Gradient Descent with Polyak's Momentum Finds Flatter Minima via Large Catapults

Prin Phunyaphibarn[*†] PRIN10517@KAIST.AC.KR
Department of Mathematical Sciences, KAIST, Daejeon, Republic of Korea

Junghyun Lee[*] JH_LEE00@KAIST.AC.KR
Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea

Bohan Wang BHWANGFY@GMAIL.COM
USTC & Microsoft Research Asia

Huishuai Zhang ZHANGHUISHUAI@PKU.EDU.CN
Wangxuan Institute of Computer Technology, Peking University, Beijing, China

Chulhee Yun CHULHEE.YUN@KAIST.AC.KR
Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea

Abstract

Although gradient descent with Polyak's momentum is widely used in modern machine and deep learning, a concrete understanding of its effects on the training trajectory remains elusive. In this work, we empirically show that for linear diagonal networks and nonlinear neural networks, momentum gradient descent with a large learning rate displays large catapults, driving the iterates towards much flatter minima than those found by gradient descent. We hypothesize that the large catapult is caused by momentum "prolonging" the self-stabilization effect [11]. We provide theoretical and empirical support for our hypothesis in a simple toy example and empirical evidence supporting our hypothesis for linear diagonal networks.

1. Introduction

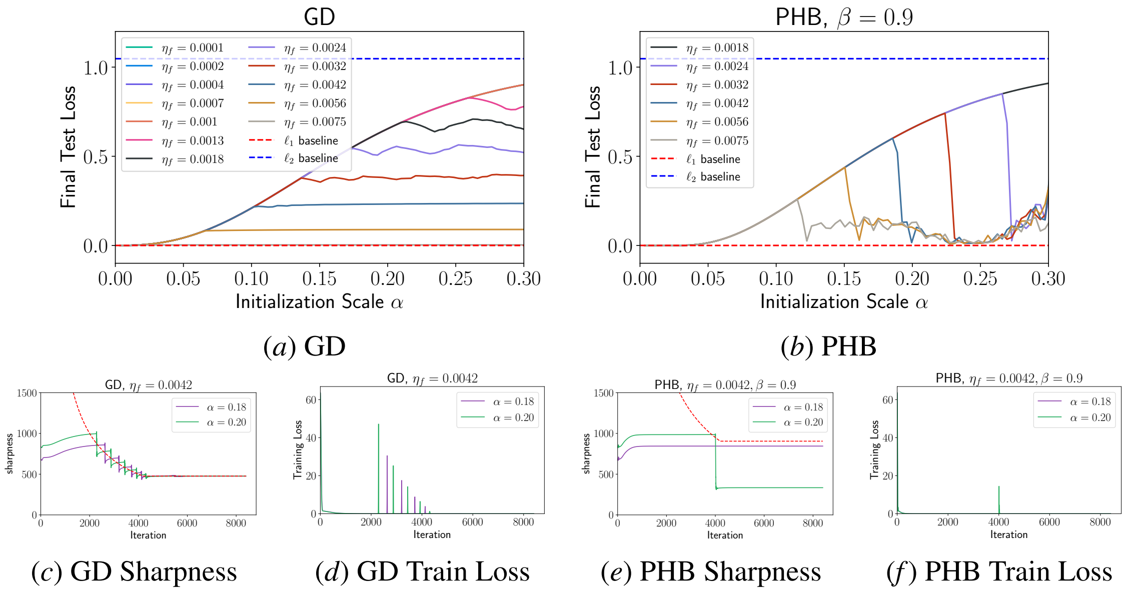
Although momentum is one of the most widely used and crucial component for training modern neural networks [42], our understanding of the effects of momentum on neural network training dynamics is still lacking. Throughout the paper, for a loss function $\mathcal{L}(\boldsymbol{w})$, we consider Polyak's heavy-ball momentum (PHB; Polyak [40]) method given as follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{w}_t) + \beta(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}), \tag{1}$$

where η_t denotes the learning rate that may change over time and $\beta \in [0, 1)$ is the momentum parameter where $\beta = 0$ for gradient descent (GD). From here on, we will refer to Polyak's momentum simply as "momentum" or "PHB". For momentum, we are not even sure yet why and when it accelerates (stochastic) gradient descent [14, 15, 25], and perhaps more importantly, how it changes the implicit bias and thus the resulting model's generalization capability [16, 23, 43, 44].

Many works consider training dynamics under the large learning rate regime. One such work closely related to ours is the *catapult mechanism* introduced by Lewkowycz et al. [28]. Classical

[*] Equal contributions
[†] Work done as an undergraduate intern at KAIST AI.



(*a*) GD (*b*) PHB

(*c*) GD Sharpness (*d*) GD Train Loss (*e*) PHB Sharpness (*f*) PHB Train Loss

Figure 1: Experiments following the same setting as Nacson et al. [35]. In (a) and (b), "ℓ_1 baseline" and "ℓ_2 baseline" respectively stand for the solution with the minimal ℓ_1 norm and the solution with the minimal ℓ_2 norm to the regression problem. We use $\beta = 0.9$ for PHB.

optimization theory suggests that if the *sharpness* $S := \lambda_{max}(\nabla^2 \mathcal{L})$, defined as the maximum eigenvalue of the Hessian of the training loss, is larger than a certain threshold, then training should be unstable and divergent. We refer to this threshold as the *maximum stable sharpness (MSS)*, and it is equal to $2/\eta$ for GD and $2(1+\beta)/\eta$ for momentum [18]. However, Lewkowycz et al. [28] show that rather than diverging, GD initialized with a large learning rate η satisfying $2/S < \eta < 4/S$ displays a loss spike and a simultaneous sharpness drop, which drives the iterates towards a flat region with stable sharpness $S < 2/\eta$. Following this observation, throughout this paper, we define a **catapult** as *a drastic sharpness reduction coupled with a single spike in the training loss*. For more discussion on related works, please refer to Appendix A.

2. Large Catapults in Momentum Gradient Descent

2.1. Motivating Example: Linear Diagonal Networks

The linear diagonal network (LDN) is known to be one of the simplest non-linear models that display rich and non-trivial implicit bias [47] while still being mathematically tractable; see Appendix A for related work on the implicit bias of LDNs. Here, we focus on the depth-2 linear diagonal network defined by

$$f(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) := \langle \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}, \boldsymbol{x} \rangle, \quad \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \tag{2}$$

where $\boldsymbol{x}$ is the input vector, $(\boldsymbol{u}, \boldsymbol{v})$ are the trainable parameters, and $\boldsymbol{w} := \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}$ is the linear coefficient vector. We investigate the implicit bias effect of momentum in LDNs in the sparse regression experiment of Nacson et al. [35] over various initializations $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha\mathbf{1}$ and learning rate warmups from $\eta_i = 10^{-8}$ to η_f; the precise setting is deferred to Appendix B.

Results. Over the varying initialization scales α and (final) learning rates η_f, we report the final resulting test losses in Figure 1. Compared to GD, whose final test loss saturates after some α as reported in Nacson et al. [35] (Figure 1(*a*)), it is clear that *PHB displays a fundamentally different implicit bias in the large learning rate regime.* For PHB, the final test loss initially increases monotonically as α increases like GD. However, once α becomes larger than some threshold $\bar{\alpha}(\eta_f)$ (dependent on η_f), the test loss sharply drops close to zero. We also note that $\bar{\alpha}(\eta_f)$ appears to decrease as η_f increases.[1]

Momentum Induces Larger Catapults. To find the cause for our observed phenomena, we plot the evolution of sharpness for PHB in Figure 1(*e*) at the α and η at which the test loss suddenly drops. Note that as the warmup proceeds, the MSS decreases monotonically, and as soon as the sharpness of the iterates "touches" the MSS curve, it goes through a rapid sharpness reduction, coupled with a loss spike (Figure 1(*f*)), with the final sharpness being well below the MSS of the final learning rate. This is in contrast to GD which goes through an incremental, step-wise sharpness reduction (Figure 1(*c*)) with multiple loss spikes (Figure 1(*d*)), and the final sharpness stays just below the MSS corresponding to the final learning rate; this phenomenon for SGD was also observed in Zhu et al. [50]. Here, one could make an educated guess that momentum induces much larger catapults that bias the solution towards flatter minima.

One observation of note is that the MSS $\frac{2(1+\beta)}{\eta}$ for PHB is higher than the MSS $\frac{2}{\eta}$ of GD with the same η, so the α that causes a catapult for GD may not for PHB. With this observation in mind, from this point on, whenever we compare GD and PHB starting from the same initialization, we always match the MSS of GD and PHB by properly rescaling the momentum learning rates $\eta_{PHB} = (1+\beta)\eta_{GD}$ for fair comparison. We only specify the GD learning rate in the text as η.

2.2. Nonlinear Neural Networks

To show that the phenomenon is not limited to LDNs, we also conduct experiments with narrow and wide fully-connected networks (FCN) on 1k and 5k-datapoint subsets of CIFAR10 [27] for both MSE *and* cross entropy (CE) losses. The results are shown in Figure 2. It is clear that the catapults observed in PHB reduce the sharpness farther below the MSS than GD. Additional experiments, including results for ResNet20 and for $\beta = 0.99$, can be found in Appendices E.3 and E.4.

