Learning Dynamics and Implicit Bias of Gradient Flow in Overparameterized Linear Models

Enrique Mallada

JOHNS HOPKINS UNIVERSITY



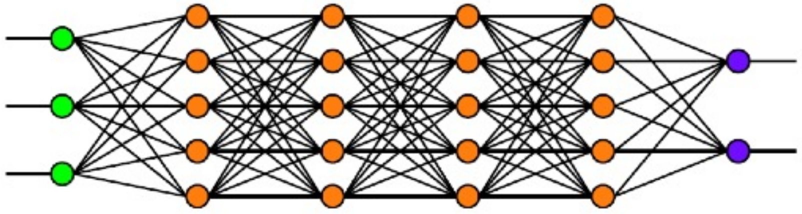


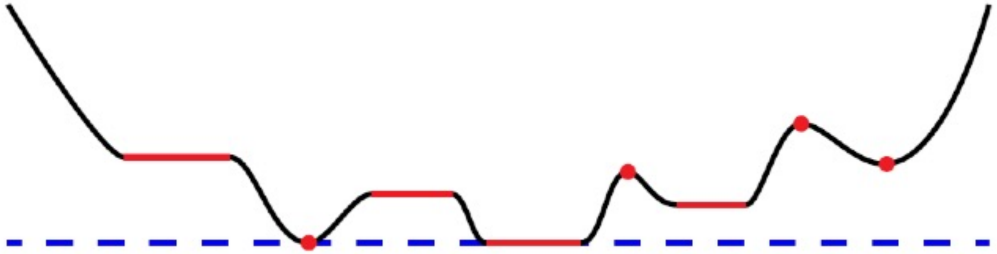
S. Tarmoun H. Min G. Franca B. Haeffele R. Vidal

Joint Mathematics Meeting, Jan 5, 2023

Optimization, Machine Learning and Dynamical Systems

- Optimization has become the workhorse of machine learning
  - Training problem is non-convex and large-scale
  - First order methods: SGD, Momentum, Nesterov, Adagrad, Adam, RMSprop





- Deep neural networks are typically overparameterized
  - Highly underdetermined learning problem with many possible solutions
  - Variants of gradient descent often find one of these solutions

- **Question**: what is the effect of overparameterization on the learning dynamics of optimization algorithms?

Prior Work: Analysis of GD/GF for Overparametrized Models

- In the overparametrized regime, specific initialization may:
  - Promote generalization $\Longrightarrow$ Implicit regularization
  - Accelerate convergence $\Longrightarrow$ Implicit acceleration

- **NTK initialization [1]:** Large hidden layer width, random initialization
  - Exponential convergence for GF
  - "lazy regime": rarely seen in practice [2]

- **Small initialization [3]**: All weights are initialized close to zero
  - Interesting studies on implicit bias: low-rank, sparse models
  - Slow convergence (initialized close to origin, a stationary point) [4]

[1] A Jacot, F Gabriel, and C Hongler. Neural tangent kernel: Convergence and generalization in neural networks. NeurIPS 2018
[2] L Chizat, E Oyallon, and F Bach. On lazy training in differentiable programming. NeurIPS 2019.
[3] D Stöger and M Soltanolkotabi. Small random initialization is akin to spectral learning. NeurIPS 2021.
[4] J Li, T V Nguyen, C Hegde, and R K. W. Wong. Implicit sparse regularization: The impact of depth and early stopping. NeurIPS 2021.

Prior Work: Analysis of GD/GF for Linear Networks

- Non-NTK, non-small initialization is mostly studied for linear networks
- Existing analysis of convergence of GD/GF for two-layer linear networks requires strong assumptions on the initialization

	Spectral	Nonspectral (with sufficient margin)
Balanced	Saxe '14 Gidel '19	Arora '18
Sufficiently imbalanced	Yun '21 Tarmoun '21	Tarmoun '21 Min '21

A Saxe, J Mcclelland, and S Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural network." ICLR 2014
G Gidel, F Bach, and S Lacoste-Julien. "Implicit regularization of discrete gradient dynamics in linear neural networks." NeurIPS 2019
S Arora, N Cohen, N Golowich, and W Hu. "A convergence analysis of gradient descent for deep linear neural networks." ICLR 2018
S Tarmoun, G França, B D Haeffele, and R Vidal. "Understanding the dynamics of gradient flow in overparameterized linear models." ICML 2021
C Yun, S Krishnan, and H Mobahi. A unifying view on implicit bias training linear neural networks. ICLR 2020

