Learning Dynamics of Deep Linear Networks Beyond the Edge of Stability

Avrajit Ghosh*[1], Soo Min Kwon*[2], Rongrong Wang[1], Saiprasad Ravishankar[1], Qing Qu[2]

[1] Computational Mathematics Science and Engineering, Michigan State University, [2] Department of EECS, University of Michigan

Motivation

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \cdot \nabla_{\mathbf{w}} f(\mathbf{w})$$

- How does the step-size $\eta > 0$ affect learning dynamics?
- What kind of solutions are found when η is large?
- GD typically occurs at Edge of Stablity (EoS) [1], where sharpness $\|\nabla^2 f(\boldsymbol{w})\|_2$ hovers about $2/\eta$.

Our Contributions

We focus on deep linear networks (DLNs) at EoS and show that:

- Oscillations only occur within a low-dimensional subspace.
- Subspace dimension depends on the step-size.
- DLN oscillates towards and periodically about the flattest (balanced) minima.
- Conservation law in DLNs (balancing) breaks at EoS.

Loss Optimization and Initialization

- Deep Matrix Factorization Loss (where $\text{rank}(\boldsymbol{M}_\star) = r$):

$$\underset{\boldsymbol{\Theta}}{\text{argmin}}\, f(\boldsymbol{\Theta}) = \frac{1}{2}\|\mathbf{W}_L \cdot \ldots \cdot \mathbf{W}_1 - \mathbf{M}_\star\|_{\mathsf{F}}^2.$$

- Update using GD with $\eta > 0$:

$$\mathbf{W}_\ell(t) = \mathbf{W}_\ell(t-1) - \eta \cdot \nabla_{\mathbf{W}_\ell} f(\boldsymbol{\Theta}(t-1)), \quad \forall \ell \in [L].$$

- Initialization ($\alpha > 0$ is small):

$$\mathbf{W}_L(0) = \mathbf{0}, \quad \mathbf{W}_\ell(0) = \alpha \mathbf{I}_d, \quad \forall \ell \in [L-1],$$

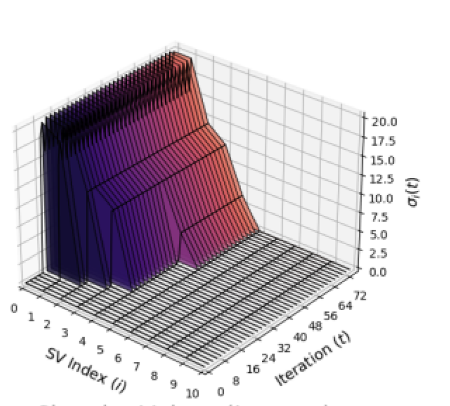
Implicit Bias in Singular Vectors of DLN:

Singular vectors remain static:

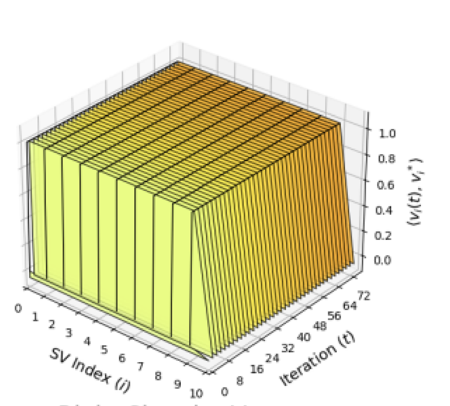
$$\boldsymbol{W}_L(t) = \boldsymbol{U}_\star \boldsymbol{\Sigma}_L(t) \boldsymbol{V}_\star^\top, \quad \boldsymbol{W}_\ell(t) = \boldsymbol{V}_\star \boldsymbol{\Sigma}_\ell(t) \boldsymbol{V}_\star^\top.$$

This simplifies the loss:

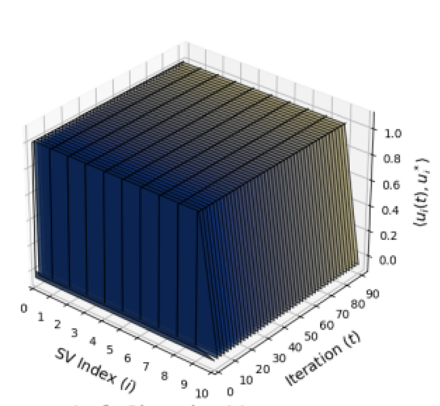
$$\frac{1}{2}\|\boldsymbol{W}_{L:1}(t) - \boldsymbol{M}^\star\|_{\mathsf{F}}^2 = \frac{1}{2}\sum_{i=1}^{r}\left(\sigma_i(\boldsymbol{\Sigma}_{L:1}(t)) - \sigma_{\star,i}\right)^2, \tag{1}$$