3. Why Large Catapult? Because Momentum Prolongs Self-Stabilization

In this section, we propose a hypothesis for the mechanism that momentum causes the large catapults. To do that, we first review the *self-stabilization* [11] mechanism of GD. We use $\boldsymbol{w}_{\max}(\boldsymbol{\theta}_\star)$ to denote the eigenvector corresponding to the sharpness S (leading eigenvalue of the training loss Hessian) at some minimum $\boldsymbol{\theta}_\star$. Self-stabilization consists of four stages:

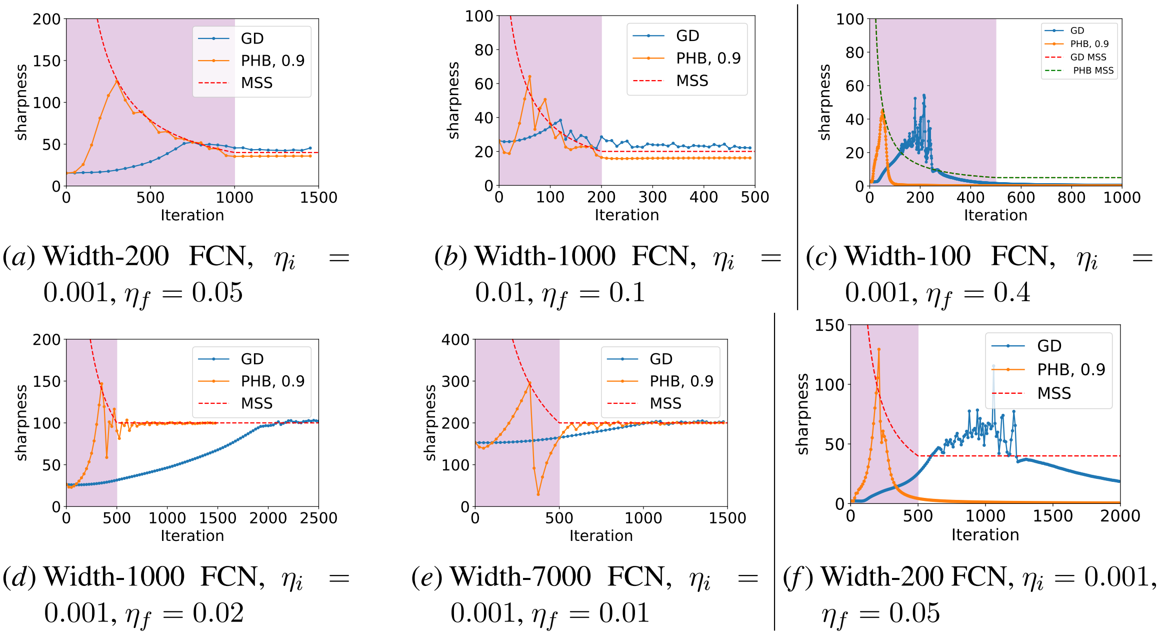
Stage 1 (Progressive Sharpening[2]). The sharpness of the iterates increases to reach the MSS.

Stage 2 (Blowup). Once the sharpness becomes larger than the MSS, the iterates oscillate and diverge along $\boldsymbol{w}_{\max}$. Then, the loss starts to increase sharply, depending on the amount of divergence.

Stage 3 (Self-Stabilization). *Simultaneously*, the divergence along $\boldsymbol{w}_{\max}$ induces a drift along the direction of $-\nabla S$ which decreases the sharpness.

[1] In fact, we can see in Figure 1(*b*) that the final test loss slightly increases (in a noisy fashion) with α after the sharp drop at $\bar{\alpha}(\eta_f)$. We attribute this phenomenon to *overshooting*; see Appendix E.2 for more discussion.

[2] Although Damian et al. [11] assume that PS always occurs, it may not for simple settings [9, 50].



(a) Width-200 FCN, $\eta_i = 0.001$, $\eta_f = 0.05$

(b) Width-1000 FCN, $\eta_i = 0.01$, $\eta_f = 0.1$

(c) Width-100 FCN, $\eta_i = 0.001$, $\eta_f = 0.4$

(d) Width-1000 FCN, $\eta_i = 0.001$, $\eta_f = 0.02$

(e) Width-7000 FCN, $\eta_i = 0.001$, $\eta_f = 0.01$

(f) Width-200 FCN, $\eta_i = 0.001$, $\eta_f = 0.05$

Figure 2: Neural Networks trained on (a-c) 1k and (d-f) 5k subset of CIFAR10 [27]. For (a,b,d,e), we use the MSE loss, and for (c,f) we the CE loss. All FCNs are 3-layer and use ReLU activation. The shaded region is the linear warmup period.

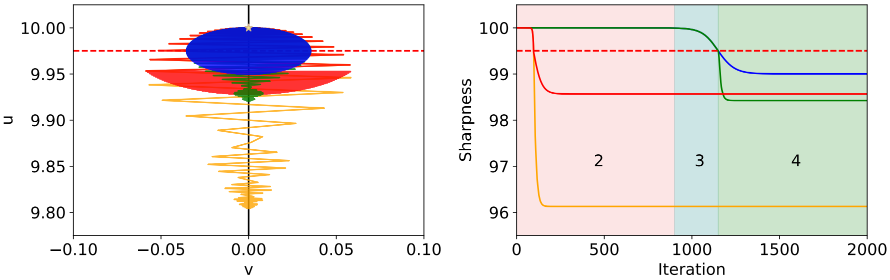


Figure 3: (Left) Trajectories of GD, PHB, GD → PHB, PHB → GD with $\beta = 0.9$, $\eta = (2 + \epsilon)/u_0^2$ where $\epsilon = 0.01$, $(u_0, v_0) = (10, 10^{-6})$, and no warmup. (Right) The self-stabilization stages for GD are highlighted and labeled in the sharpness plot. The MSS is shown as the red dotted line.

Stage 4 (Return to Stability). When the sharpness drops below the MSS, the oscillation in the $\boldsymbol{w}_{\max}$ direction dampens, and dynamics become stable again.

One can draw a connection between self-stabilization and catapult by viewing *a catapult as a single round of self-stabilization (Stages 2–4).* Drawing from known intuitions for momentum dynamics [18, 34, 37], we propose the following hypothesis:

Hypothesis 1 *Polyak's momentum prolongs self-stabilization in the following sense:*

1. *As the iterates oscillate and diverge along $\mathbf{w}_{\max}$, the momentum in the direction of $-\nabla S$ "builds up" (Stages 2–3).*

2. *Even after the sharpness drops below the MSS, momentum prolongs oscillation along $\boldsymbol{w}_{\max}$, which in turn prolongs the movement in the $-\nabla S$ direction (Stage 4).*

3.1. Empirical Verification of Hypothesis 1 for ReLU Scalar Network

We now verify Hypothesis 1 in a simple ReLU scalar network by providing qualitative empirical results and theoretical characterization of the sharpness reduction for GD and PHB. Here, we consider the loss function $\mathcal{L}(u, v) = \frac{1}{2}u^2v^2\mathbb{1}[u \geq 0]$, where $\mathbb{1}[\cdot]$ is the indicator function. We consider constant learning rate $(1 + \beta)\eta$ to simplify the analysis and initialize close to an unstable minimum: (u_0, v_0) with $u_0^2 = {}^{(2+\epsilon)}/_{\eta}$ for some small $\epsilon, |v_0| \in (0, 1)$. In Appendix E.5, we provide additional empirical evidence for our hypothesis in LDNs.

The GD dynamics (blue in Figure 3) show the interplay between oscillation and stabilization, which is precisely the self-stabilization for GD [11]. The PHB dynamics (orange in Figure 3) shows that momentum does not simply accelerate the GD dynamics but shows a *qualitatively* different behavior, breaking the symmetry around $(\sqrt{{}^{2}/_{\eta}}, 0)$ that was present for GD.

To further empirically validate our Hypothesis 1, we consider two additional settings: GD → PHB and PHB → GD, where $A \to B$ means that we first run A, then switch to B once the sharpness crosses the MSS. If Hypothesis 1.1 (resp. 2) holds, then due to the effect of momentum before crossing the MSS, one would expect PHB → GD (resp. GD → PHB) to experience a larger sharpness reduction than GD. As shown in Figure 3, in increasing order of sharpness reduction, we have GD < PHB → GD < GD → PHB < PHB. This shows that although the effect of each part of our hypothesis is sufficient to display a larger sharpness reduction, the *combination* of these two factors results in an even larger sharpness reduction for PHB.

3.2. Theoretical Verification of Hypothesis 1 for ReLU Scalar Network

Let $\{(u_t, v_t)\}$ be the GD/PHB iterates with learning rate $\eta > 0$ and momentum parameter $\beta \in [0, 1)$. The proofs for all the statements provided here are deferred to Appendix D.

For that, we first introduce the following quantities:

$$\tau_u := \inf\left\{t \geq 0 : u_t^2 < \frac{2-\epsilon}{\eta}\right\}, \quad C_u := \frac{u_{\tau_u} - \beta u_{\tau_u - 1}}{1-\beta}, \quad C_v := \frac{1+\beta}{1-\beta} \textstyle\sum_{t=\tau_u}^{\infty} v_t^2. \tag{3}$$

The following theorem characterizes an upper bound on u_∞:

Theorem 2 *Suppose that* $\inf_{t\geq 0} u_t \geq 0$ *and* $\epsilon, |v_0| < 1$. *Then,* $u_\infty = \lim_{t\to\infty} u_t \leq \frac{C_u}{1+\eta C_v} =: \overline{u}_\infty$.

According to Theorem 2, the RHS decreases (i.e., larger displacement in the $-u$ direction) with larger C_v and smaller C_u. This observation lends support to Hypothesis 1 since C_u corresponds to the "build up" of momentum in the $-\nabla S$ direction from Stages 2-3 in Hypothesis 1.1, and C_v measures the movement in the $-\nabla S$ direction caused by divergence along $w_{\max}$ *after* the iterates cross the MSS (Stage 4) which captures the "prolonging" effects of momentum. We show that for our model, Theorem 2 is numerically tight; see Appendix D.4 for more discussions.