Contributions: Analysis of Gradient Flow for Linear Networks

- **Convergence Analysis**
  - **Tarmoun ’21**: spectral or homogeneously imbalanced initializations
    - Closed form solution via Ricatti equations

$$Convergence\ Rate = \sqrt{(Imbalance)^2 + 4\sigma_{min}(Data)^2}$$

  - **Min ’21**: initialization with sufficient imbalance or sufficient margin
    - Grönwall's inequality

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

- **Implicit Bias:** orthogonal initialization leads to min-norm solution [2]

- **Random initialization + large network width** approximately satisfies the two conditions above, allowing us to find near minimum norm solution efficiently [2]

[1] Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21
[2] Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Overparametrized Linear Regression & Matrix Factorization

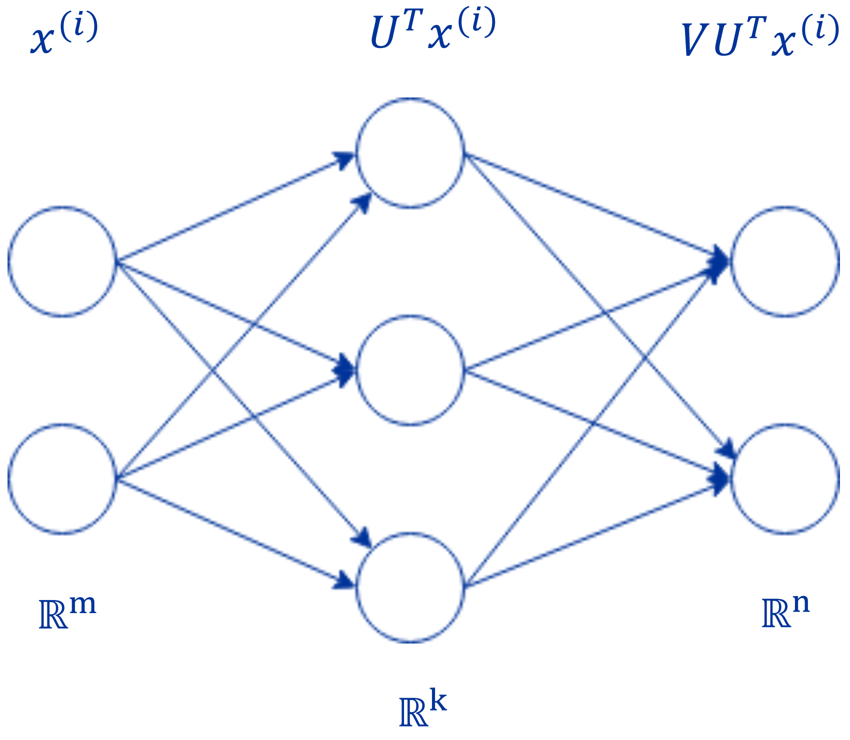
- Linear regression with squared loss

$$\min_{W \in \mathbb{R}^{m \times n}} \frac{1}{2}\|Y - XW\|_F^2$$

Data: $X \in \mathbb{R}^{N \times m}, Y \in \mathbb{R}^{N \times n}$

- Regression with two-layer linear network

$$\min_{\substack{U \in \mathbb{R}^{m \times k} \\ V \in \mathbb{R}^{n \times k} \\ k \geq \min\{m,n\}}} \frac{1}{2}\|Y - XUV^\top\|_F^2$$



- If data are whitened ($\Sigma_x = I$), the problem becomes matrix factorization

$$\min_{U,V} \frac{1}{2}\|\Sigma_{xy} - UV^\top\|_F^2 \quad \text{or} \quad \min_{U,V} \frac{1}{2}\|Y - UV^\top\|_F^2$$

Warm-up: Scalar Case

- Objective function:

$$\ell(x) = \frac{1}{2}(y - x)^2, x \in \mathbb{R}$$

- Gradient flow:

$$\dot{x} = -\nabla \ell_x = y - x$$

- Solution:

$$x(t) = y + (x_0 - y)e^{-t}$$

- Convergence rate:

$$O(e^{-t})$$

- Objective function:

$$\ell(x) = \frac{1}{2}(y - uv)^2, u, v \in \mathbb{R}$$

- Gradient flow:

$$\dot{u} = -\nabla \ell_u = (y - uv)v$$
$$\dot{v} = -\nabla \ell_v = (y - uv)u$$

- Solution: ?

- Convergence rate: ?

