Large Catapults in Momentum Gradient Descent with Warmup: An Empirical Study

Prin Phunyaphibarn*[1*], *Junghyun Lee*[2*], Bohan Wang[3,4], Huishuai Zhang[4], Chulhee Yun[2]

[1] Department of Mathematical Sciences, KAIST
[2] Kim Jaechul Graduate School of AI, KAIST
[3] University of Science and Technology of China
[4] Microsoft Research Asia

NEURAL INFORMATION PROCESSING SYSTEMS

KAIST AI Kim Jaechul Graduate School

OptiML Optimization & Machine Learning

OSI Optimization and Statistical Inference LAB

Microsoft Research 微软亚洲研究院

中国科学技术大学 University of Science and Technology of China

Contributions

- We provide empirical evidence suggesting that gradient descent (GD) with *momentum* with *learning rate warmup* induces a **large catapult** (compared to vanilla GD).

  → larger sharpness reduction → ***flatter minima***
- We show this holds for a wide range of settings.
- We relate this to the *self-stabilization* mechanism (Damian et al., 2023).

Preliminaries

- **Heavy-ball momentum (PHB)**:

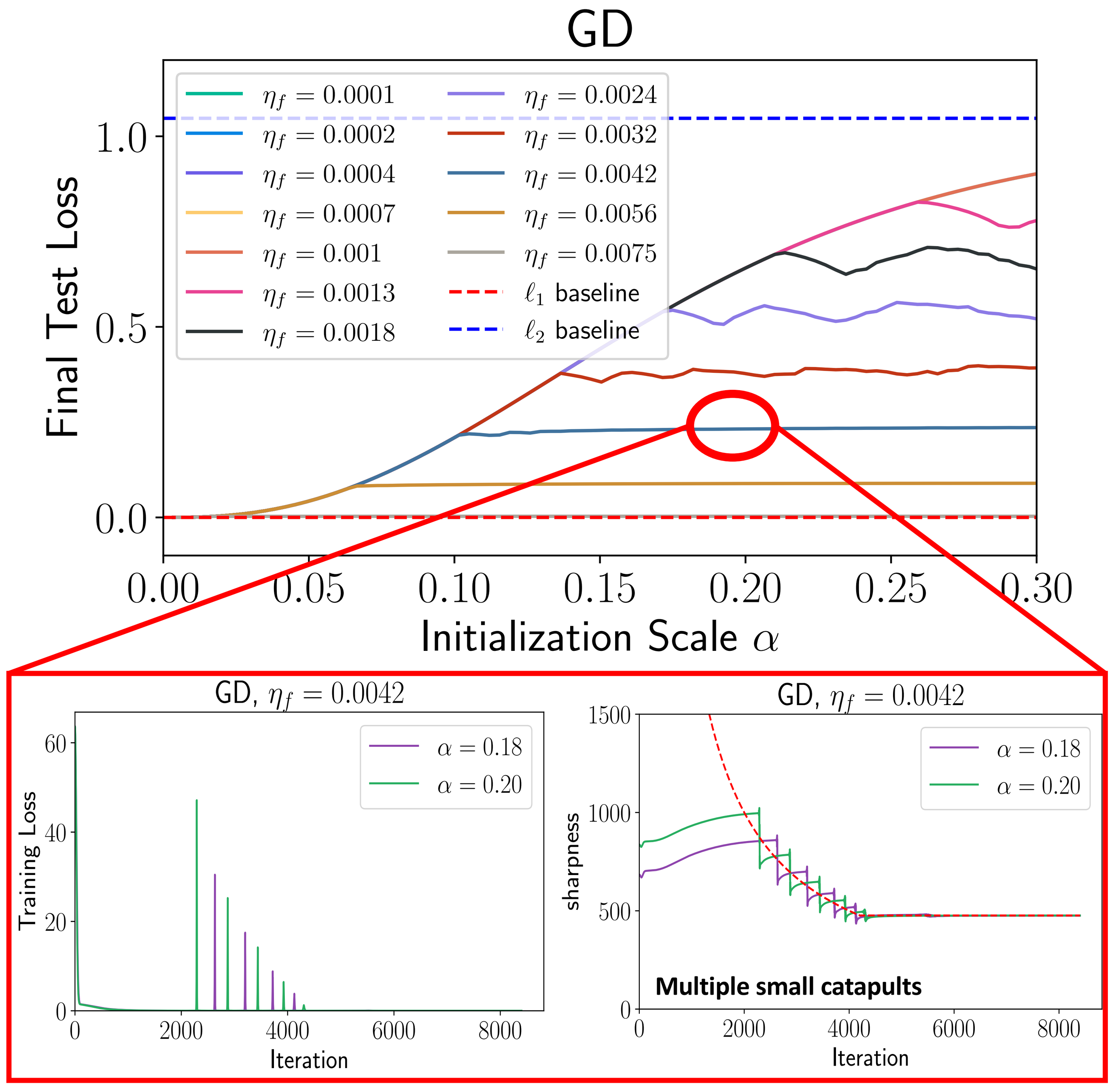
$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla f(\boldsymbol{w}_t) + \beta(\boldsymbol{w}_t - \boldsymbol{w}_{t-1})$$