As a counterpart, we provide an asymptotic lower bound on u_∞ for *GD*:

Theorem 3 (Informal) *For GD* ($\beta = 0$)*, suppose that* $\inf_{t\geq 0} u_t \geq 0$, $\epsilon = o(1)$, $v_0^2 = \mathcal{O}(\epsilon)$, *and* $\eta = \Theta(1)$. *Then, we have that* $u_\infty \geq \sqrt{\frac{2}{\eta}} - \mathcal{O}(\sqrt{\epsilon})$.

For the significance of our theoretical results, we provide some discussions in Appendix D.5 and D.6.

4. Conclusion

In this paper, for the first time, we show that PHB with large learning rate induces large catapults, resulting in a much larger sharpness reduction than that of GD. We first provide empirical evidence for this on linear diagonal networks (LDNs) and nonlinear neural networks. We then hypothesize that the large catapult of PHB is caused by momentum *prolonging* self-stabilization [11]. We verify our hypothesis for ReLU scalar networks and LDNs and rigorously prove that it holds. This opens up numerous exciting future directions, which we defer to Appendix C.

Acknowledgements

We thank the anonymous reviewers for their helpful comments in improving the paper. We also thank Alex Damian, Eshaan Nichani, Jeremy Cohen, Jason Lee, and Molei Tao for helpful discussions at NeurIPS 2023 M3L Workshop. This project was funded by the 2023 Microsoft Research Asia Collaborative Research grant funded by Microsoft. J. Lee and C. Yun were supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)).

References

[1] Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 169–195. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/agarwala23b.html.

[2] Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 247–257. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ahn22a.html.

[3] Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability". In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=9cQ6kToLnJ.

[4] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-Aware Minimization Leads to Low-Rank Features. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=29WbraPk8U.

[5] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A Modern Look at the Relationship between Sharpness and Generalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 840–902. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/andriushchenko23a.html.

[6] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding Gradient Descent on the Edge of Stability in Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 948–1024. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

[7] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 468–477. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/azulay21a.html.

[8] David Barrett and Benoit Dherin. Implicit Gradient Regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3q5IqUrkcF.

[9] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

[10] Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive Gradient Methods at the Edge of Stability. *arXiv preprint arXiv:2207.14484*, 2022. URL https://arxiv.org/abs/2207.14484.

[11] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

[12] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

[13] Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over Diagonal Linear Networks: Implicit Regularisation, Large Stepsizes and Edge of Stability. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=uAyElhYKxg.

[14] Jingwen Fu, Bohan Wang, Huishuai Zhang, Zhizheng Zhang, Wei Chen, and Nanning Zheng. When and Why Momentum Accelerates SGD: An Empirical Study. *arXiv preprint arXiv:2306.09000*, 2023. URL https://arxiv.org/abs/2306.09000.

[15] Swetha Ganesh, Rohan Deb, Gugan Thoppe, and Amarjit Budhiraja. Does Momentum Help in Stochastic Optimization? A Sample Complexity Analysis. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of*

Machine Learning Research, pages 602–612. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/ganesh23a.html.

[16] Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. Implicit regularization in Heavy-ball momentum accelerated stochastic gradient descent. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZzdBhtEH9yB.

[17] Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A Loss Curvature Perspective on Training Instabilities of Deep Learning Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=OcKMT-36vUs.

[18] Gabriel Goh. Why Momentum Really Works. *Distill*, 2017. doi: 10.23915/distill.00006. URL http://distill.pub/2017/momentum.

[19] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r14EOsCqKX.

[20] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv preprint arXiv:1706.02677*, 2017. URL https://arxiv.org/abs/1706.02677.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1):1–42, 01 1997. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

[22] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8580–8589. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[23] Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9965–10040. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jelassi22a.html.

[24] Dayal Singh Kalra and Maissam Barkeshli. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=Al9yglQGKj.

[25] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for Stochastic Optimization. In *International Conference*

on Learning Representations, 2018. URL https://openreview.net/forum?id=rJTutzbA-.

[26] Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient Descent Monotonically Decreases the Sharpness of Gradient Flow Solutions in Scalar Networks and Beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17684–17744. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kreisler23a.html.

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[28] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. URL https://arxiv.org/abs/2003.02218.

[29] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgz2aEKDr.

[30] Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. Dying ReLU and Initialization: Theory and Numerical Examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020. ISSN 1991-7120. doi: https://doi.org/10.4208/cicp.OA-2020-0165. URL http://global-sci.org/intro/article_detail/cicp/18393.html.

[31] Bochen Lyu and Zhanxing Zhu. On the Role of Momentum in the Implicit Bias of Gradient Descent for Diagonal Linear Networks, 2024. URL https://openreview.net/forum?id=014CgNPAGy.

[32] David Meltzer and Junyu Liu. Catapult Dynamics and Phase Transitions in Quadratic Nets. *arXiv preprint arXiv:2301.07737*, 2023. URL https://arxiv.org/abs/2301.07737.

[33] Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy. In *Advances in Neural Information Processing Systems*, volume 33, pages 22182–22193. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fc2022c89b61c76bbef978f1370660bf-Paper.pdf.

[34] Michael Muehlebach and Michael I. Jordan. Optimization with Momentum: Dynamical, Control-Theoretic, and Symplectic Perspectives. *Journal of Machine Learning Research*, 22(73):1–50, 2021. URL http://jmlr.org/papers/v22/20-207.html.

[35] Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit Bias of the Step Size in Linear Diagonal Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16270–16295. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nacson22a.html.

[36] Hristo Papazov, Scott Pesme, and Nicolas Flammarion. Leveraging Continuous Time to Understand Momentum When Training Diagonal Linear Networks. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3556–3564. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/papazov24a.html.

[37] Fabian Pedregosa. A Hitchhiker's Guide to Momentum. In *The Second Blogpost Track at ICLR 2023*, 2023. URL https://openreview.net/forum?id=kUk79nBY__2.

[38] Scott Pesme and Nicolas Flammarion. Saddle-to-Saddle Dynamics in Diagonal Linear Networks. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=iuqCXg1Gng.

[39] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/f4661398cb1a3abd3ffe58600bf11322-Paper.pdf.

[40] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

[41] Minhak Song and Chulhee Yun. Trajectory Alignment: Understanding the Edge of Stability Phenomenon via Bifurcation Theory. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=PnJaA0A8Lr.

[42] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

[43] Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does Momentum Change the Implicit Regularization on Separable Data? In *Advances in Neural Information Processing Systems*, volume 35, pages 26764–26776. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ab3f6bbe121a8f7a0263a9b393000741-Paper-Conference.pdf.

[44] Li Wang, Zhiguo Fu, Yingcong Zhou, and Zili Yan. The Implicit Regularization of Momentum Gradient Descent in Overparametrized Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10149–10156, Jun. 2023. doi: 10.1609/aaai.v37i8.26209. URL https://ojs.aaai.org/index.php/AAAI/article/view/26209.

[45] Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *arXiv preprint arXiv:2310.17087*, 2023. URL https://arxiv.org/abs/2310.17087.

[46] Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability. In *Advances in Neural Information Processing Systems*, volume 35, pages 9983–9994. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/40bb79c081828bebdc39d65a82367246-Paper-Conference.pdf.

[47] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

[48] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit Bias of Gradient Descent for Logistic Regression at the Edge of Stability. In *Advances in Neural Information Processing Systems*, volume 36, pages 74229–74256. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=IT9mWLYNpQ.

[49] Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large Stepsize Gradient Descent for Logistic Loss: Non-Monotonicity of the Loss Improves Optimization Efficiency. *arXiv preprint arXiv:2402.15926*, 2024. URL https://arxiv.org/abs/2402.15926.

[50] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024. URL https://arxiv.org/abs/2306.04815.

[51] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Quadratic models for understanding catapult dynamics of neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PvJnX3dwsD.

[52] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p7EagBsMAEO.

Contents

Appendix A. Related Works

Catapults in (S)GD. Lewkowycz et al. [28] are the first to describe the catapult mechanism of GD, the phenomenon of momentary spikes in training loss resulting in lower sharpness. They then analytically prove that catapults occur for $f = d^{-1/2}\boldsymbol{u}^T\boldsymbol{v}$, where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, when d is sufficiently large. Since then, the theoretical analysis of the catapult mechanism has been extended to other models such as (neural) quadratic models [1, 32, 51] and matrix factorization [43, 45] (from the perspective of "balancing" effect). On the empirical side, Zhu et al. [50] study the loss spikes that occur during the catapults of SGD by decomposing the loss into components corresponding to different eigenspaces of the neural tangent kernel (NTK; Jacot et al. [22]) and show that catapults improve generalization through feature learning. Recent work by Kalra and Barkeshli [24] explores the training dynamics of SGD for neural networks for various settings and uncovers four distinct regimes controlled by critical values of the sharpness. Despite an abundance of works on the catapult mechanism, to the best of our knowledge, we are the first to report the substantial differences in the catapult mechanism and sharpness reduction between GD and PHB.

Edge of Stability. A phenomenon related to the catapult phenomenon that has garnered much attention in recent years is the *edge of stability* or *EoS* phenomenon. Cohen et al. [9] report two surprising phenomena that consistently occur during neural network training. First, when the sharpness is below the MSS, the sharpness tends to increase due to a phenomenon known as *progressive sharpening (PS)*. Generally, progressive sharpening will increase the sharpness until it is above the MSS. However, instead of diverging as classical optimization theory suggests, the sharpness oscillates around the MSS while the training loss decreases non-monotonically. In a sense, the EoS phenomenon can be viewed as a cycle of a catapult and PS: once PS drives the sharpness above the MSS, a catapult occurs and decreases the sharpness below the MSS while the training loss spikes. Once the sharpness is lower than the MSS, PS increases it again, and the cycle continues. This cycle can be viewed together as sharpness oscillation and a non-monotonic decrease in training loss as reported by Cohen et al. [9]. Subsequent works have formalized this intuition [1, 11].

