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

Hancheng Min, Salma Tarmoun, René Vidal, Enrique Mallada
Mathematical Institute for Data Science, Johns Hopkins University

ABSTRACT

In this paper, we present a novel analysis of single-hidden-layer linear networks trained under gradient flow, which connects initialization, optimization, and overparametrization, specifically, we show

- The squared loss converges exponentially to its optimum at a rate that depends on the level of imbalance and the margin of the initialization.
- Proper initialization constrains the dynamics of the network parameters to lie within an invariant set. In turn, minimizing the loss over this set leads to the min-norm solution.
- Large hidden layer width, together with (properly scaled) random initialization, ensures proximity to such an invariant set during training, allowing us to derive a novel non-asymptotic upper-bound on the distance between the trained network and the min-norm solution.

PRELIMINARIES

We study the squared loss on a single-hidden-layer linear network

$$\mathcal{L}(U,V)=\frac{1}{2}\|Y-XUV^T\|_F^2$$

where $X\in\mathbb{R}^{P\times n}$, $Y\in\mathbb{R}^{P\times m}$, $U\in\mathbb{R}^{n\times k}$, $V\in\mathbb{R}^{m\times k}$ with $rank(X)=r$

n:	Input dimension	m:	Output dimension
k:	Hidden layer width	P:	# of data points
r:	rank of input data matrix		

and consider the gradient flow dynamics

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

We decompose the weights of the first layer as

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

We also define the imbalance (invariant under gradient flow)

$$Q=U_1^TU_1-V^TV$$

MAIN RESULTS

Exponential convergence guarantees

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

Theorem 1. *(Exponential convergence) Let $\tilde{Y}=W^TY$, and*

$$\alpha=-\bar{\lambda}_++\underline{\lambda}_-+\sqrt{(\bar{\lambda}_++\underline{\lambda}_-)^2+4\left(max\left\{\sigma_m(\tilde{Y})-\left\|\tilde{Y}-\Sigma_x^{1/2}U_1V^T\right\|_F,0\right\}\right)^2/\lambda_1(\Sigma_x)}$$

$$-\bar{\lambda}_-+\underline{\lambda}_++\sqrt{(\bar{\lambda}_-+\underline{\lambda}_+)^2+4\left(max\left\{\sigma_r(\tilde{Y})-\left\|\tilde{Y}-\Sigma_x^{1/2}U_1V^T\right\|_F,0\right\}\right)^2/\lambda_1(\Sigma_x)},$$

where

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

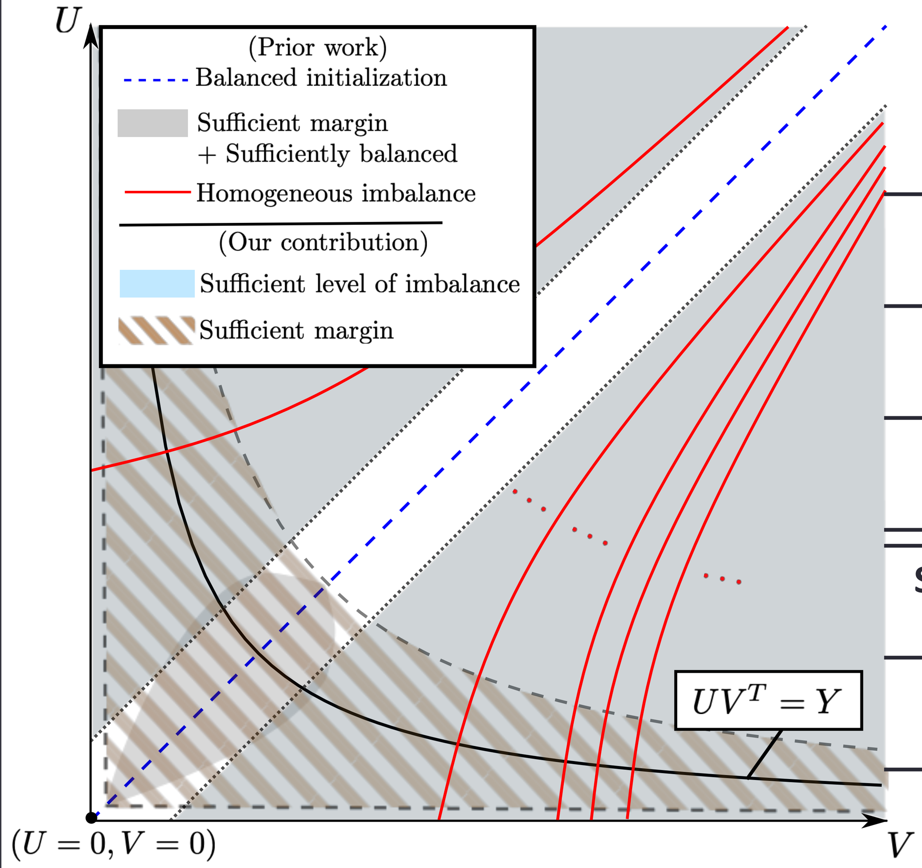
$$\bar{\lambda}_-=max\{\lambda_1(-Q),0\},\qquad \underline{\lambda}_+=max\{\lambda_n(Q),0\}$$