Singular Values (Large η)

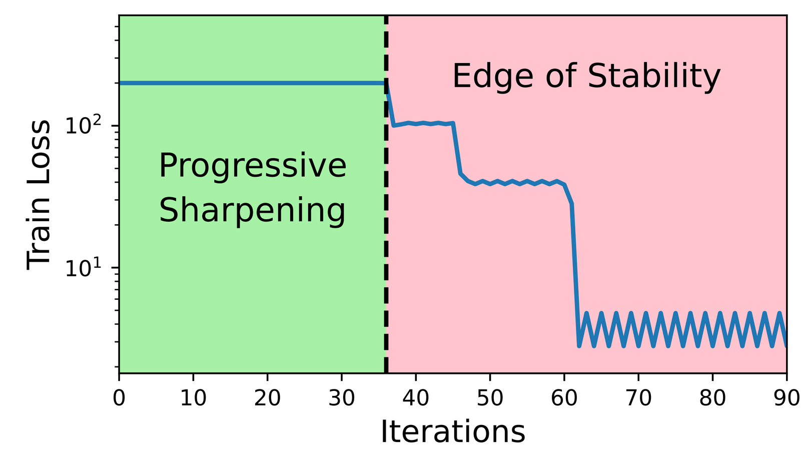


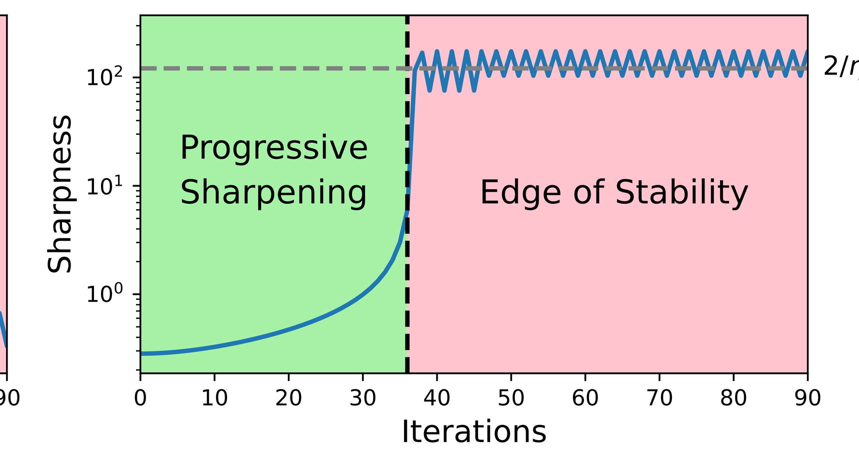
Right Singular Vectors



Left Singular Vectors

Progressive Sharpening and EoS in DLNs





Broken Conservation Law: Balancing

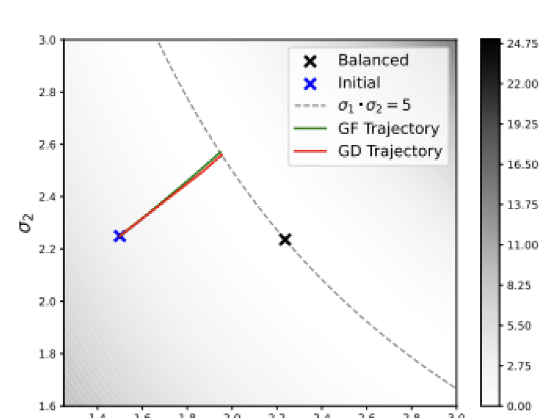
Theorem (Balancing of Singular Values). Consider the simplified loss in (1) and define $S_i := L\sigma_{\star,i}^{2-\frac{2}{L}}$. If we run GD with $\eta > \frac{2}{S_i}$ and α satisfies

$$0 < \alpha < \left(\ln\left(\frac{2\sqrt{2}}{\eta S_i}\right) \cdot \frac{\sigma_{\star,i}^{4/L}}{L^2 \cdot 2^{\frac{2L-3}{L}}}\right)^{1/4},$$

