^{D\times h}, V \in \mathbb{R}^{m\times h}$$

  - Underdetermined linear regression: $D > n$
  - Overparametrized model: $h \geq min\{D, m\}$

- Gradient flow dynamics

$$\dot{U} = -\frac{\partial L}{\partial U} = (Y - XUV^T)V^T$$

$$\dot{V} = -\frac{\partial L}{\partial V} = (Y - XUV^T)^T U$$



Problem Setup

- Suppose $rank(X) = r$, we decompose the weight U using the SVD of X

$$U = \Phi_1 \overbrace{\Phi_1^T U}^{:=U_1} + \Phi_2 \overbrace{\Phi_2^T U}^{:=U_2}, \qquad X = W \begin{bmatrix} \Sigma_x^{1/2} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \end{bmatrix}$$

- We have $U_1 \in \mathbb{R}^{r \times h}$, $U_2 \in \mathbb{R}^{m \times h}$, and

$$\dot{U}_1 = \Sigma_x^{1/2} \left(W^T Y - \Sigma_x^{1/2} U_1 V^T \right) V^T, \qquad \dot{U}_2 = 0,$$

$$\dot{V} = \left(W^T Y - \Sigma_x^{1/2} U_1 V^T \right)^T \Sigma_x^{1/2} U$$

- For **convergence**, it suffices to study the flow of U_1 and V, which is exactly the gradient flow dynamics on $\tilde{L}(U_1, V) = \frac{1}{2} \left\| W^T Y - \Sigma_x^{1/2} U_1 V^T \right\|_F^2$

Problem Setup

- Suppose $rank(X) = r$, we decompose the weight U using the SVD of X

$$U = \Phi_1 \overbrace{\Phi_1^T U}^{:=U_1} + \Phi_2 \overbrace{\Phi_2^T U}^{:=U_2}, \qquad X = W\begin{bmatrix} \Sigma_x^{1/2} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \end{bmatrix}$$

- We have $U_1 \in \mathbb{R}^{r \times h}$, $U_2 \in \mathbb{R}^{m \times h}$, and

$$\dot{U}_1 = \Sigma_x^{1/2} \left(W^T Y - \Sigma_x^{1/2} U_1 V^T \right) V^T, \qquad \dot{U}_2 = 0,$$

$$\dot{V} = \left(W^T Y - \Sigma_x^{1/2} U_1 V^T \right)^T \Sigma_x^{1/2} U$$

- We also study the **implicit bias towards the min-norm solution**

$$\widehat{\Theta} = \min_{\Theta} \left\{ \|\Theta\|_F : \ \|Y - X\Theta\|_F = \min_{\Theta} \|Y - X\Theta\|_F \right\}$$

Overview

- **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence

- **Orthogonal initialization** leads to min-norm solution

- **Random initialization** + **large network width** approximately satisfies the two conditions above, allowing us to find near minimum norm solution efficiently

Overview

- **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence
- **Orthogonal initialization** leads to min-norm solution
- **Random initialization** + **large network width** approximately satisfies the two conditions above, allowing us to find near minimum norm solution efficiently