The gradient flow satisfies

$$L(t)-L^*\leq\exp(-\lambda_r(\Sigma_x)\alpha(0)t)(L(0)-L^*),t\geq 0$$

i.e., if $\alpha(0)>0$, $L(t)$ converges to its global minimum exponentially.

- Theorem 1 suggests that either **sufficient level of imbalance** or **sufficient margin** guarantees exponential convergence



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

Balanced initialization	$Q:=U^TU-V^TV=0$
Margin + approx. balanced [Arora'18]	$\sigma_{min}(Y)-\|Y-UV^T\|_F>\delta$ $\|Q\|_F\leq C\delta^2$
Homogeneous imbalance [Tarmoun'21]	$Q=\lambda_0I_h,\lvert\lambda_0\rvert>0$
Sufficient level of imbalance [Min'21]	$\underline{\lambda}_-+\underline{\lambda}_+>0$
Sufficient margin	$\sigma_{min}(Y)-\|Y-UV^T\|_F>0$

S Arora, N Cohen, and E Hazan. "On the optimization of deep networks: Implicit acceleration by overparameterization." ICML 2018
S Tarmoun, G França, B D Haeffele, and R Vidal. "Understanding the dynamics of gradient flow in overparameterized linear models." ICML 2021
H Min, S Tarmoun, R Vidal, and E Mallada. "On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks." ICML 2021.

MAIN RESULTS

Orthogonal Initialization Leads to Min-norm Solution

The min-norm solution is given by

$$\hat{\Theta}=argmin_\Theta\{\|\Theta\|_F:\|Y-X\Theta\|_F=min_\Theta\|Y-X\Theta\|_F\}$$

Proposition 1. (Informal) *If at initialization, we have*

$$V(0)U_2^T(0)=0,\qquad U_1(0)U_2^T(0)=0,$$

then the gradient flow, if converges, finds the min-norm solution.

- Extension of "initializing Θ within the span of the input data leads to min-norm solution" in standard linear regression problem
- For single-hidden-layer model, initialization within the span of the data $VU_2^T=0$ is not sufficient

Under **random initialization + large hidden layer width k**, w.h.p.:

- There is sufficiently positive level of imbalance
- Orthogonality conditions are approximately satisfied

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

SUMMARY

We study the gradient flow on single-hidden-layer linear networks:

- **Sufficient Imbalance or Sufficient Margin** ⇒ exponential convergence
- **Orthogonal initialization** ⇒ exact min-norm solution
- **Random initialization + large network width** finds near min-norm solution efficiently

Future work:

- Deep Linear Networks and potentially nonlinear networks