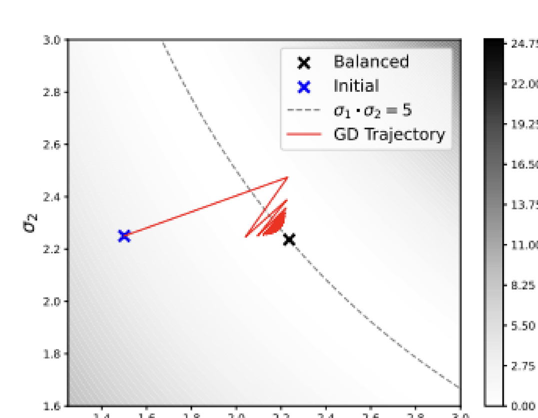
then for all $\ell \in [L-1]$, there exists a $c \in (0,1]$:

$$\left|\sigma_{L,i}^2(t+1) - \sigma_{\ell,i}^2(t+1)\right| < c \cdot \left|\sigma_{L,i}^2(t) - \sigma_{\ell,i}^2(t)\right|.$$

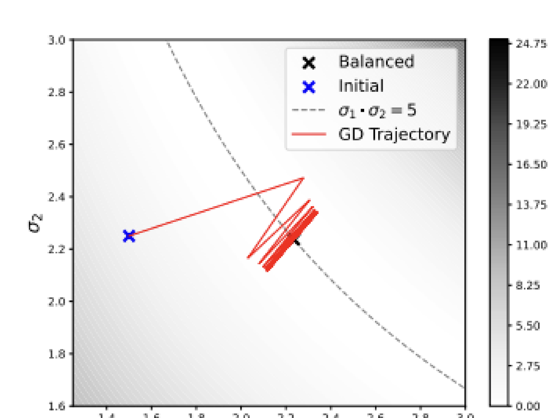
$$\text{Example: } f(\sigma_1, \sigma_2) = \frac{1}{2}(\sigma_2 \cdot \sigma_1 - \sigma_*)^2$$



(a) GF and Stable GD



(b) GD at EOS



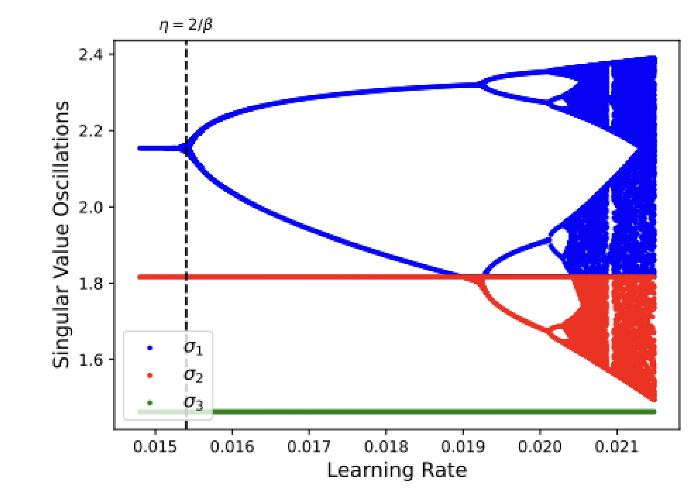
(c) GD Beyond EOS

- Gradient flow ($\eta \to 0$): $|\sigma_1^2(t) - \sigma_2^2(t)|$ stays constant.

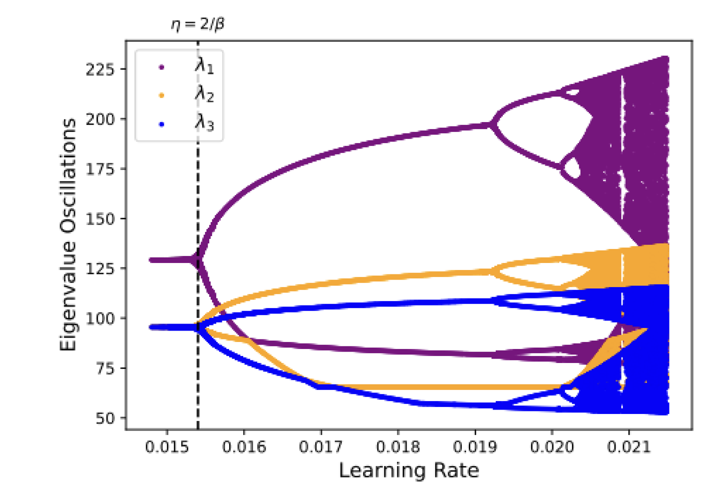
Define $S_i := L\sigma_{\star,i}^{2-\frac{2}{L}}$ to be stability limit:

- Gradient descent at EoS ($\eta < 2/S_i$): $|\sigma_1^2(t) - \sigma_2^2(t)|$ monotonically reduces to $\mathcal{O}(\alpha)$, where α is initialization scale.
- Gradient descent beyond EoS ($\eta > 2/S_i$): $|\sigma_1^2(t) - \sigma_2^2(t)|$ monotonically reduces to zero due to oscillations!