Many works have been devoted to the theoretical analysis of PS and EoS [2, 3, 6, 11, 41, 46, 48, 49, 52] as well as their systematic empirical analyses [9, 10, 17]. Ahn et al. [3] rigorously characterize the EoS phenomenon in a single-neuron network. They show that the single-neuron training dynamics under the large learning rate regime display oscillatory "bouncing" behavior accompanied by a decrease in sharpness so that the sharpness converges below the MSS. Song and Yun [41] extend the results of Ahn et al. [3] to two-layer fully connected linear networks trained using a single data point. Damian et al. [11] explain the EoS phenomenon by proving a "self-stabilization" property in GD. By utilizing cubic Taylor expansions of the loss function, they prove that if the sharpness is above the MSS, GD will begin to oscillate and diverge in the leading eigenvector direction. However, the oscillations induce a "self-stabilization" effect which moves the iterates in the $-\nabla S$ direction thus reducing the sharpness and stabilizing the dynamics. This interplay between the oscillations and self-stabilization result in the EoS behavior. There are also works explaining the effect of optimization tricks by considering their interaction with EoS [14, 17]. Lastly, we note that although the main focus of Cohen et al. [9] is on GD, the authors also include PHB experiments, mostly in their Appendix N. Their experiments also display occasional larger catapults, but this difference is not highlighted.

Implicit Bias of LDNs. The training dynamics and implicit bias of the LDN have been rigorously investigated in the literature. Still, most works focus on the continuous regime with vanishing learning rate, such as the (stochastic) gradient flow [7, 33, 38, 39, 47]. Although [36] and [31] study the impact of momentum on training dynamics, they also focus on the flow regime, which cannot capture phenomena such as the edge of stability or catapults that are inherently unique to the large learning rate regime. Recently, there has been some progress for (S)GD with finite step sizes [13, 35], but they do not consider momentum at all.

Woodworth et al. [47] prove that for the sparse regression problem where the true $\boldsymbol{w}_\star$ is assumed to be sparse, the gradient flow for LDN initialized at $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha \cdot \mathbf{1}$ converges to the minimum ℓ_1-norm solution as $\alpha \to 0$ and minimum ℓ_2-norm solution as $\alpha \to \infty$. It is well-known that the minimum ℓ_1-norm solution has a better generalization capability, due to sparsity of $\boldsymbol{w}_\star$. Subsequently, [35] show that GD with finite learning rate consistently recovers solutions with smaller test loss, even for initializations with large α's.

Role of Momentum in Generalization. Closely related works are Ghosh et al. [16], Jelassi and Li [23], which also consider the positive effects of momentum for nonlinear neural networks. Jelassi and Li [23] prove that a binary classification setting exists where PHB provably generalizes better than GD for a one-hidden-layer convolutional network. Ghosh et al. [16] derive an implicit gradient regularizer [8] for PHB that biases the solution towards flatter minima, and they showed that the momentum regularizer term is stronger than that of GD by a factor of $\frac{1+\beta}{1-\beta}$. In contrast, we report that the behaviors of PHB and GD are fundamentally different.

Appendix B. Deferred Experimental Setting for Section 2.1 (LDN experiments)

The training set is $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ generated as $\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)$ and $\boldsymbol{y}_n = \langle \boldsymbol{w}_\star, \boldsymbol{x}_n \rangle$, where $N = 50, \sigma^2 = 5, d = 100, \boldsymbol{\mu} = 5 \cdot \mathbf{1}$ and $\boldsymbol{w}_\star = (\delta_{i \leq 5}/\sqrt{5})_i$. We use the mean squared error (MSE) loss, and we initialize $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha \cdot \mathbf{1}$, where $\alpha \in [0, 0.3]$ is the initialization scale. Following Nacson et al. [35], for each η_f, we use linear warmup (linearly increasing the learning rate) for the first $\eta_f \cdot 10^6$ steps, starting from $\eta_i = 10^{-8}$. For the momentum parameter of PHB, we use $\beta = 0.9$.

Remark 4 (Learning Rate Warmup) *Learning rate warmup is a standard practice in deep learning [20], with known benefits such as variance reduction [19, 29] and preconditioning [17]. We and Nacson et al. [35] consider learning rate warmup to avoid possible instabilities from using large learning rates.*

Remark 5 (Role of Warmup in Large Catapults) *Although we used linear warmup schedule in the experiments, warmup is not a strict requirement for (large) catapults to occur. Indeed, we observe that catapults consistently occur as long as (1) the iterate is initially close to a global minimum and (2) the learning rate is large enough so that the sharpness is slightly above the MSS but small enough so that the iterates do not completely diverge. Linear warmup provides a natural way to satisfy these two criteria, whereas using large constant learning rates from the beginning usually results in severe instability. We remark that one could use different scheduling to induce the catapults; see Appendix E.1 for ablations on different warm-up schedules.*

Appendix C. Future Works

Firstly for PHB, we sometimes observe a lack of PS after the large catapults, as shown in Figure 2(a–c). However, with larger datasets, PS appears to occur again, as shown in Figure 2(d–f). We conjecture that the lack of PS in ReLU networks is due to catapults aggressively reducing the number of active ReLU neurons [4, 30]; refer to Appendix E.2 for a detailed discussion. Understanding the presence and absence of PS in different setups is an interesting direction. Secondly, although PHB drives iterates towards flatter minima via the large catapults, we do not always observe better generalization. Further exploring the connection between catapults and generalization is an interesting area for future research and may contribute to the growing literature on the connection between flatness and generalization [5, 12, 13, 21].

Appendix D. Missing Proofs and Discussions for Section 3.1

Throughout, recall the following quantities:

$$\tau_u := \inf\left\{t \geq 0 : u_t^2 < \frac{2-\epsilon}{\eta}\right\}, \; C_u := \frac{u_{\tau_u} - \beta u_{\tau_u - 1}}{1-\beta}, \; C_v := \frac{1+\beta}{1-\beta}\sum_{t=\tau_u}^{\infty} v_t^2. \tag{4}$$

The following lemma states that for our dynamics, $\{u_t\}$ is monotone decreasing and convergent:

Lemma 6 $u_\infty := \lim_{t\to\infty} u_t \leq u_{t+1} \leq u_t$ *for all* $t \geq 0$. *Furthermore, if* $\tau_0 := \inf\{t \geq 0 : u_t < 0\} < \infty$*, then we have that* $u_\infty = \frac{u_{\tau_0} - \beta u_{\tau_0 - 1}}{1-\beta}$.

Notice that our experiments correspond to the $\tau_0 = \infty$. So the question remains: how to characterize u_∞ when $\tau_0 = \infty$, which is what Theorem 2 and 3 are doing.

D.1. Proof of Lemma 6

Recall the update rule for u-coordinate:

$$u_{t+1} - u_t = \beta(u_t - u_{t-1}) - \eta(1+\beta)u_t v_t^2 \mathbb{1}[u_t \geq 0]. \tag{5}$$

We proceed by induction. For $t = 1$, it is trivial. For $1 < t < \tau_0$, we have that

$$u_{t+1} - u_t = \beta(u_t - u_{t-1}) - \eta(1+\beta)u_t v_t^2 \leq 0,$$

as $u_t \leq u_{t-1}$ and $\eta u_t v_t^2 > 0$. If $\tau_0 = \infty$, then by monotone convergence theorem, u_t converges to some $u_\infty \geq 0$. If not, then for $t \geq \tau_0$, we can solve the recursion to obtain

$$u_t = u_{\tau_0} + \frac{\beta(u_{\tau_0 - 1} - u_{\tau_0})}{1-\beta}(\beta^{t-\tau_0} - 1).$$

As $u_{\tau_0} < u_{\tau_0 - 1}$ by the definition of τ_0, u_t also monotonically decreases for $t \geq \tau_0$ (in a geometric speed), and we conclude by again applying the monotone convergence theorem. □

D.2. Proof of Theorem 2

By telescoping, we have that for any $t \geq \tau_u + 1$,

$$u_t - u_{\tau_u} = \beta(u_{t-1} - u_{\tau_u - 1}) - \eta(1+\beta)\sum_{k=\tau_u}^{t-1} u_k v_k^2.$$

First, let $N \geq \tau_u$ be fixed and define $C_v(N) := \frac{1+\beta}{1-\beta}\sum_{t=\tau_u}^{N} v_t^2$. Then, for $t \geq N+1$,

$$u_t - \beta u_{t-1} = u_{\tau_u} - \beta u_{\tau_u - 1} - \eta(1+\beta)\sum_{k=\tau_u}^{t-1} u_k v_k^2 \leq (1-\beta)C_u - \eta(1-\beta)C_v(N)u_t.$$

We can rewrite the recursive inequality as

$$u_t - \frac{C_u}{1+\eta C_v(N)} \leq \frac{\beta}{1+\eta(1-\beta)C_v(N)}\left(u_{t-1} - \frac{C_u}{1+\eta C_v(N)}\right).$$

To deal with possibly changing sign, we consider the first time in which the iterates pass another point:

$$\tau_u'(N) := \inf\left\{t \geq N+1 : u_t < \frac{C_u}{1+\eta C_v(N)}\right\}. \tag{6}$$

If $\tau_u'(N) = \infty$, then we have that for all $t \geq N+1$,

$$u_t \leq \frac{C_u}{1+\eta C_v(N)} + \left(\frac{\beta}{1+\eta(1-\beta)C_v(N)}\right)^{t-N-1}\left(u_N - \frac{C_u}{1+\eta C_v(N)}\right). \tag{7}$$

If not, then we have that for all $t \geq \tau_u'(N)$,

$$u_t \leq \frac{C_u}{1+\eta C_v(N)} - \left(\frac{\beta}{1+\eta(1-\beta)C_v(N)}\right)^{t-\tau_u'(N)}\left(\frac{C_u}{1+\eta C_v(N)} - u_{\tau_u'(N)}\right), \tag{8}$$

