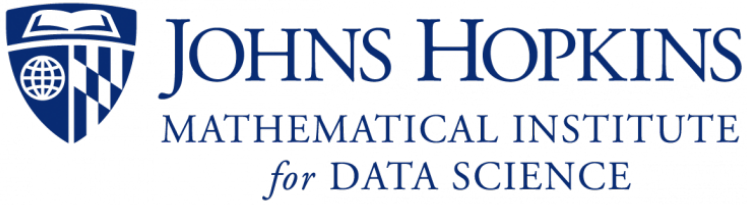
On the Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks

Hancheng Min, Salma Tarmoun, René Vidal and Enrique Mallada



ICML2021, Virtual Conference, July 18th – 24th

Introduction

- Theoretical Understanding of Deep Learning:

  Overparametrization / Initialization ??? → **Convergence** (Efficient Training), **Implicit Bias** (Generalization)

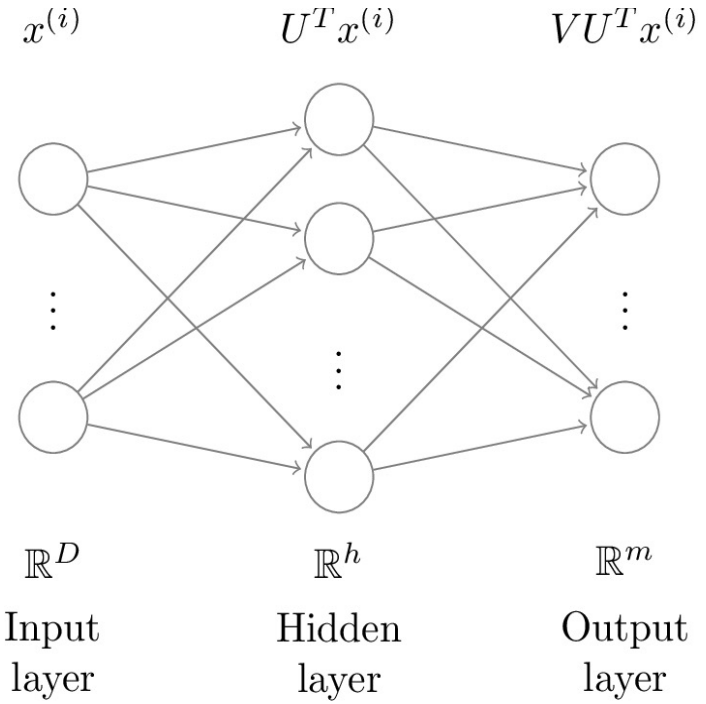
- Analysis for Linear Networks

Gradient flow on two-layer linear networks:

Training data: $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - VU^T x^{(i)}\right)^2 = \frac{1}{2}\|Y - XUV^T\|_F^2$$

$$\dot{U} = -\frac{\partial \mathcal{L}}{\partial U},\ \dot{V} = -\frac{\partial \mathcal{L}}{\partial V}$$



Introduction

- Theoretical Understanding of Deep Learning:

  Overparametrization / Initialization ??? → **Convergence** (Efficient Training) / **Implicit Bias** (Generalization)

- Analysis for Linear Networks: Prior Works
  - **Convergence**
    - Spectral Initialization (Saxes'14), Balanced Initialization (Arora'19)
      Do not work for random initialization
    - Kernel Regime (Du&Hu'19)
      Random initialization, but requires large network width
  - **Implicit Bias**
    - Vanishing Initialization (Gunasekar'17)
      Initialization close to zero implies slow convergence

We study more general types of initialization

Main Results

We decompose the weights of the first layer according to the SVD of the input data:

$$U = \Phi_1 \underbrace{\Phi_1^T U}_{:=U_1} + \Phi_2 \underbrace{\Phi_2^T U}_{:=U_2}, \quad \text{where } X \overset{\text{(SVD)}}{=} W \begin{bmatrix} \underbrace{\Sigma_x^{1/2}}_{\mathbb{R}^{r\times r}} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \end{bmatrix}$$