Period Doubling Route to Chaos in DLNs



Singular Values of Weights



Eigenvalues of Hessian

We focus on the period-2 case! Here $\beta = L\sigma_*^{2-\frac{2}{L}}$ (i.e., the stability limit).

Periodic Subspace Oscillations Beyond EoS

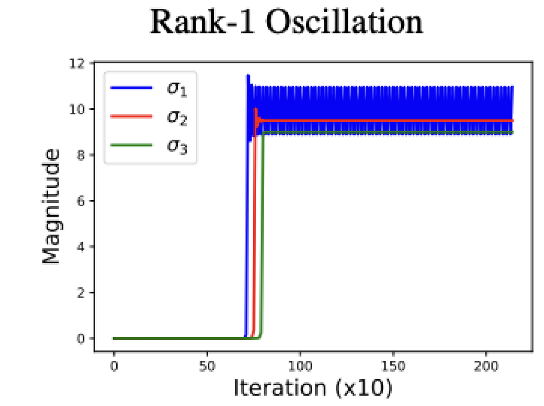
Theorem (Rank-p Periodic Oscillations). Let $\boldsymbol{M}_* = \boldsymbol{U}_*\boldsymbol{\Sigma}_*\boldsymbol{V}_*^\top$. Define $S_p := L\sigma_{\star,p}^{2-\frac{2}{L}}$. If we run GD with $\eta > \frac{2}{S_p}$, then the top-p singular values of the end-to-end DLN oscillate in a 2-period orbit ($j \in \{1,2\}$) around the balanced minimum and admit the following decomposition:

$$\boldsymbol{W}_{L:1} = \underbrace{\sum_{i=1}^{p} \rho_{i,j}^L \cdot \boldsymbol{u}_{\star,i}\boldsymbol{v}_{\star,i}^\top}_{\text{oscillation subspace}} + \underbrace{\sum_{k=p+1}^{d} \sigma_{\star,k} \cdot \boldsymbol{u}_{\star,k}\boldsymbol{v}_{\star,k}^\top}_{\text{stationary subspace}}, \quad j \in \{1,2\} \tag{2}$$

where $\rho_{i,1} \in \left(0, \sigma_{\star,i}^{1/L}\right)$ and $\rho_{i,2} \in \left(\sigma_{\star,i}^{1/L}, (2\sigma_{\star,i})^{1/L}\right)$ are the two real roots of a $(2L-2)(2L-1)$ order polynomial.

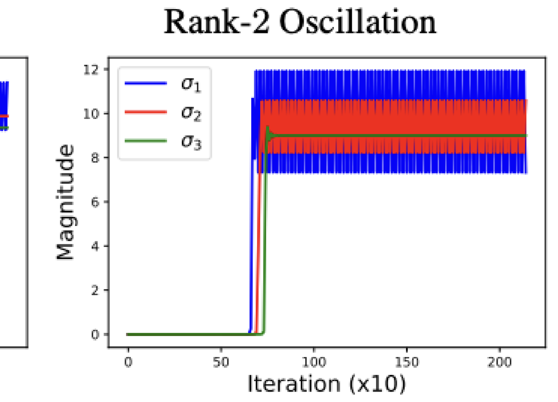
$$\text{Example: } \text{rank}(\mathbf{M}_* \in \mathbb{R}^{10\times 10}) = 3$$

Rank-1 Oscillation



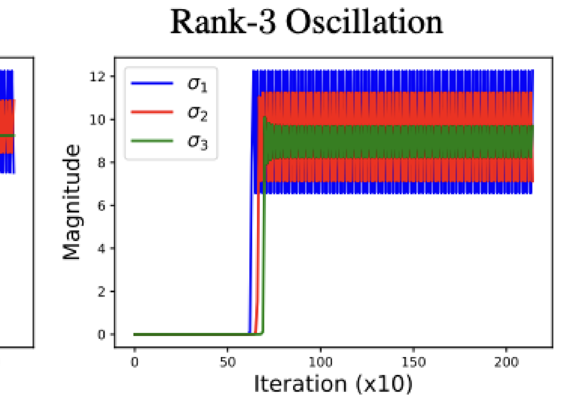
$\frac{2}{S_2} > \eta > \frac{2}{S_1}$

Rank-2 Oscillation



$\frac{2}{S_3} > \eta > \frac{2}{S_2}$

Rank-3 Oscillation



$\frac{2}{S_4} > \eta > \frac{2}{S_3}$

- i-th singular value oscillates in a two-period orbit if $\eta > \frac{2}{S_i}$
- Oscillation range depends on how large the learning rate is!
- Does all Symmetry induced Conservation laws break at Edge of Stability?

References

[1] Cohen et al., "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability". ICLR 2021.