In either case, we obtain the desired conclusion by taking the limit $\min\{N,t\} \to \infty$ with $t \geq N+1$. □

D.3. Proof of Theorem 3

We start by providing a nonasymptotic version of Theorem 3:

Theorem 7 *For sufficiently small $0 < v_0^2 < \epsilon \ll 1$, we have that*

$$\lim_{t\to\infty} u_t \geq \sqrt{\frac{2}{\eta}} - \sqrt{P_{\tau_u} \exp\left(\frac{4\eta^2 u_{\tau_u}^2 P_{\tau_u}}{\epsilon(2-\epsilon)}\right)}, \tag{9}$$

where P_{τ_u} is a quantity satisfying

$$P_{\tau_u} \leq \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2(2+\epsilon)\sqrt{\frac{2+\epsilon}{2-\epsilon}} + 2(2+\epsilon)\sqrt{\frac{2\epsilon}{2-\epsilon}}\right). \tag{10}$$

Proof [Proof of Theorem 7] In contrast to the proof technique used for Theorem 2, we utilize an energy argument that works for GD. Inspired by the empirical observation that the GD iterates roughly form an ellipse centered around the point $\left(\sqrt{\frac{2}{\eta}}, 0\right)$, we consider the following "elliptical energy" function:

$$P_t := \left(u_t - \sqrt{\frac{2}{\eta}}\right)^2 + \frac{1}{2}v_t^2. \tag{11}$$

Note that $P_0 \leq \frac{\epsilon}{\eta} + \frac{1}{2}v_0^2$. We will first prove that the elliptical energy is well-bounded and then use that fact to lower bound u_∞.

Let us first fix $\epsilon, v_0^2 \in (0,1)$. The following key lemma, whose proof is provided at the end of this section, states that the energy is approximately well-bounded, given that u_t is sufficiently lower bounded:

Lemma 8 $P_{t+1} \leq P_t \exp\left(2\eta^2 u_t^2 v_t^2\right)$ *for any $t \geq 0$ satisfying $u_t \geq \frac{1}{\sqrt{\eta}}$.*

As $u_t^2 \geq u_{\tau_u-1}^2 \geq \frac{2-\epsilon}{\eta} > \frac{1}{\eta}$ for any $t \leq \tau_u - 1$ (due to Lemma 6), we have:

$$\begin{aligned}
P_{\tau_u} &\leq \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2\eta^2 \sum_{t=0}^{\tau_u-1} u_t^2 v_t^2\right) && \text{(telescoping with Lemma 8)}\\
&\leq \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2\eta^2 u_0^2 \sum_{t=0}^{\tau_u-1} v_t^2\right) && \text{(Lemma 6)}\\
&= \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2(2+\epsilon)\eta v_{\tau_u-1}^2 + 2(2+\epsilon)\sum_{t=0}^{\tau_u-2}\frac{u_t - u_{t+1}}{u_t}\right)\\
& && \left(u_0 = \sqrt{\tfrac{2+\epsilon}{\eta}},\ u_t - u_{t+1} = \eta v_t^2 u_t\right)\\
&\leq \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2(2+\epsilon)\eta v_{\tau_u-1}^2 + \frac{2\sqrt{\eta}(2+\epsilon)}{\sqrt{2-\epsilon}}\sum_{t=0}^{\tau_u-2}(u_t - u_{t+1})\right)\\
& && \left(u_t^2 \geq \tfrac{2-\epsilon}{\eta} \text{ for } t \leq \tau_u - 1\right)\\
&= \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2(2+\epsilon)\eta v_{\tau_u-1}^2 + \frac{2\sqrt{\eta}(2+\epsilon)}{\sqrt{2-\epsilon}}(u_0 - u_{\tau_u-1})\right)\\
&\leq \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2(2+\epsilon)\eta v_{\tau_u-1}^2 + 2(2+\epsilon)\sqrt{\frac{2\epsilon}{2-\epsilon}}\right). && \left(\sqrt{a+b} - \sqrt{b} \leq \sqrt{a}\right)
\end{aligned}$$

Also, we have that

$$v_{\tau_u-1}^2 = \frac{u_{\tau_u-1} - u_{\tau_u}}{\eta u_{\tau_u-1}} \leq \frac{u_0}{\eta\sqrt{\frac{2-\epsilon}{\eta}}} = \sqrt{\frac{2+\epsilon}{2-\epsilon}}\frac{1}{\eta}. \tag{12}$$

Thus, we have that

$$P_{\tau_u} \leq \left(\frac{\epsilon}{\eta} + \frac{1}{2}v_0^2\right)\exp\left(2(2+\epsilon)\sqrt{\frac{2+\epsilon}{2-\epsilon}} + 2(2+\epsilon)\sqrt{\frac{2\epsilon}{2-\epsilon}}\right). \tag{13}$$

We now claim that

$$u_t \geq \frac{1}{\sqrt{\eta}} \ \text{ and } \ P_t \leq P_{\tau_u}\exp\left(\frac{4\eta^2 u_{\tau_u}^2 P_{\tau_u}}{\epsilon(2-\epsilon)}\right), \quad \forall t \geq \tau_u - 1, \tag{14}$$

which then implies our desired statement.

We proceed by induction. The base case ($t = \tau_u - 1$) is trivial. For $t' \geq \tau_u$, suppose the statement holds for all $t < t'$. Again, using Lemma 8, we have that

$$P_{t'} \leq P_{\tau_u}\exp\left(2\eta^2 \sum_{t=\tau_u}^{t'-1} u_t^2 v_t^2\right) \leq P_{\tau_u}\exp\left(2\eta^2 u_{\tau_u}^2 \sum_{t=\tau_u}^{t'-1} v_t^2\right).$$

Let us now bound $\sum_{t=\tau_u}^{t'-1} v_t^2$. For $t \in [\tau_u, t'-1]$, we have that $(\eta u_t^2 - 1)^2 < (1-\epsilon)^2$ by the induction hypothesis, and thus, $v_{t+1}^2 < (1-\epsilon)^2 v_t^2$. This implies that

$$\sum_{t=\tau_u}^{t'-1} v_t^2 < v_{\tau_u}^2 \sum_{t=0}^{t'-\tau_u-1}(1-\epsilon)^{2t} \leq v_{\tau_u}^2 \sum_{t=0}^{\infty}(1-\epsilon)^{2t} = \frac{v_{\tau_u}^2}{\epsilon(2-\epsilon)} \leq \frac{2P_{\tau_u}}{\epsilon(2-\epsilon)},$$

and thus, we have that $P_{t'} \leq P_{\tau_u} \exp\left(\frac{4\eta^2 u_{\tau_u}^2 P_{\tau_u}}{\epsilon(2-\epsilon)}\right)$. From the definition of our elliptical energy, this then implies that

$$u_{t'} \geq \sqrt{\frac{2}{\eta}} - \sqrt{P_{\tau_u} \exp\left(\frac{4\eta^2 u_{\tau_u}^2 P_{\tau_u}}{\epsilon(2-\epsilon)}\right)}. \tag{15}$$

As $P_{\tau_u} = \mathcal{O}(\epsilon)$ for small ϵ and $v_0^2 = \mathcal{O}(\epsilon)$, with suitable choices we can conclude that $u_{t'} \geq \sqrt{\frac{1}{\eta}}$, and we are done. ■

Proof [Proof of Lemma 8] This is shown via a brute-force computation:

$$\begin{aligned}
P_{t+1} &= \left(u_{t+1} - \sqrt{\frac{2}{\eta}}\right)^2 + \frac{1}{2}v_{t+1}^2 \\
&= \left(u_t - \eta u_t v_t^2 - \sqrt{\frac{2}{\eta}}\right)^2 + \frac{1}{2}\left(v_t - \eta v_t u_t^2\right)^2 && \text{(GD update)} \\
&= P_t - 2\eta u_t v_t^2 \left(u_t - \sqrt{\frac{2}{\eta}}\right) + \eta^2 u_t^2 v_t^4 - \eta v_t^2 u_t^2 + \frac{\eta^2}{2} v_t^2 u_t^4 \\
&= P_t + u_t^2 v_t^2 \left(\eta^2 v_t^2 + \frac{\eta^2}{2} u_t^2 - 3\eta + \frac{2\sqrt{2\eta}}{u_t}\right).
\end{aligned}$$

We then have the following helpful inequality, whose proof is deferred to the end:

Lemma 9 *For $z \geq \frac{1}{\sqrt{\eta}}$, $\frac{\eta^2}{2}z^2 - 3\eta + \frac{2\sqrt{2\eta}}{z} \leq 2\eta^2\left(z - \sqrt{\frac{2}{\eta}}\right)^2$.*

Using this and the given assumption that $u_t \geq \frac{1}{\sqrt{\eta}}$, we then have the desired statement as follows:

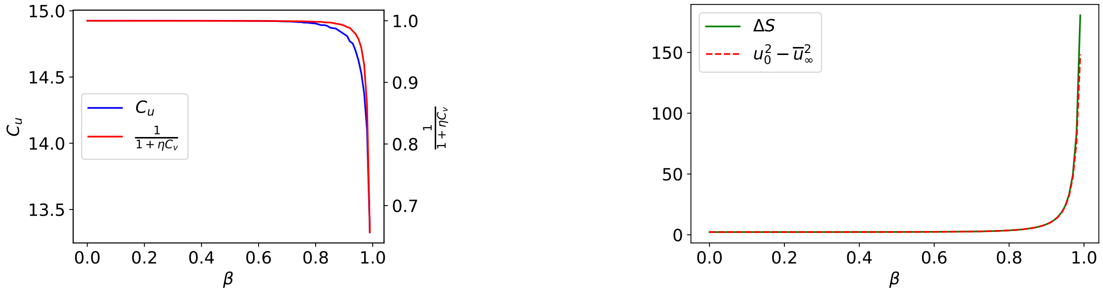
$$P_{t+1} \leq P_t + u_t^2 v_t^2 \left(\eta^2 v_t^2 + 2\eta^2\left(u_t - \sqrt{\frac{2}{\eta}}\right)^2\right) = (1 + 2\eta^2 u_t^2 v_t^2) P_t.$$

■

Proof [Proof of Lemma 9] By reparametrizing $z \leftarrow z/\sqrt{\eta}$, it suffices to prove that for $z \geq 1$, $\frac{1}{2}z^2 - 3 + \frac{2\sqrt{2}}{z} \leq 2\left(z - \sqrt{2}\right)^2$. By rearranging, this is equivalent to

$$f(z) \triangleq 3z^3 - 8\sqrt{2}z^2 + 14z - 4\sqrt{2} \geq 0, \quad \forall z \geq 1.$$

This is then obvious, as f is a cubic function with $f(0) > 0$, the local minimum of $(\sqrt{2}, 0)$ and the local maximum of $\left(\frac{7\sqrt{2}}{9}, \frac{8\sqrt{2}}{243}\right)$. ■



(a) Change of of C_u and $\frac{1}{1+\eta C_v}$

(b) Comparing sharpness reduction

Figure 4: Numerical verification of Theorem 2.