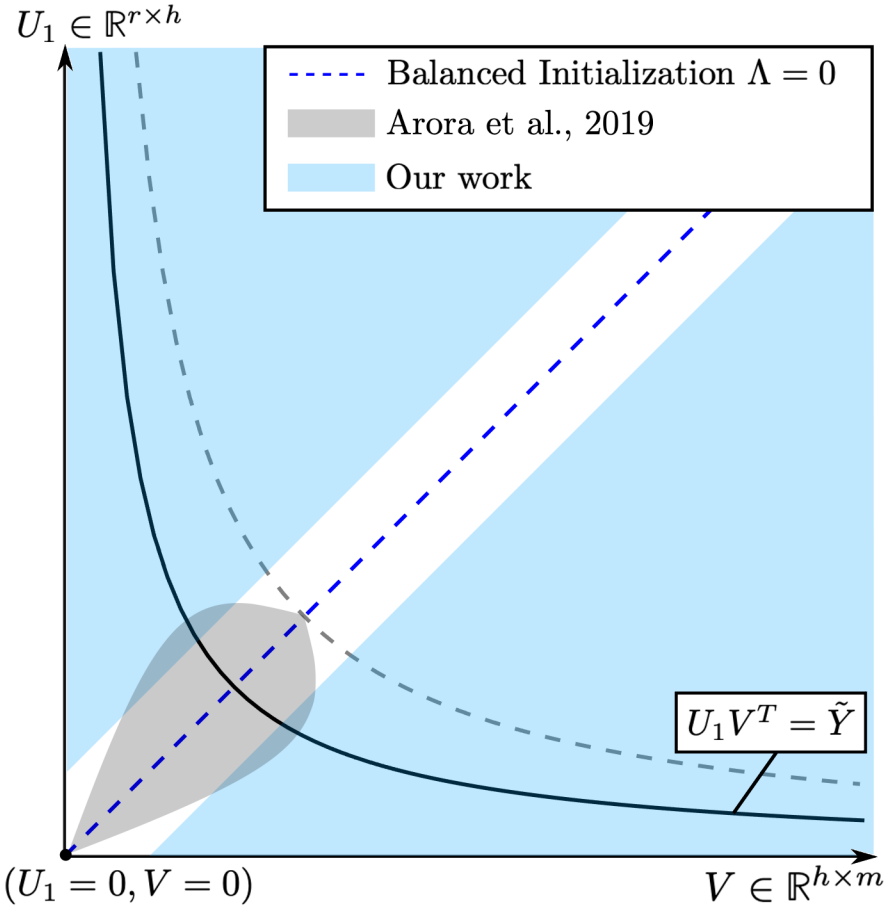
- Assume $\Sigma_x = I$, rewrite the loss:

  $$\mathcal{L} = \tfrac{1}{2}\|W^T Y - U_1 V^T\|_F^2$$

  Matrix factorization for target $\tilde{Y} = W^T Y$

  Gradient flow on U_1, V

- Imbalance $\Lambda = U_1^T U_1 - V^T V$
- Previous work (Arora'19) studied the balanced initialization $\Lambda = 0$



Main Results

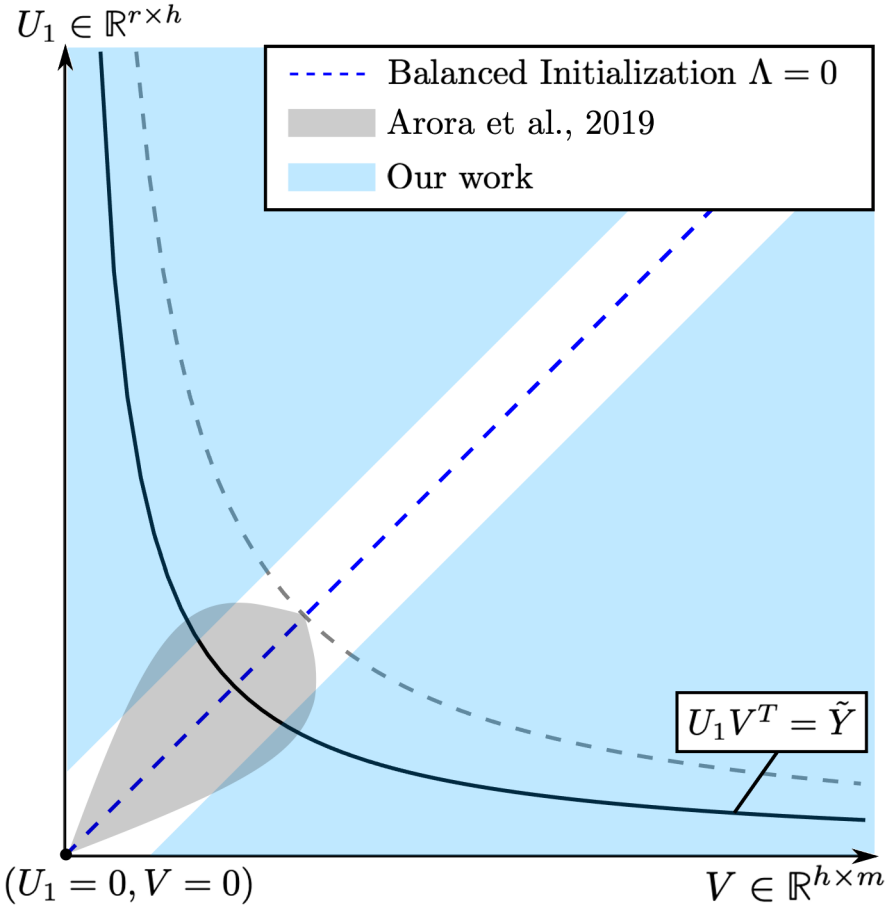
Imbalanced Initialization Guarantees Exponential Convergence