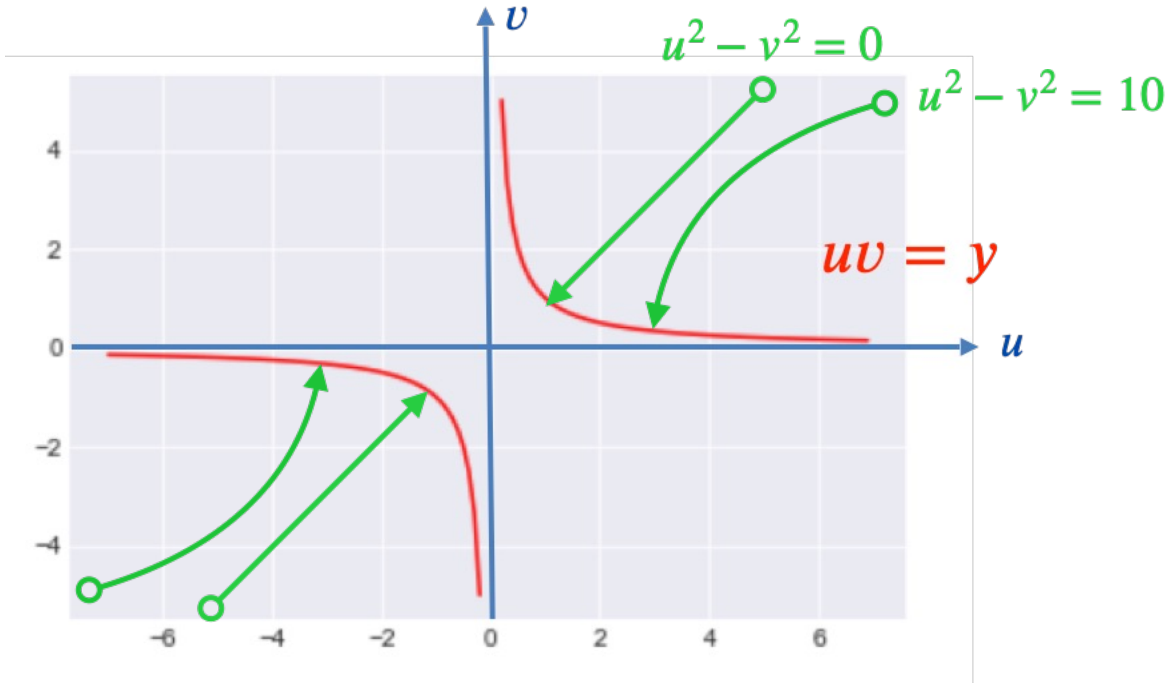
[1] Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21
[2] Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Conservation Law = Hyperbolic Trajectories

- Gradient flow induces conservation law

$$\begin{matrix}\dot{u} = (y - uv)v \\ \dot{v} = (y - uv)u\end{matrix} \Longrightarrow \frac{d}{dt}(u^2 - v^2) = 0 \Longrightarrow u_t^2 - v_t^2 = u_0^2 - v_0^2 = \lambda_0$$

- Law arises due to scaling symmetry $u \to \alpha u$, $v \to v/\alpha$



Noether's theorem explains connections between symmetries and conservation laws.

Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21

Overparametrized Gradient Flow Dynamics

- What are the induced dynamics on $x = uv$?

Gradient flow on (u, v)

$$\dot{u} = (y - uv)v$$
$$\dot{v} = (y - uv)u$$

Induced dynamics on x

$$\dot{x} = \frac{d}{dt}(uv) = (y - uv)(u^2 + v^2)$$

x ?

- Can we get rid of (u, v)?

Imbalance

$$u^2 - v^2 = \lambda_0$$

$$u^2 = \tfrac{1}{2}\left(\lambda_0 + \sqrt{\lambda_0^2 + 4(uv)^2}\right)$$
$$v^2 = \tfrac{1}{2}\left(-\lambda_0 + \sqrt{\lambda_0^2 + 4(uv)^2}\right)$$

Induced flow on x

$$\dot{x} = (y - x)\sqrt{\lambda_0^2 + 4x^2}$$

Gradient flow on x

$$\dot{x} = y - x$$

Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21

Convergence Rate of Overparametrized GF

- How does $x = uv$ behave under gradient flow on (u, v)

$$\dot{x} = (y - x)\sqrt{\lambda_0^2 + 4x^2}$$

- Closed form solution

$$x(t) = \frac{ye^{2t\sqrt{\lambda_0^2+4y^2}} - 2c\lambda_0^2 e^{t\sqrt{\lambda_0^2+4y^2}} - 4y\lambda_0^2 c^2}{e^{2t\sqrt{\lambda_0^2+4y^2}} + 8yce^{t\sqrt{\lambda_0^2+4y^2}} - 4\lambda_0^2 c^2}$$