  - η_t is the learning rate (possibly scheduled)
  - $\beta \in [0, 1)$ is the momentum parameter
- **Maximum stable sharpness (MSS):** $\frac{2(1+\beta)}{\eta_t}$
  - Minima with sharpness above MSS are *unstable* (Cohen et al., 2021)
- **Linear warmup from η_i to η_f:** $\eta_t = \eta_i + \frac{\eta_f - \eta_i}{T_{warmup}} t$
  - This allows for stable training with large learning rate η_f

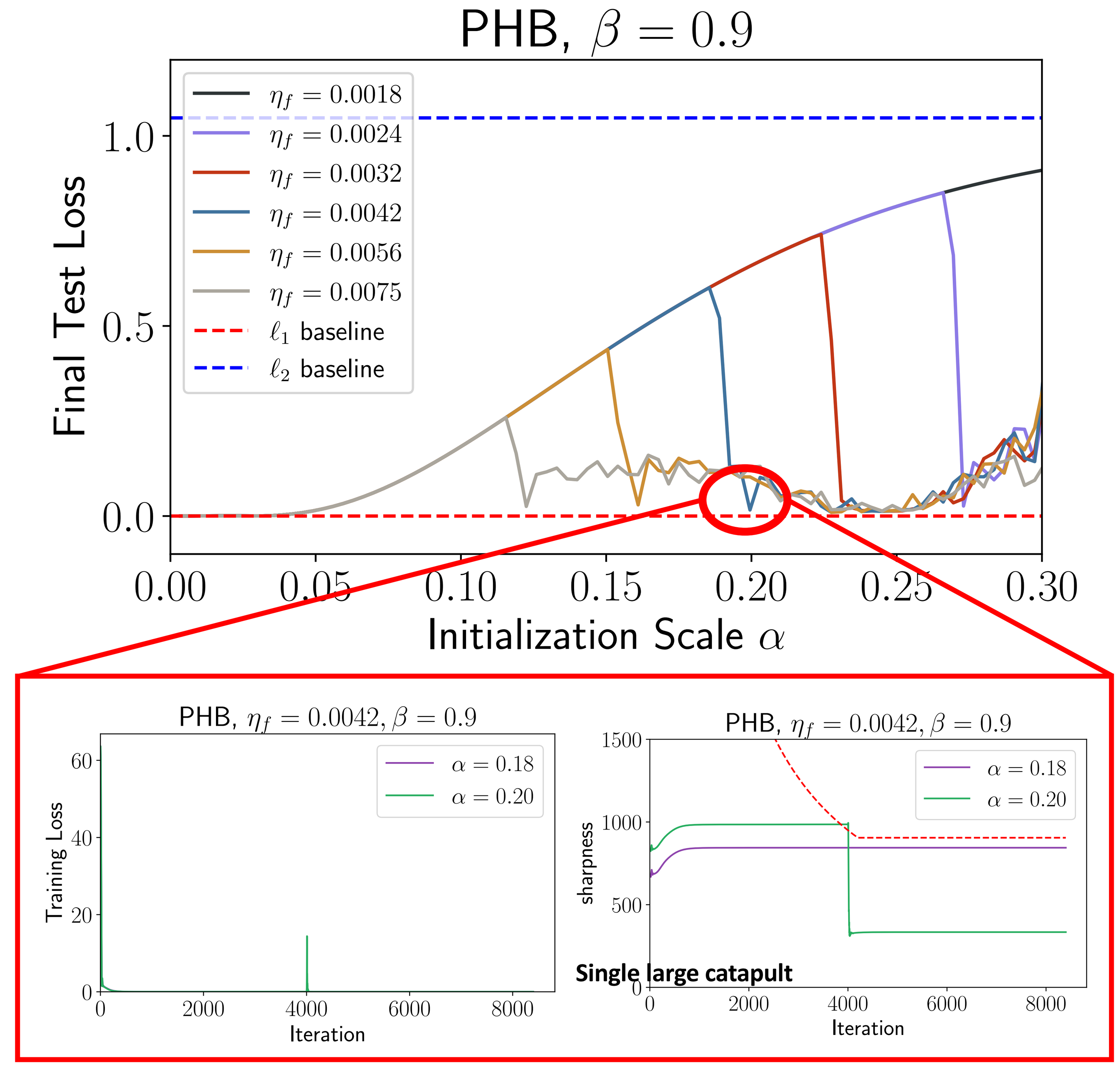
Motivation: Linear Diagonal Networks

Linear Diagonal Networks (LDNs). $f(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) := \langle \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}, \boldsymbol{x} \rangle = \langle \boldsymbol{w}, \boldsymbol{x} \rangle, \quad \boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha \cdot \mathbf{1}$

GD



PHB, $\beta = 0.9$



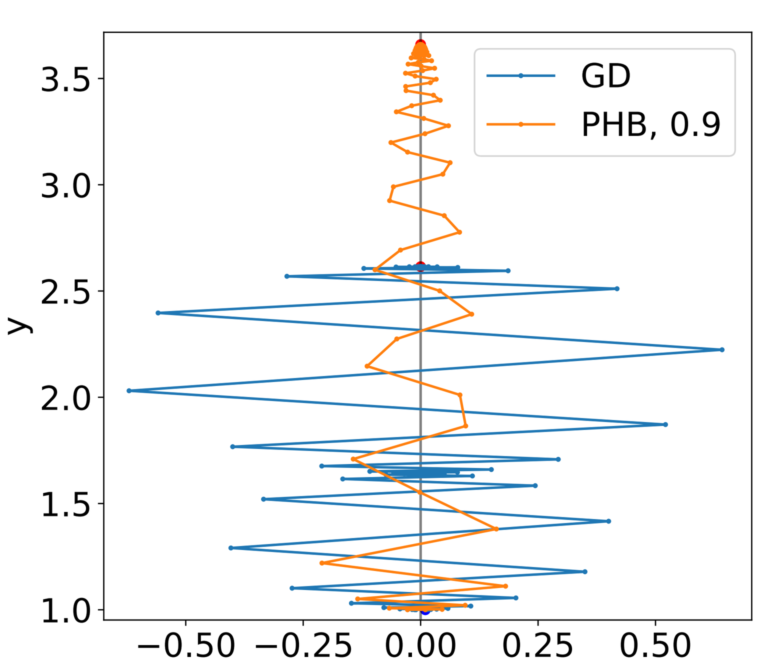
- Final test loss increases with α until saturation, consistent with the observations of Nacson et al. (2022).
- Sharpness closely follows the MSS curve with multiple small catapults.
- After certain α, the final test loss suddenly decreases due to a ***large catapult.***
- Sharpness significantly deviates from the MSS after a ***large catapult.***