D.4. Numerical Verification of Theorem 2

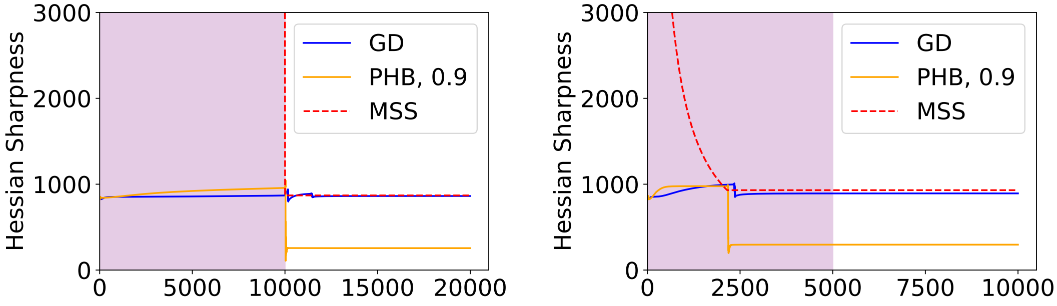
We show numerically that Theorem 2 is tight by rerunning the scalar ReLU network experiment with PHB for varying $\beta \in \{0.0, 0.01, \cdots, 0.99\}$ for $T = 10^5$ iterations. In Figure 4(a), we plot C_u and $\frac{1}{1+\eta C_v}$ across β. We expect that both terms will decrease with increasing β, which turns out to be true; this confirms Hypothesis 1 for the scalar ReLU network case. Connecting this observation to our previous experiment in Figure 3, PHB → GD corresponds to smaller C_u due to the effects of momentum in stages 2-3 while GD → PHB corresponds to larger C_v (hence a smaller $\frac{1}{1+\eta C_v}$) due to momentum prolonging the oscillations in stage 4. Each of these terms explains the increased sharpness reduction in PHB → GD and GD → PHB when compared to GD. The combined effect of these two quantities is shown in Figure 4(b) where we plot the theoretical sharpness reduction $u_0^2 - \overline{u}_\infty^2$ and $\Delta S = u_0^2 - u_T^2$. Note that $u_0^2 - \overline{u}_\infty^2$ is a tight lower bound on ΔS.

D.5. Comparisons to Prior Works

Our scalar ReLU network is essentially identical to the 1D uv-model analyzed in Kalra and Barkeshli [24], Lewkowycz et al. [28]. However, neither work considers the *PHB* dynamics in the parameter space (u_t, v_t). Furthermore, [28] additionally require NTK scaling with a finite but sufficiently large width. Our characterization of the GD dynamics share many characteristics with the single-neuron neural network described and analyzed by Ahn et al. [3], like the quasi-static principle (an ellipse-like envelope) and final resulting sharpness being close to the MSS. However, their analysis cannot be directly applied to our scenario, as $\ell(u) = \frac{1}{2}u^2$ does not satisfy their assumptions (ℓ is globally Lipschitz and $\ell'(u)/u$ decays locally away from $u = 0$). Although we focus on a simple model, our Theorems 2 and 3 provide a rigorous characterization of the parameter dynamics of GD/PHB with large learning rates beyond the previously considered assumptions.

D.6. Tightness of Theorem 2 and 3

In Theorem 2, for GD ($\beta = 0$), we have that $u_\infty \leq \frac{1}{1+\mathcal{O}(1)}\sqrt{\frac{2-\epsilon}{\eta}}$. This follows from the fact that $C_v = \mathcal{O}(1)$ for GD, which we show in the process of proving Theorem 3. This then implies that $\Delta S_\infty \geq \Omega(\epsilon)$, off by a factor of $\sqrt{\epsilon}$ compared to Theorem 3. We leave extending Theorem 3 to PHB for future work. The key difficulty is that the energy argument fails; this can be seen empirically in the PHB trajectory (orange) in Figure 3.



(*a*) Step warmup.

(*b*) Terminating the warmup.

Figure 5: Ablations on the learning rate warmup. (a) Step warmup is used instead of linear warmup. (b) The warmup period is initially set to 5000 iterations, but the warmup is terminated at iteration 2150 around the iteration where the sharpness crosses the MSS.

Appendix E. Additional Experimental Results

Additional Experimental Details Experiments on nonlinear networks were carried out using an A6000. All other experiments were run locally on CPU. Experiments on nonlinear networks are based on code from https://github.com/locuslab/edge-of-stability [9].

E.1. Is Linear Warmup Necessary?

Although we use linear warmup in our experiments, we emphasize that linear warmup is *not* necessary to induce the catapults. As mentioned in the main text, we observe that the main criteria for inducing the catapults are (1) for the iterates to be in a neighborhood of a stable minimum and (2) for the current learning rate to be large enough that the minimum is unstable (in that the sharpness of the minimum is above the MSS) but not so unstable that training diverges. Linear warmup satisfies these two criteria by (1) stably moving the iterates towards a minimum under a low learning rate and (2) automatically finding a suitably large learning rate (that does not lead to divergence) by gradually increasing the learning rate. However, as long as the two criteria are met, catapults can be induced without linear warmup.

Other Warmups in the LDN. To show that the specific form of the warmup is not essential in inducing the catapults, we train an LDN using a step warmup scheduler. We use the learning rate of 10^{-5} for the first 10000 iterations and then 0.0023 for the remaining 10000 iterations. Here, it is necessary to use a sufficiently long warmup period to ensure that the pre-catapult training loss is close to zero. As shown in Figure 5(*a*), this setting also induces a catapult despite not using linear warmup. It should be noted that, unlike linear warmups, the final learning rate must be carefully tuned to prevent training from diverging.

To show the effectiveness of the linear warmup in finding the appropriate scale of the learning rate for inducing catapults, we terminate the warmup as soon as the sharpness crosses the MSS, even before the prescribed warmup period ends. As shown in Figure 5(*b*), this is enough to induce catapults for PHB, supporting our claim that linear warmup has the advantage of "smoothly" finding a suitable learning rate for inducing catapults.

Linear Warmup in the Toy Model. Conversely, although we use a fixed learning rate for the toy model, we show in Figure 6 that catapults still occur even when using linear warmup.

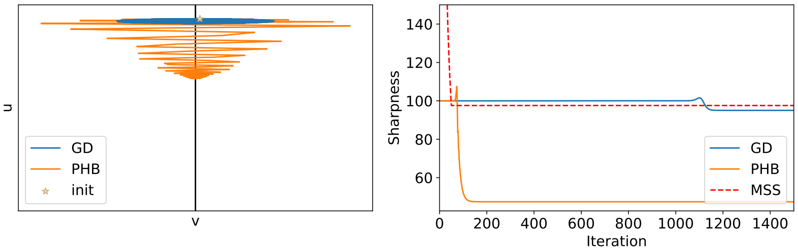
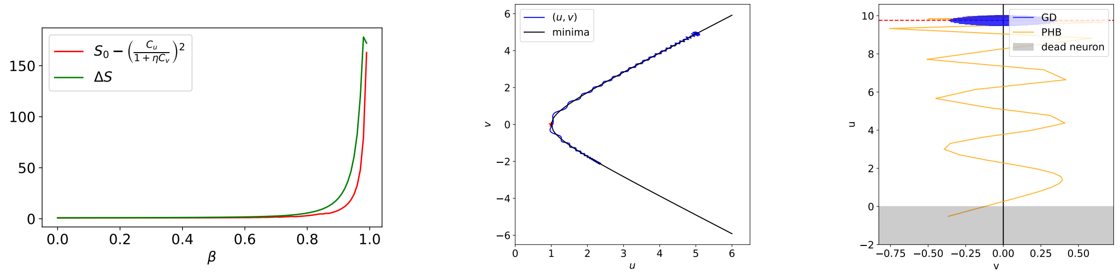


Figure 6: Training the toy model using linear learning rate warmup still induces a catapult

E.2. Effects of "Overshooting" by Larger Catapults

When the momentum parameter β is large, and an even larger catapult can happen and PHB may carry the iterates farther than the flattest minima, resulting in "overshooting." A good example can be found in the simple LDN loss $\mathcal{L}(u, v) = \frac{1}{2}(u^2 - v^2 - 1)^2$. As shown in Figure 7($b$), momentum with $\beta = 0.99$ causes the iterates to move past the flattest minima $(u, v) = (1, 0)$ and converge at a sharper solution. Overshooting may also explain why the final sharpness increases again in Figure 1(b) as α increases further: there is some optimal α for which momentum allows the iterate to reach the flattest possible minima, but increasing α even further may result in overshooting.