- Convergence rate

$$|x(t) - y| \approx 2c(\lambda_0^2 + 4y^2)e^{-t\sqrt{\lambda_0^2+4y^2}}$$

$$Rate = \sqrt{(Imbalance)^2 + 4(Data)^2}$$

Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21

Convergence Rate of Overparametrized GF

- Grönwall's inequality: $\dot{\ell}(t) \leq -\alpha\ell(t) \implies \ell(t) \leq \exp(-\alpha t)\ell(0)$

- What are the induced dynamics on $\ell = (y - uv)^2/2$?

Gradient flow on (u, v)

$$\dot{u} = (y - uv)v$$
$$\dot{v} = (y - uv)u$$

→ Induced dynamics on the loss ℓ

$$\dot{\ell} = -(y - uv)^2(u^2 + v^2)$$

$(y - uv)^2$ ↓ 2ℓ $\qquad$ $(u^2 + v^2)$ ↓ $\sqrt{\lambda_0^2 + 4(uv)^2}$

- Margin $\quad |uv| \geq |y| - |y - uv| \geq |y| - |y - u_0 v_0| = Margin$

- Loss convergence rate $\quad Rate \geq 2\sqrt{(Imbalance)^2 + 4(Margin)^2}$

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Summary of Convergence Rate in Scalar Case

- Loss convergence rate

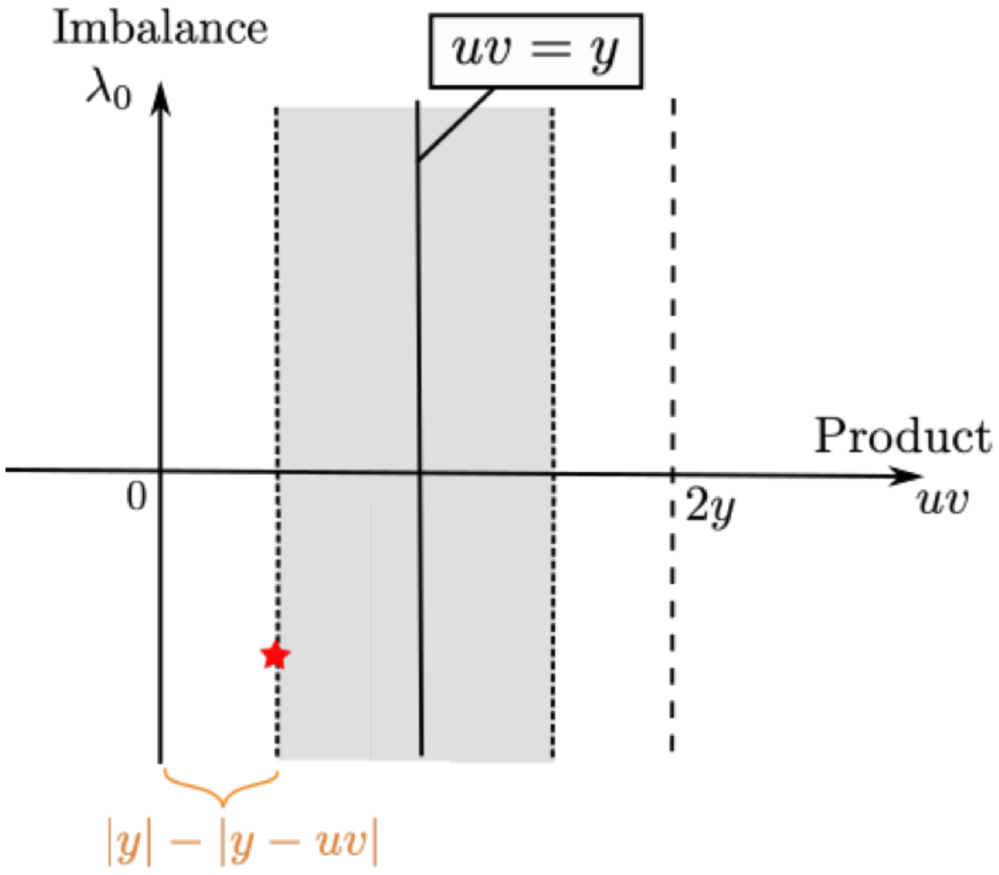
$$Rate \geq 2\sqrt{(Imbalance)^2 + 4(Margin)^2}$$