Theorem (Informal). *If the initialization satisfies* $c := [\lambda_r(\Lambda)]_+ + [\lambda_m(-\Lambda)]_+ > 0$, *then the gradient flow converges to global optimum exponentially with a rate at least* c.

- Level of imbalance

  $$c = [\lambda_r(\Lambda)]_+ + [\lambda_m(-\Lambda)]_+$$

  measures how U_1, V are different in terms of their singular values and row spaces
- Random initialization almost surely has positive level of imbalance



Main Results

We decompose the weights of the first layer according to the SVD of the input data:

$$U = \Phi_1 \underbrace{\Phi_1^T U}_{:=U_1} + \Phi_2 \underbrace{\Phi_2^T U}_{:=U_2}, \quad \text{where } X \overset{\text{(SVD)}}{=} W \begin{bmatrix} \underbrace{\Sigma_x^{1/2}}_{\mathbb{R}^{r\times r}} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \end{bmatrix}$$

Orthogonal Initialization Leads to Min-norm Solution

Proposition (Informal). *If the initialization satisfies* $VU_2^T = 0,\ U_1 {U_2}^T = 0$, *then the gradient flow, if converges, finds the min-norm solution.*

- Min-norm solution $\hat{\Theta} = \underset{\Theta\in\mathbb{R}^{D\times m}}{\arg\min}\{\|\Theta\|_F : \|Y - X\Theta\|_F^2 = \min_{\Theta}\|Y - X\Theta\|_F^2\}$
- Extension of "initializing Θ within the span of the input data leads to min-norm solution" in standard linear regression problem

Main Results

Imbalanced Initialization Guarantees Exponential Convergence

$$c = [\lambda_r(\Lambda(0)]_+ + [\lambda_m(-\Lambda(0)]_+ > 0$$

Orthogonal Initialization Leads to Min-norm Solution

$$VU_1^T = 0$$

$$U_1U_2^T = 0$$

Random initialization

$$[U]_{ij} \sim \mathcal{N}\left(0, h^{-1}\right), \ [V]_{ij} \sim \mathcal{N}\left(0, h^{-1}\right)$$

+ Large hidden layer width h

With high probability, we have

Sufficient level of imbalance

$$c \geq 1$$

Approximate Orthogonality

$$\|VU_1^T\|_F \leq \mathcal{O}(h^{-1/2})$$

$$\|U_1U_2^T\|_F \leq \mathcal{O}(h^{-1/2})$$

Theorem (Informal). *With random initialization and large hidden layer width, the gradient flow finds a solution within* $\mathcal{O}(h^{-1/2})$ *spectral norm distance to the minimum norm solution with high probability.*

Conclusion

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

- **Imbalanced Initialization** Guarantees Exponential Convergence (*Convergence*)
- **Orthogonal Initialization** Leads to Min-norm Solution (*Implicit bias*)
- **Random initialization** + **large network width** finds near minimum norm solution efficiently

Thank you!

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.