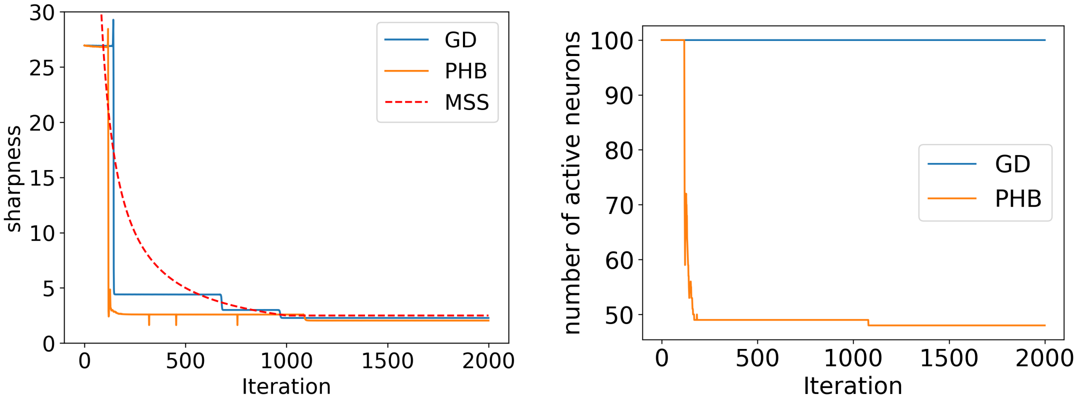
(a) Theoretical lower bound on sharpness reduction based on toy example theory. At $\beta = 0.99$, ΔS slightly decreases due to overshooting.

(b) Trajectory plot for $\beta = 0.99$. The iterates overshoot past $(u, v) = (1, 0)$ which is the flattest minima.

(c) In the **ReLU Scalar Network**, overshooting cause by momentum can lead to the neuron dying.

Figure 7: Overshooting in simple models

As the loss landscape and manifold of minima become increasingly complex, overshooting may result in different outcomes, depending on the architectures. Another effect of overshooting relates to the lack of progressive sharpening in some of the experiments. In some of our experiments (Figures 2(a) and 2(b)), we observed an interesting phenomenon where PHB does not display progressive sharpening after a few large catapults. We conjecture this lack of progressive sharpening to be a result of large catapult inducing more dead neurons. Indeed, through a synthetic experiment with ReLU FCN, we show that large catapults due to PHB induce more dead neurons than GD; see Figure 8. Although seemingly unrelated, overshooting is one possible explanation as to why momentum induces more dead neurons. To illustrate this point, consider the case of the ReLU scalar network in Figure 7(c). Under certain settings, momentum can cause the iterates to overshoot the flattest possible minima $(u = 0)$ and land in the region $u < 0$ which kills the ReLU neuron. A similar phenomenon could occur in more realistic networks as well.



(a) Sharpness.

(b) Number of active neurons.

Figure 8: 2-layer FCN of width 100 trained with MSE loss and rank-2 synthetic dataset [50], generated as follows: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{100}$ with $\boldsymbol{x}_i \sim \mathcal{N}_3(\boldsymbol{0}, \boldsymbol{I})$ and $y_i = (\boldsymbol{x}_i)_1(\boldsymbol{x}_i)_2$.

E.3. More experiments on Nonlinear Neural Networks

E.3.1. ADDITIONAL FCN EXPERIMENTS

For the nonlinear neural network experiment in Figure 9, we follow the setting of Zhu et al. [50]. We train a fully-connected 3-layer ReLU network of width 64 on the synthetic rank-2 dataset. The synthetic rank-2 dataset is generated by i.i.d. sampling data $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and generating outputs $y_i = \boldsymbol{x}_i^{(1)}\boldsymbol{x}_i^{(2)}$ (product of the first two coordinates of $\boldsymbol{x}_i$). A rank2-D-N dataset refers to the synthetic rank-2 dataset generated using $d = D$ whose training set consists of N data points; in our experiment, we used a rank2-400-200 dataset. For both experiments, we used a momentum rate of $\beta = 0.9$ and MSE loss.

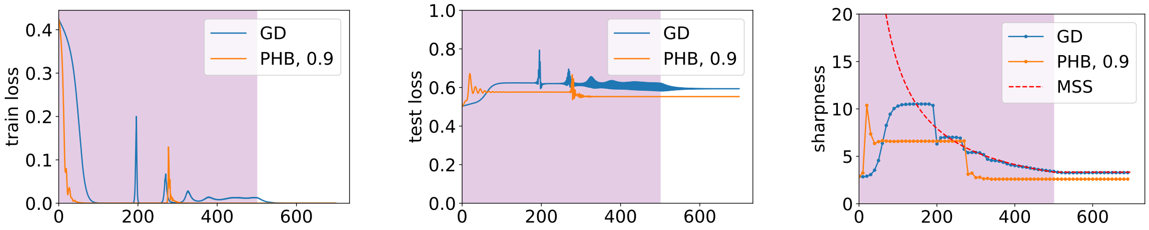
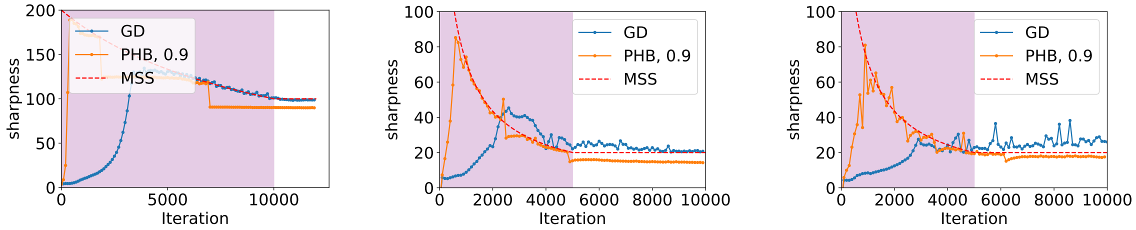


Figure 9: FCN trained on the Rank2-400-200 dataset, $\eta_i = 0.02$, $\eta_f = 0.6$.

E.3.2. RESNET20 EXPERIMENTS

In Figure 10, we provide additional experiments on deep neural networks using the ResNet20 architecture. We observe that large catapults also occur for ResNet20 as well.



(*a*) 2-class 128 subset of CIFAR10, $\eta_i = 0.01$, $\eta_f = 0.02$

(*b*) 1k subset of CIFAR10, $\eta_i = 0.01$, $\eta_f = 0.1$

(*c*) 5k subset of CIFAR10, $\eta_i = 0.01$, $\eta_f = 0.1$

Figure 10: Results for ResNet20. The shaded region is the linear warmup period.

E.3.3. CROSS-ENTROPY LOSS EXPERIMENTS

We provide additional experiments showing that large catapults still occur for networks trained using cross-entropy loss. As shown in Figure 11, for PHB, large catapults occur during early training for FCN trained using Tanh activation and cross-entropy loss. Due to using cross-entropy loss, the iterates also converge quickly to a minimum with near-zero sharpness after the catapult.

Furthermore, for PHB, the iterates quickly converge to a minimum with near-zero sharpness right after the large catapult whereas for GD the iterates do not immediately converge to a flat minimum after the catapult. Instead, the sharpness of GD iterates settles at the MSS after the catapult before slowly converging towards flatter minima as training progresses.

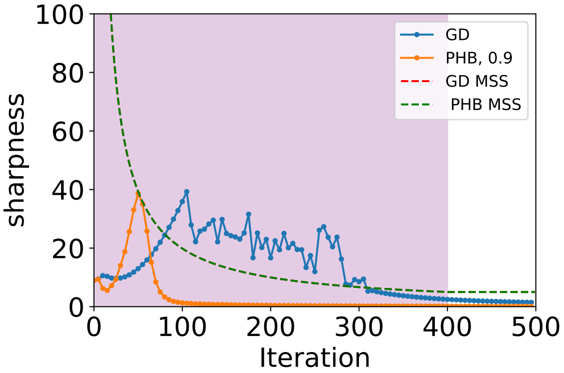


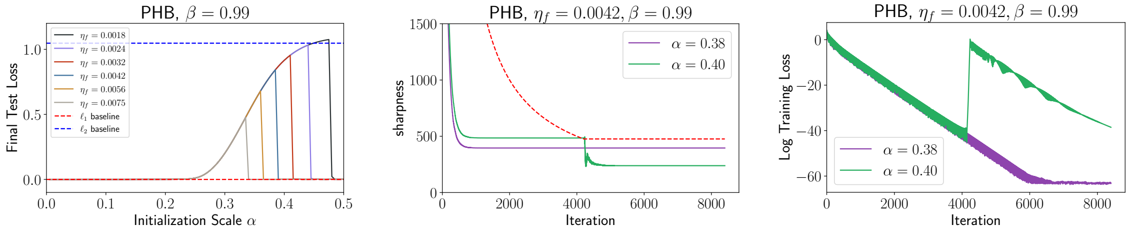
Figure 11: FCN Experiments using *cross-entropy loss* on CIFAR10-1k, with Tanh activation, width 100, $\eta_i = 0.001$, $\eta_f = 0.4$

E.4. Additional Results for $\beta = 0.99$

We redo experiments with $\beta = 0.99$. Overall, the trend is the same, with the effect of momentum more amplified. We provide the necessary details and some discussions for each experiment redone.

E.4.1. LINEAR DIAGONAL NETWORKS

Here, we redo the experiments of Section 2.1 with $\beta = 0.99$, where the results are reported in Figure 12. Note how we expanded the range of α's to see the effect of momentum, which seems to be a bit "delayed". But, at the same time, there is little instability in the trend in that once the curve reaches zero test loss, it stays there; this is in contrast to our $\beta = 0.9$ experiment (Figure 1(*b*) of Section 2.1), where there were some instabilities over α's.