- Sufficient imbalance:

$$\lambda_0^2 = (u_0^2 - v_0^2)^2 > 0$$

- Sufficient margin:

$$|y| - |y - u_0 v_0| > 0$$



From Scalar to Matrix Case

	Scalar case
Objective function	$\ell(u,v) = \frac{1}{2}(y - uv)^2$
Gradient flow	$\dot{u}(t) = (y - uv)v$ $\dot{v}(t) = (y - uv)u$
Conservation law	$u^2 - v^2 = \lambda_0$
Induced flow	$\dot{x} = (y - x)\sqrt{\lambda_0^2 + 4x^2}$
Convergence rate	$O\left(e^{-t\sqrt{\lambda_0^2 + 4y^2}}\right)$

From Scalar to Matrix Case

	Scalar case	Matrix case
Objective function	$\ell(u,v) = \frac{1}{2}(y-uv)^2$	$\ell(U,V) = \frac{1}{2}\|Y - UV^\top\|_F^2$
Gradient flow	$\dot{u}(t) = (y-uv)v$ $\dot{v}(t) = (y-uv)u$	$\dot{U} = (Y - UV^\top)V$ $\dot{V} = (Y - UV^\top)^\top U$
Conservation law	$u^2 - v^2 = \lambda_0$	$U^\top U - V^\top V = \Lambda_0$
Induced flow	$\dot{x} = (y-x)\sqrt{\lambda_0^2 + 4x^2}$	Ricatti equation on X
Convergence rate	$O\left(e^{-t\sqrt{\lambda_0^2+4y^2}}\right)$	$O\left(e^{-t\sqrt{(Imbalance)^2+4(Data)^2}}\right)$

Spectral Initialization:

$$Rate = \sqrt{(Imbalance)^2 + 4\sigma_{min}(Data)^2}$$

- What are the induced dynamics on (U, V)?

Gradient flow on (U, V)

$$\dot{U} = (Y - UV^\top)V$$

$$\dot{V} = (Y - UV^\top)^\top U$$

→

Induced dynamics on X

$$\dot{X} = (Y - UV^\top)VV^\top + UU^\top(Y - UV^\top)$$

- Spectral initialization

$$Y = \Phi\Sigma\Psi^\top \implies X_0 = \Phi\Sigma_0\Psi^\top$$

- Spectral solution

s-vectors remain constant

$$X(t) = \Phi\Sigma(t)\Psi^\top$$

s-values follow scalar dynamics

$$\dot{\sigma}_i(t) = (\sigma_i - \sigma_i(t))\sqrt{\lambda_{0,i}^2 + 4\sigma_i(t)^2}$$

- Convergence rate
  - Large s-values converge faster
  - Large imbalance, faster convergence