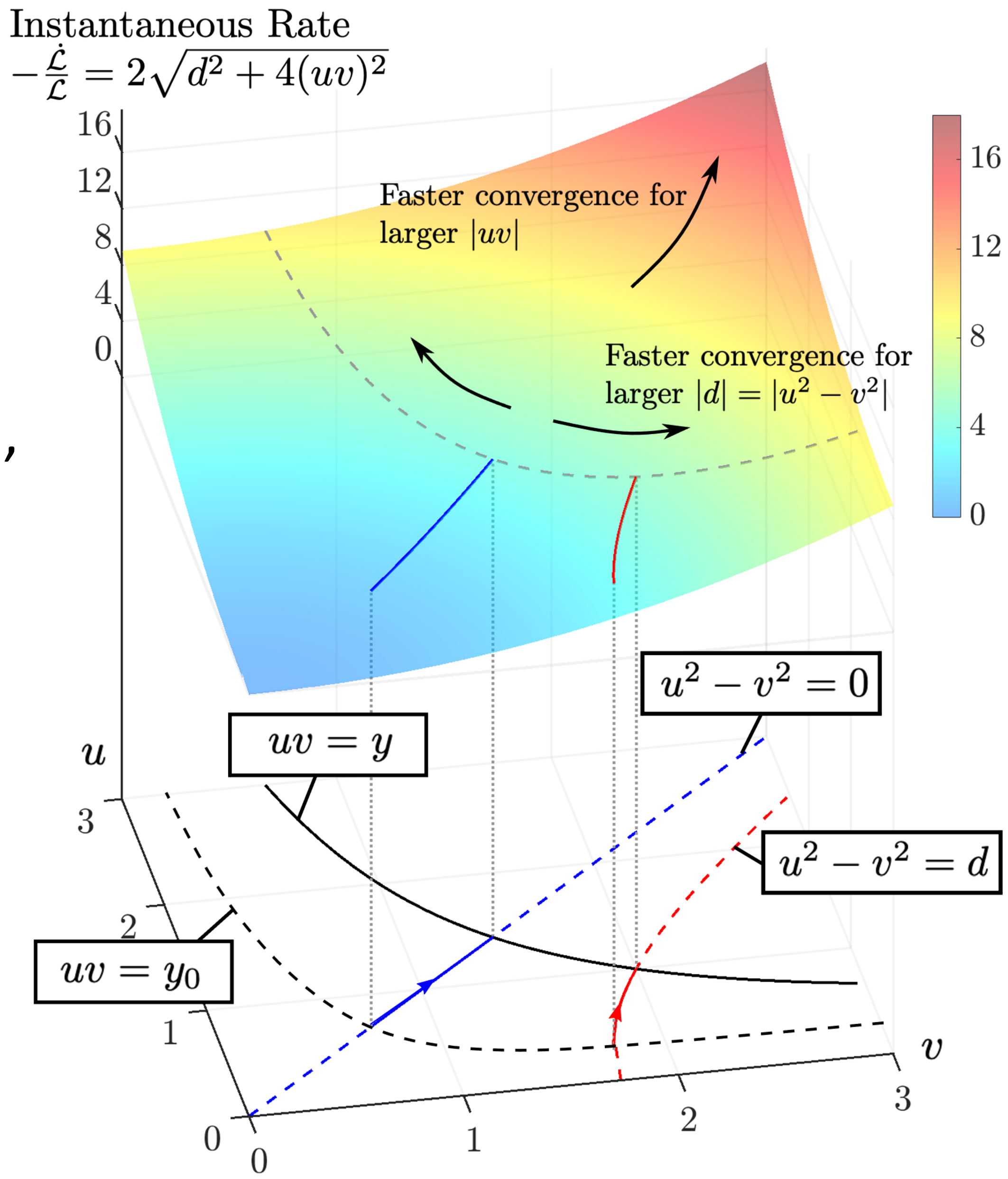
Convergence Analysis: Insights from Scalar Dynamics