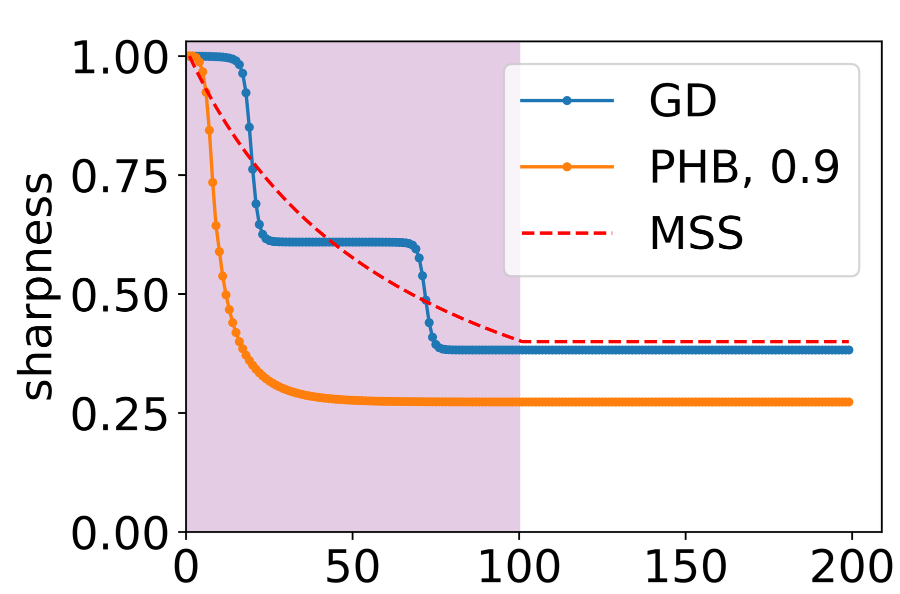
Toy Example

- Consider the following toy loss function:

$$f(x, y) = \frac{x^2}{2y}, \quad y > 0$$

- Trajectory & sharpness plots of GD vs. PHB:





- Resembles *self-stabilization* (Damian et al., 2023):

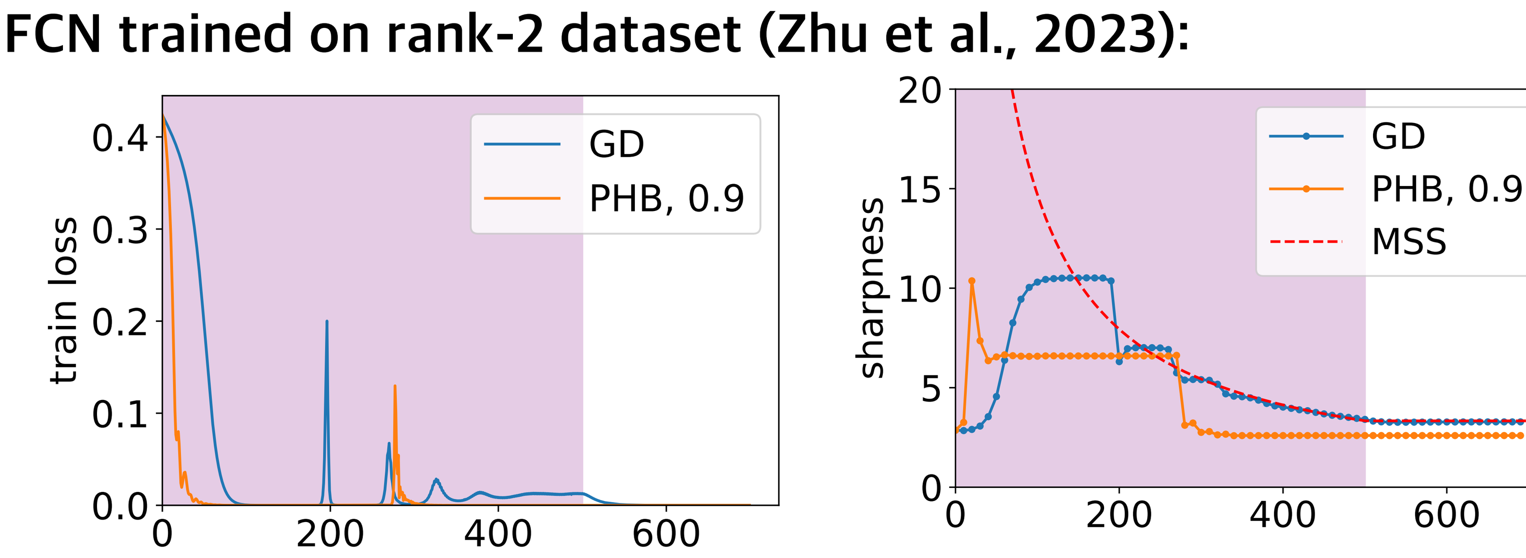
1) **Progressive Sharpening***. Stable training, Sharpness increases
2) **Blowup.** Sharpness > MSS, divergent dynamics
3) **Self-Stabilization.** Movement in $+y$ direction stabilizes dynamics in the x direction and decreases sharpness
4) **Return to Stability.** Sharpness < MSS