$$O\left(e^{-t\sqrt{\lambda_{0,i}^2 + 4\sigma_i(Y)^2}}\right)$$

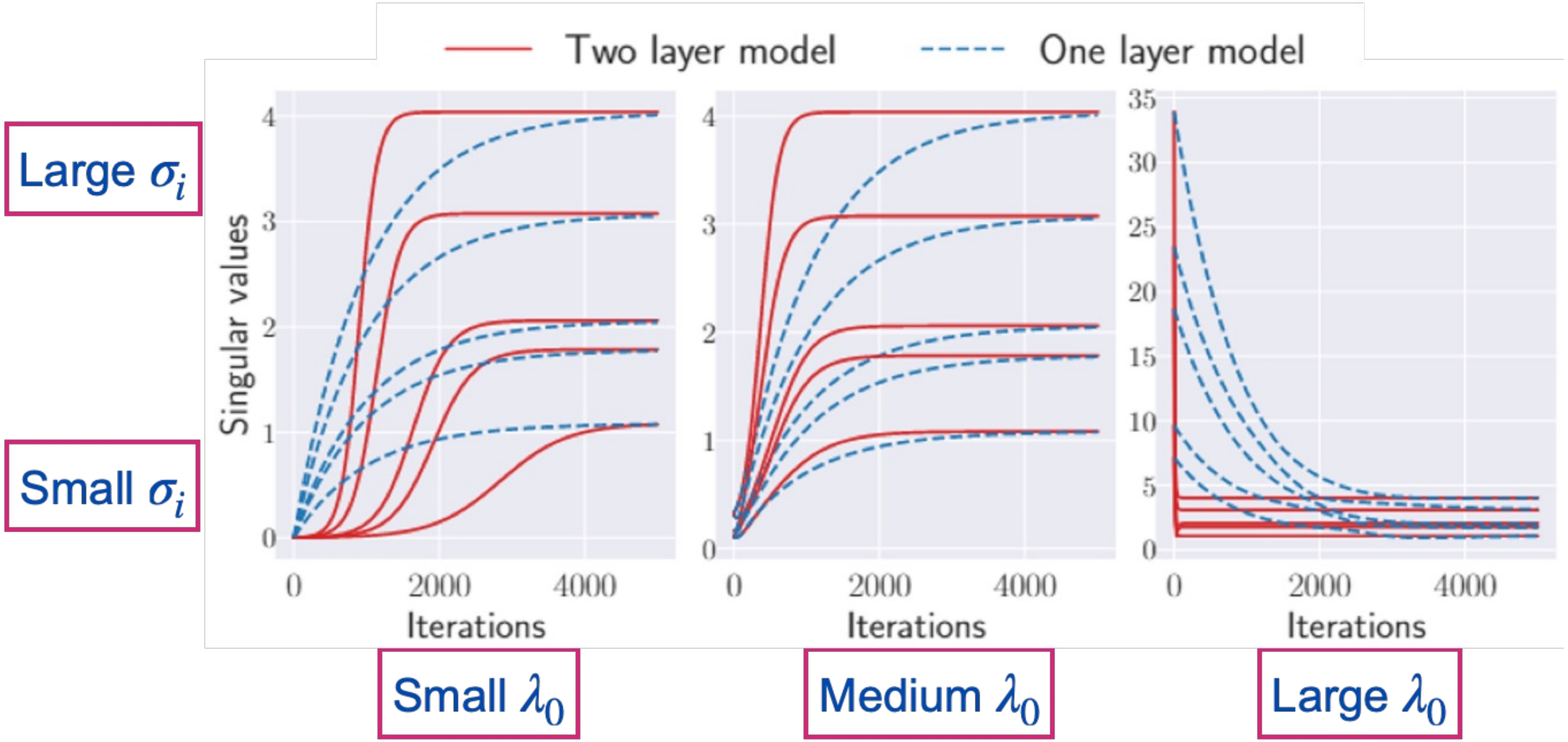
Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21

Spectral Initialization:

$$Rate = \sqrt{(Imbalance)^2 + 4\sigma_{min}(Data)^2}$$

- Convergence rate
  - Large s-values converge faster
  - Large imbalance, faster convergence

$$O\left(e^{-t\sqrt{\lambda_{0,i}^2 + 4\sigma_i(Y)^2}}\right)$$



General Initialization:

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

- Scalar Case

Induced dynamics on the loss ℓ

$$\dot{\ell} = -2\ell(u^2 + v^2)$$

Imbalance

$$\lambda_0 = u^2 - v^2$$

- Matrix Case

Induced dynamics on the loss ℓ

$$\dot{\ell} \leq -2\ell\left(\lambda_n(UU^\top) + \lambda_m(VV^\top)\right)$$

Imbalance matrix eigenvalues

$$\{l_i\} = f(\Lambda_0), \quad \Lambda_0 = U^\top U - V^\top V$$

$$\bar{\lambda}_+ = \max(\lambda_1(\Lambda_0), 0) \qquad l_1 = -\bar{\lambda}_+ + \underline{\lambda}_-$$

$$\bar{\lambda}_- = \max(\lambda_1(-\Lambda_0), 0) \qquad l_2 = \bar{\lambda}_+ + \underline{\lambda}_-$$

$$\underline{\lambda}_+ = \max(\lambda_n(\Lambda_0), 0) \qquad l_3 = -\bar{\lambda}_- + \underline{\lambda}_+$$

$$\underline{\lambda}_- = \max(\lambda_m(-\Lambda_0), 0) \qquad l_4 = \bar{\lambda}_- + \underline{\lambda}_+$$

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

General Initialization:

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

- Scalar Case

Induced dynamics on the loss ℓ

$$\dot{\ell} = -2\ell(u^2 + v^2)$$

Imbalance

$$\lambda_0 = u^2 - v^2$$

$$u^2 = \frac{1}{2}\left(\lambda_0 + \sqrt{\lambda_0^2 + 4(uv)^2}\right)$$

$$v^2 = \frac{1}{2}\left(-\lambda_0 + \sqrt{\lambda_0^2 + 4(uv)^2}\right)$$

Margin

$$|y| - |y - u_0 v_0| \leq |uv|$$

- Matrix Case

Induced dynamics on the loss ℓ

$$\dot{\ell} \leq -2\ell\left(\lambda_n(UU^\top) + \lambda_m(VV^\top)\right)$$

Imbalance matrix eigenvalues

$$\{l_i\} = f(\Lambda_0), \quad \Lambda_0 = U^\top U - V^\top V$$

$$\lambda_n(UU^\top) \geq \frac{1}{2}\left(l_1 + \sqrt{l_2^2 + 4\sigma_n(UV^\top)}\right)$$

$$\lambda_m(VV^\top) \geq \frac{1}{2}\left(l_3 + \sqrt{l_4^2 + 4\sigma_m(UV^\top)}\right)$$