Consider the gradient flow on

$$L(u, v) = |y - uv|^2/2$$

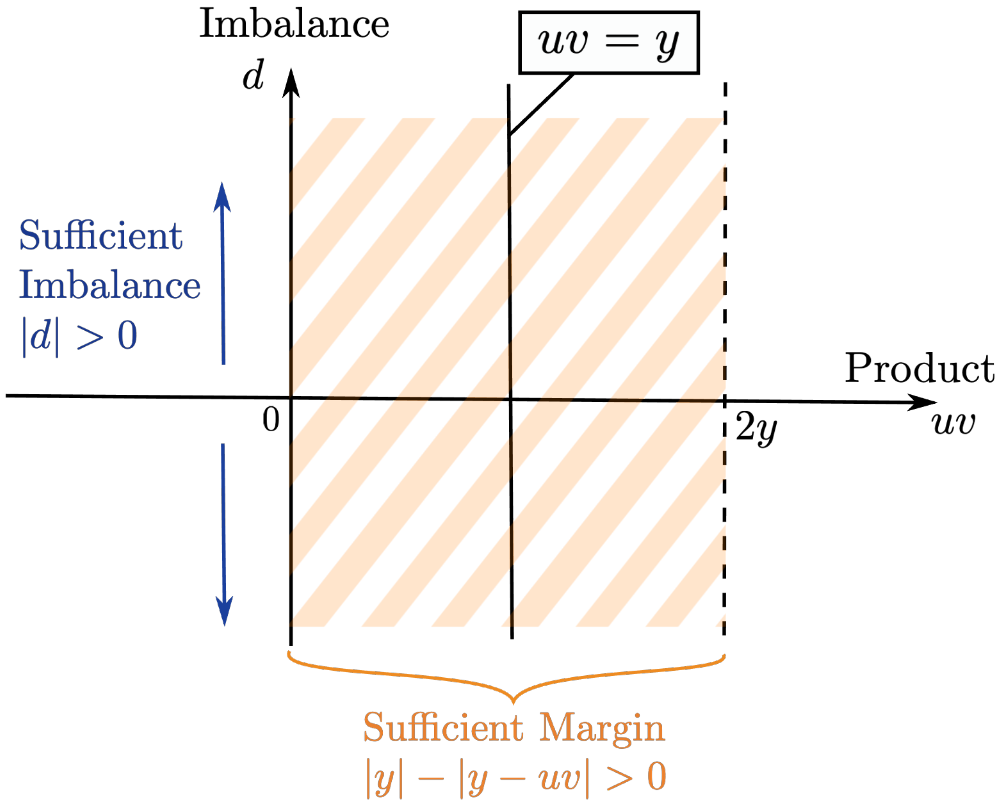
- The imbalance $d = u^2 - v^2$ is time invariant
- Start with same initial product $uv = y_0$, different imbalance leads to different trajectory (solid lines)

- The instantaneous rate $-\dot{L}/L$ is closely related to the exponential convergence
- Instantaneous rate depends on the **imbalance** d and the **product** uv



Convergence Analysis: Insights from Scalar Dynamics

Instantaneous rate depends on the **imbalance** $|d|$ and the **product** $|uv|$



Proper initialization controls d and uv for the entire trajectory:

- $|d|$ is time invariant
- Positive margin $|y| - |y - uv| > 0$ ensures $|uv|$ stays above margin

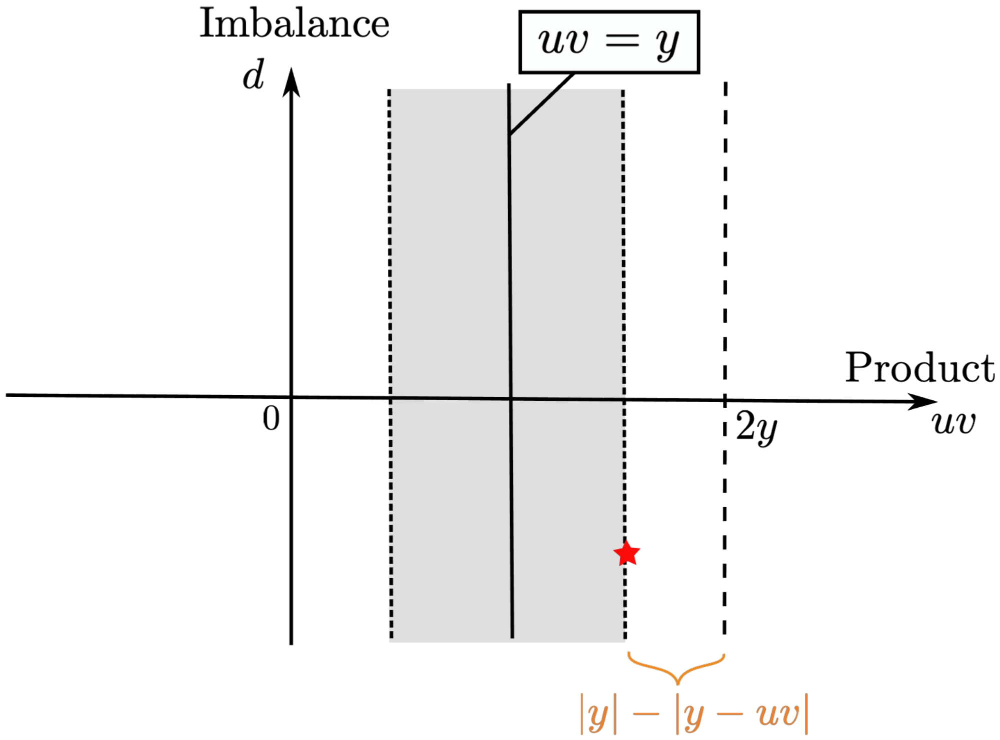
A lower bound on Instantaneous rate leads to exponential convergence