*This stage may not occur depending on the scenario (e.g., initialization).

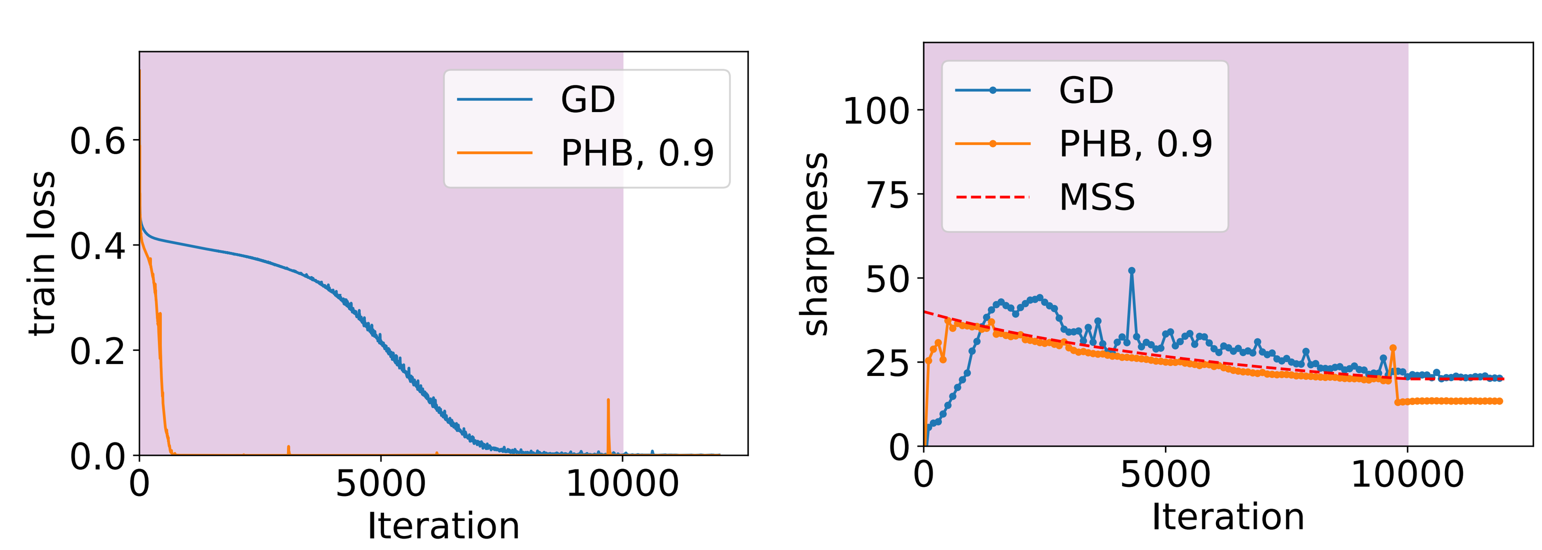
Momentum prolongs self-stabilization effect in the direction of negative gradient of the sharpness

Nonlinear Networks

FCN trained on rank-2 dataset (Zhu et al., 2023):



ResNet20 trained on 1k subset of CIFAR10:



References

A. Damian, E. Nichani, and J. D. Lee. "Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability." In ICLR 2023.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability." In ICLR 2021.

L. Zhu, C. Liu, A. Radhakrishnan, and M. Belkin. "Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning." In arXiv:2306.04815, 2023.

M. S. Nacson, K. Ravichandran, N. Srebro, and D. Soudry. "Implicit Bias of the Step Size in Linear Diagonal Networks." In ICML 2022.