Margin

$$\sigma_i(Y) - \|Y - U_0 V_0^\top\|_F \leq \sigma_i(UV^\top)$$

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Summary General Initialization

- **Theorem:** Gradient flow on $\ell = \frac{1}{2}||Y - UV^T||_F^2$ satisfies

$$\ell_t \leq \ell_0 \exp(-\alpha t)$$

- **Rate**

$$\alpha = l_1 + \sqrt{l_2^2 + 4\max(\sigma_n(Y) - \|Y - U_0V_0^\top\|_F, 0)^2} + l_3 + \sqrt{l_4^2 + 4\max(\sigma_m(Y) - \|Y - U_0V_0^\top\|_F, 0)^2}$$

Imbalance

Margin

- **Corollary**
  - sufficient imbalance
  - sufficient margin

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

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Problem Setup: Linear Regression

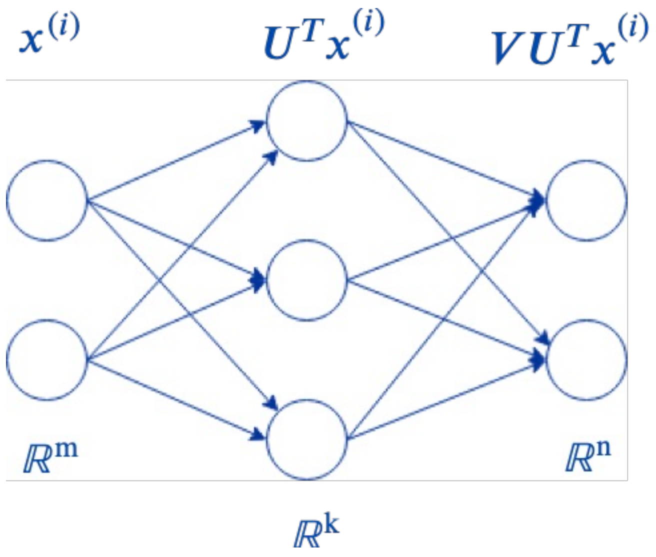
- Two-layer linear network, square loss

$$\min_{U,V} \frac{1}{2}\|Y - XUV^\top\|_F^2$$

$$X \in \mathbb{R}^{N\times m}, Y \in \mathbb{R}^{N\times n}$$

$$U \in \mathbb{R}^{m\times k}, V \in \mathbb{R}^{n\times k}$$

$$k \geq \min\{m, n\}$$



- So far, we have assumed data whitened, i.e., $\Sigma_x = I$
  - Equivalent to matrix factorization
  - $\min_{U,V} \frac{1}{2}\|\Sigma_{xy} - UV^\top\|_F^2$ or $\min_{U,V} \frac{1}{2}\|Y - UV^\top\|_F^2$

- **Question:** How does X affect convergence rate of GF?

Matrix Factorization vs Linear Regression

- **Theorem:** Gradient flow on $\ell = \frac{1}{2}||Y - UV^T||_F^2$ satisfies

$$\ell_t \leq \ell_0 \exp(-\alpha t) \quad \text{with} \quad \alpha \geq 2\sqrt{(Imbalance)^2 + 4(Margin)^2}$$

- **Theorem:** Gradient flow on $\ell = \frac{1}{2}||Y - XUV^T||_F^2$ satisfies

$$\ell_t - \ell^* \leq (\ell_0 - \ell^*)\exp(-\alpha t)$$

where

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

Contributions: Analysis of Gradient Flow for Linear Networks

- **Convergence Analysis**
  - **Tarmoun '21**: spectral or homogeneously imbalanced initializations
    - Closed form solution via Ricatti equations

$$Convergence\ Rate = \sqrt{(Imbalance)^2 + 4\sigma_{min}(Data)^2}$$

  - **Min '21**: initialization with sufficient imbalance or sufficient margin
    - Grönwall's inequality