$$-\dot{L}(t)/L(t) \geq c \implies \int_0^t \dot{L}(t)/L(t)dt \leq -ct$$

$$\implies \log \frac{L(t)}{L(0)} \leq -ct \implies L(t) \leq \exp(-ct)\, L(0)$$

Convergence Analysis: Insights from Scalar Dynamics

Instantaneous rate depends on the **imbalance** $|d|$ and the **product** $|uv|$



Proper initialization controls d and uv for the entire trajectory:

- $|d|$ is time invariant
- Positive margin $|y| - |y - uv| > 0$ ensures $|uv|$ stays above margin

A lower bound on Instantaneous rate leads to exponential convergence

$$-\dot{L}(t)/L(t) \geq c \implies \int_0^t \dot{L}(t)/L(t)dt \leq -ct$$

$$\implies \log\frac{L(t)}{L(0)} \leq -ct \implies L(t) \leq \exp(-ct)\, L(0)$$

Convergence Analysis: Our Contribution

In our problem setting, we study the gradient flow on $U_1 \in \mathbb{R}^{r\times h}, U_2 \in \mathbb{R}^{m\times h}$ with loss function

$$\tilde{L}(U_1, V) = \frac{1}{2}\left\|W^T Y - \Sigma_x^{1/2} U_1 V^T\right\|_F^2 = L(U,V) - L^*$$

We show

- A lower bound on the instantaneous rate $-\dot{\tilde{L}}/\tilde{L}$ that depends on the **imbalance** $D = U_1^T U_1 - V^T V$ and the **product** $U_1 V^T$
- Two types of initialization that guarantees initialization
  - Sufficient level of imbalance
  - Sufficient margin

Convergence Analysis: Our Contribution

In our problem setting, we study the gradient flow on $U_1 \in \mathbb{R}^{r\times h}, U_2 \in \mathbb{R}^{m\times h}$ with loss function

$$\tilde{L}(U_1, V) = \frac{1}{2}\left\| W^T Y - \Sigma_x^{1/2} U_1 V^T \right\|_F^2 = L(U,V) - L^*$$

We show

- A lower bound on the instantaneous rate $-\dot{\tilde{L}}/\tilde{L}$ that depends on the **imbalance** $D = U_1^T U_1 - V^T V$ and the **product** $U_1 V^T$
- Two types of initialization that guarantees initialization
  - Sufficient level of imbalance
  - Sufficient margin

For simplicity, we present the results for the case $W = I_r, \Sigma_x = I_r$

Convergence Analysis: Instantaneous Rate

***Proposition** 1. (Lower bound on instantaneous rate, $\Sigma_x = I_r$) Define $D = U_1^T U_1 - V^T V$. Let $\dot{\tilde{L}}(U_1, V)$ be the time derivative of $\tilde{L}(U_1, V)$ under gradient flow. Then we have*

$$-\frac{\dot{\tilde{L}}(U_1,V)}{\tilde{L}(U_1,V)} \geq -\bar{\lambda}_+ + \underline{\lambda}_- + \sqrt{\left(\bar{\lambda}_+ + \underline{\lambda}_-\right)^2 + 4\sigma_m^2(U_1V^T)}$$
$$-\bar{\lambda}_- + \underline{\lambda}_+ + \sqrt{\left(\bar{\lambda}_- + \underline{\lambda}_+\right)^2 + 4\sigma_r^2(U_1V^T)},$$

where

$$\bar{\lambda}_+ = max\{\lambda_1(D), 0\}, \qquad \underline{\lambda}_- = max\{\lambda_m(-D), 0\}$$
$$\bar{\lambda}_- = max\{\lambda_1(-D), 0\}, \qquad \underline{\lambda}_+ = max\{\lambda_r(-D), 0\}$$

- (Recall) insta. rate in scalar dynamics: $-\dot{L}/L = 2\sqrt{d^2 + 4(uv)^2}$
- Tightness: Fix imbalance D and product U_1V^T, there exists U_1, V that attains the exact lower bound for the insta. rate

Exponential Convergence Guarantee

***Theorem** 1. (Exponential convergence, $\Sigma_x = I_r$) Let*

$$c = -\bar{\lambda}_+ + \underline{\lambda}_- + \sqrt{\left(\bar{\lambda}_+ + \underline{\lambda}_-\right)^2 + 4(max\{\sigma_m(Y) - \|Y - U_1V^T\|_F, 0\})^2}$$

$$-\bar{\lambda}_- + \underline{\lambda}_+ + \sqrt{\left(\bar{\lambda}_- + \underline{\lambda}_+\right)^2 + 4(max\{\sigma_r(Y) - \|Y - U_1V^T\|_F, 0\})^2},$$

then under gradient flow satisfies

$$\tilde{L}(t) \leq \exp(-c(0)t)\tilde{L}(0)\,, t \geq 0$$

i.e., if $c(0) > 0$, $\tilde{L}(t)$ converges to zero exponentially at a rate at least $c(0)$.

- Control the Imbalance and margin at initialization
- $\tilde{L}(U_1, V) = L(U, V) - L^*$, $L(t)$ converges to its global minimum exponentially
- First exponential convergence result for non-spectral initialization with general imbalance structure

Exponential Convergence Guarantee

***Corollary** 1. (Sufficient level of imbalance* [Min'21] *) If at initialization, we have*

$$c' = \underline{\lambda}_- + \underline{\lambda}_+ > 0,$$

then $\tilde{L}(t)$ converges to zero exponentially at a rate at least $2c'$.

- $-\bar{\lambda}_+ + \underline{\lambda}_- + \sqrt{\left(\bar{\lambda}_+ + \underline{\lambda}_-\right)^2 + 4(max\{\sigma_m(Y) - \|Y - U_1V^T\|_F, 0\})^2} \geq 2\underline{\lambda}_-$

***Corollary** 2. (Sufficient margin) If at initialization, we have*

$$\sigma_{min}(Y) - \|Y - U_1V^T\|_F > 0,$$

then $c(0) > 0$ and $\tilde{L}(t)$ converges to zero exponentially at a rate at least $c(0)$.

- Convergence with positive margin was studied for sufficiently balanced initialization [Arora'18]
- No requirement on imbalance here