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

- **Implicit Bias:** orthogonal initialization leads to min-norm solution [2]

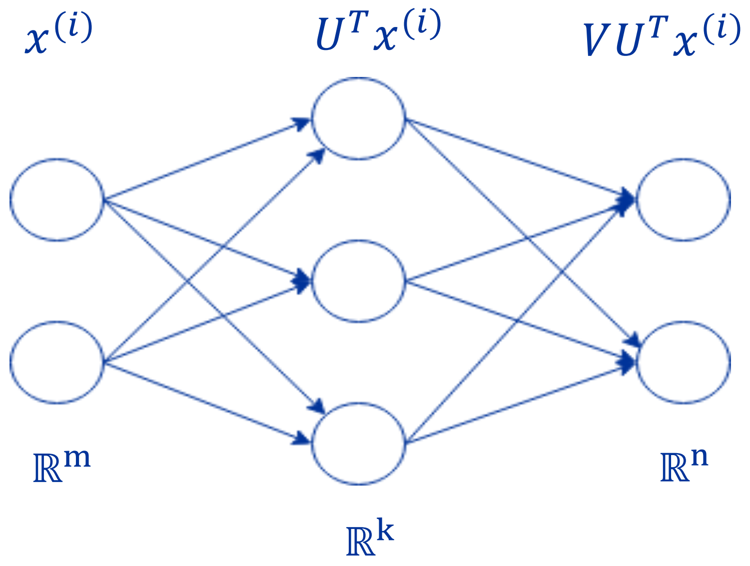
- **Random initialization + large network width** approximately satisfies the two conditions above, allowing us to find near minimum norm solution efficiently [2]

[1] Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21
[2] Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Implicit Bias of Overparametrized Gradient Flow

- Regression with a two-layer linear network

$$\min_{\substack{U\in\mathbb{R}^{m\times k}\\ V\in\mathbb{R}^{n\times k}\\ k\geq\min\{m,n\}}} \quad \frac{1}{2}\|Y - XUV^\top\|_F^2$$



- Assume X is not full rank, let $X = W\begin{bmatrix}\Sigma_x^{1/2} & 0\end{bmatrix}\begin{bmatrix}\Phi_1^\top \\ \Phi_2^\top\end{bmatrix}$ and

decompose $U = \Phi_1 \underbrace{\Phi_1^\top U}_{U_1} + \Phi_2 \underbrace{\Phi_2^\top U}_{U_2}$

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Implicit Bias to Min-norm Solution

- Consider the minimum norm solution Θ^*

$$\min_{\theta \in \Theta} \| \theta \|_F \; where \;\; \Theta = \arg\min_{\theta} \| Y - X\theta \|_F$$

- **Theorem (Orthogonal Initialization)**: If $V(0)U_2(0)^\top = 0$ and $U_1(0)U_2(0)^\top = 0$, and loss converges to a global minimum, then $U(t)V(t)^\top$ converges to min-norm solution

- Orthogonal initialization may not converge, but sufficient imbalance or margin can provide convergence guarantee

- **Question**: can we get both convergence and implicit bias?

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Implicit Bias to Min-norm Solution

- **Large hidden layer width** k

- **Random initialization** $[U(0)]_{i,j}, [V(0)]_{i,j} \sim \mathcal{N}(0, k^{-1})$

- **Theorem**: Assume a random initialization. Then, with high probability $U(t)V(t)^\top$ converges exponentially and

$$\lim_{t \to \infty} \| U(t)V(t)^\top - \Theta^* \|_F = \mathcal{O}(k^{-1/2})$$

Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21

Conclusions

- **Convergence**: Sufficient imbalance or sufficient margin guarantees exponential convergence
- **Implicit Bias**: 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
- **Extensions and ongoing work**:
  - Deep linear networks (Min ’22)
  - Gradient descent (Xu ´22)
  - Imbalance in nonlinear networks (ReLU net, etc.)

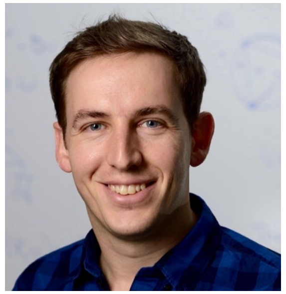
Thanks!


S. Tarmoun


H. Min


G. Franca


B. Haeffele


E. Mallada


R. Vidal

JOHNS HOPKINS UNIVERSITY

Publications:
[1] Tarmoun, França, Haeffele, Vidal. Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models, ICML 21
[2] Min, Tarmoun, Vidal, Mallada. Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks, ICML 21