Exponential Convergence Guarantee

***Corollary** 3. (Characterizing local convergence rate) If at some t_0, we have $t_0 > 0$, then*

$$\tilde{L}(t) \leq \exp(-c(t_0)t)\tilde{L}(t_0)\,, t \geq t_0.$$

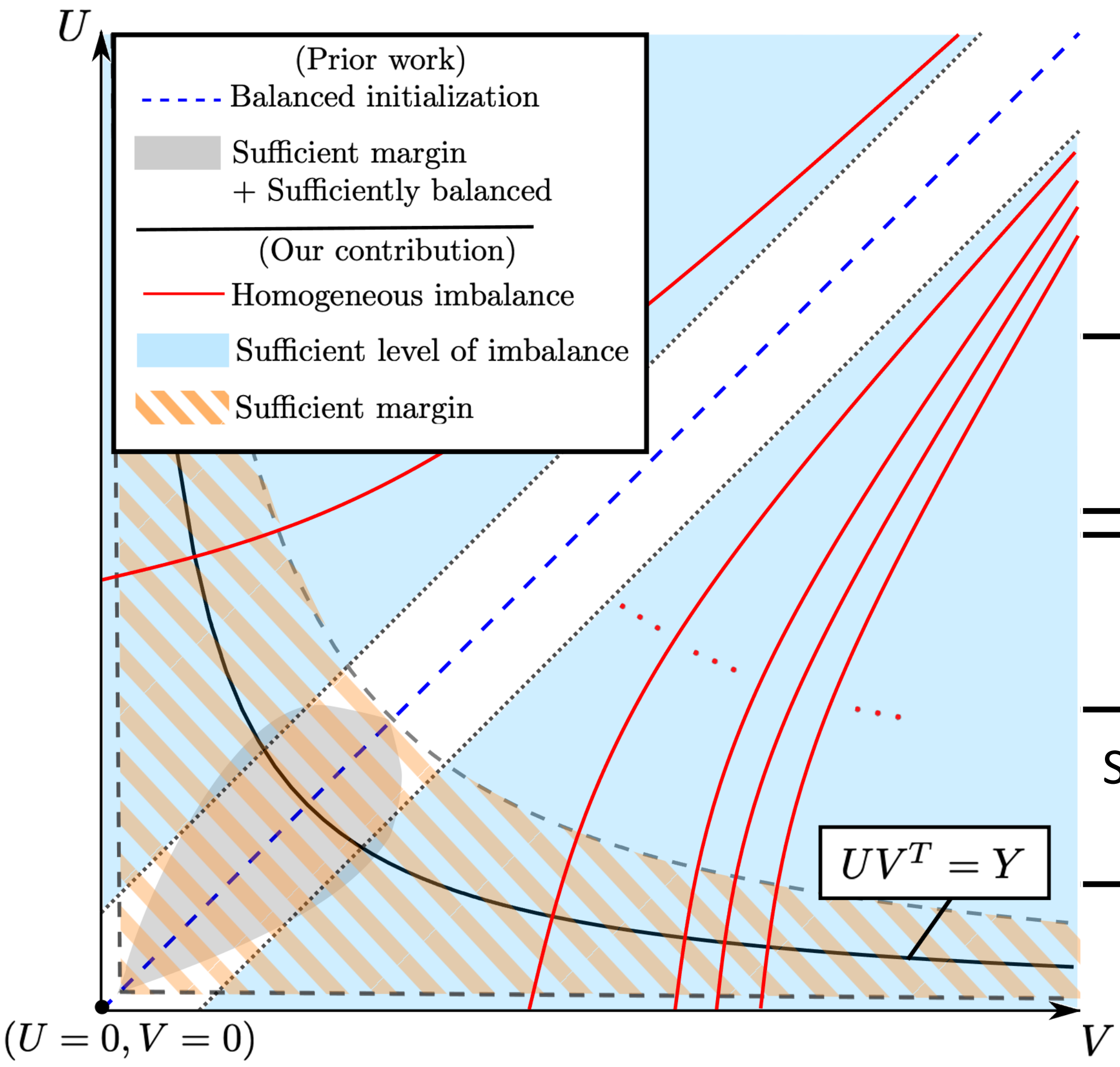
That is, after t_0, the loss converges to zero exponentially with a rate at least $c(t_0)$. Notably, for sufficiently large t_0, we have

$$c(t_0) \approx -\bar{\lambda}_+ + \underline{\lambda}_- + \sqrt{\left(\bar{\lambda}_+ + \underline{\lambda}_-\right)^2 + 4\sigma_m^2(Y)}$$

$$-\bar{\lambda}_- + \underline{\lambda}_+ + \sqrt{\left(\bar{\lambda}_- + \underline{\lambda}_+\right)^2 + 4\sigma_r^2(Y)}.$$

- Convergence rate around equilibrium: imbalance D and target Y
- [Salma'21] studies the local convergence rate when $D = \lambda_0 I_h, |\lambda_0| > 0$
- No assumption on the imbalance structure here

Exponential Convergence Guarantee: Summary



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

Balanced initialization	$D := U^TU - V^TV = 0$
Margin + approx. balanced [Arora'18]	$\sigma_{min}(Y) - \|Y - UV^T\| > \delta$ $\|D\|_F \leq C\delta^2$
Homogeneous imbalance [Salma'21]	$D = \lambda_0 I_h, \lvert\lambda_0\rvert > 0$
Sufficient level of imbalance [Min'21]	$\underline{\lambda}_- + \underline{\lambda}_+ > 0$
Sufficient margin	$\sigma_{min}(Y) - \|Y - UV^T\| > 0$

Overview

- **Sufficient imbalance** or **sufficient margin** guarantees exponential convergence
- **Orthogonal initialization** leads to min-norm solution
- **Random initialization** + **large network width** approximately satisfies the two conditions above, allowing us to find near minimum norm solution efficiently

Reference

Saxe, A. M., Mcclelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In International Conference on Learning Representations, 2014.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. In International Conference on Learning Representations, 2019.

Du, S. and Hu, W. Width provably matters in optimization for deep linear neural networks. In International Conference on Machine Learning, pp. 1655–1664, 2019.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6152–6160, 2017.