(*a*) PHB Test Loss (*b*) PHB Sharpness (*c*) PHB Log Train Loss

Figure 12: Recall from Section 2.1 that "ℓ_1 baseline" and "ℓ_2 baseline" in Figure 12(*a*) stand for the solution with the minimal ℓ_1 norm and the solution with the minimal ℓ_2 norm to the regression problem, respectively.

E.4.2. SHALLOW NONLINEAR NEURAL NETWORKS

For nonlinear networks, momentum with $\beta = 0.99$ has very unstable training dynamics when trained on a small dataset. Here, we show results for $\beta = 0.99$ on larger datasets. For the CIFAR10 experiments, we train on (1) a subset of CIFAR10 with 2 classes and *2000* training images and (2) a larger subset of CIFAR10 with 10 classes and *5000* training images. Results for the 2-class CIFAR10 are shown in Figure 13, and results for the 5k subset of CIFAR10 are shown in Figure 14. An additional observation is that although PS is exhibited after the large catapults when using $\beta = 0.9$, no PS occurs after the large catapults when using $\beta = 0.99$ For the synthetic rank-2 dataset, we use a Rank2-400-4000 dataset. Results for the synthetic rank-2 dataset are shown in Figure 15.

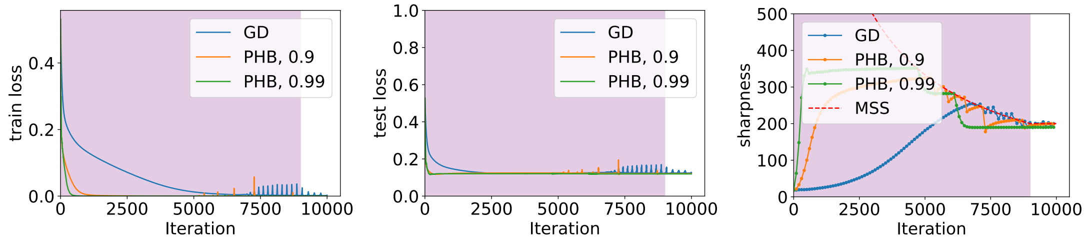
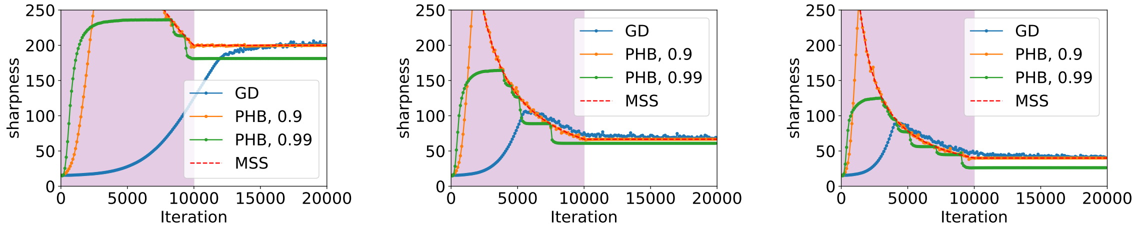


Figure 13: 3-layer width-256 FCN trained on a 2k-datapoint subset of CIFAR10 with MSE loss, $\eta_i = 0.001$, $\eta_f = 0.01$ and 9000 steps of warmup.



(*a*) $\eta_i = 0.001, \eta_f = 0.01$ (*b*) $\eta_i = 0.001, \eta_f = 0.03$ (*c*) $\eta_i = 0.001, \eta_f = 0.05$

Figure 14: 3-layer width-200 FCN trained on a 5k-datapoint subset of CIFAR10 using MSE loss and 10000 steps of warmup

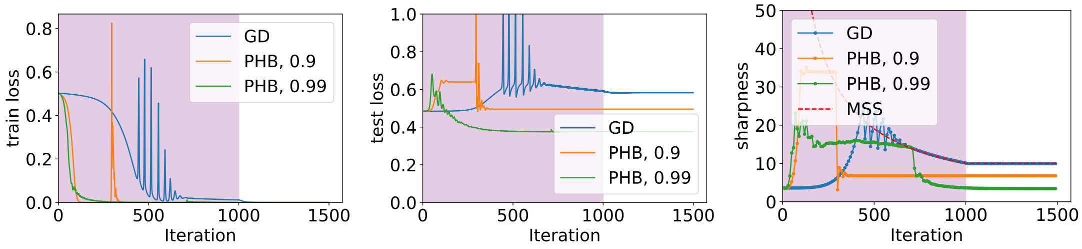
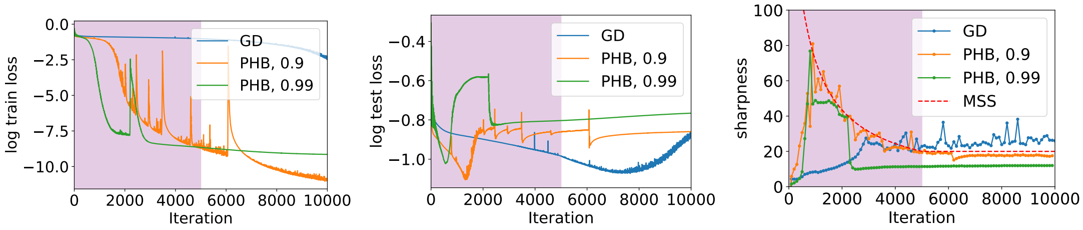


Figure 15: Rank2-400-4000, width=128, MSE loss, $\eta_i = 0.01, \eta_f = 0.2$

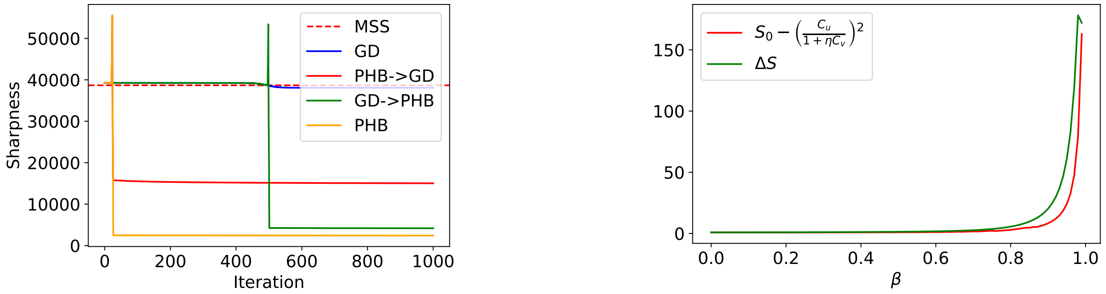
E.4.3. RESNET20

The catapults are more pronounced when training ResNet20 using momentum with $\beta = 0.99$ as shown in Figure 16.



(*a*) ResNet20 trained on a 5k-datapoint subset of CIFAR10 with $\eta_i = 0.01, \eta_f = 0.1$, and 5000 steps of warmup

Figure 16: ResNet20 Experiments with $\beta = 0.99$



(a) Sharpness of LDN for (1) GD, (2) PHB, (3) PHB $\to$ GD, and (4) GD $\to$ PHB.

(b) Numerical verification of Theorem 2 for $\mathcal{L}(u, v) = \frac{1}{2}(u^2 - v^2 - 1)^2$.

Figure 17: Empirical Validation of Hypothesis 1 and theory using LDNs

E.5. Verifying the Hypothesis for LDNs

To show the potential of our verification of hypothesis extending beyond the scalar ReLU network, we revisit the LDNs and perform similar experiments as in the ReLU scalar networks, i.e., we plot the sharpness across 4 scenarios: GD, PHB, GD $\to$ PHB, and PHB $\to$ GD. This is to show empirical evidence that Hypothesis 1 also holds for LDNs. We first initialize the weights close to a minimum by running GD until the (MSE) loss is less than 0.001. We then run each scenario from that same initialization using $\eta = \frac{(2+\epsilon)(1+\beta)}{S_0}$ (and adjusting the learning rates as needed after the sharpness crosses the MSS) where $\epsilon = 0.03$ and S_0 is the sharpness of the initialization. As shown in Figure 17(a), in increasing order of reduction, we again have GD < PHB $\to$ GD < GD $\to$ PHB < PHB.

We now turn to the question of whether Theorem 2 can be extended to LDNs as well. For a general loss function $\mathcal{L}(\boldsymbol{\theta})$, consider running GD and PHB (with rescaled learning rate) from initialization $\boldsymbol{\theta}_0$ satisfying $\lambda_{\max}(\nabla^2 \mathcal{L}(\boldsymbol{\theta}_0)) = \frac{2+\epsilon}{\eta}$ for some small $\epsilon \in (0, 1)$. Then, we extend the definition of C_u and C_v (Eqn. (3)) as the following. For the GD/PHB iterates $\{\boldsymbol{\theta}_t\}_{t \geq 0}$, we solve gradient flow (GF) starting from $\boldsymbol{\theta}_t$ to convergence, and obtain the solution $\boldsymbol{\theta}_t^*$, which can be viewed as the global minimum "closest" to $\boldsymbol{\theta}_t$. Then, for $\tau_u := \inf \left\{ t \geq 0 : \lambda_{\max}(\boldsymbol{\theta}_t^*) < \frac{2-\epsilon}{\eta} \right\}$, define

$$C_u := \frac{\sqrt{S(\boldsymbol{\theta}_{\tau_u}^*)} - \beta\sqrt{S(\boldsymbol{\theta}_{\tau_u - 1}^*)}}{1-\beta}, \quad C_v := \frac{1+\beta}{1-\beta} \sum_{t=\tau_u}^{\infty} \left\langle \boldsymbol{w}_{\max}(\boldsymbol{\theta}_t^*), \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \right\rangle^2, \tag{16}$$

where $\boldsymbol{w}_{\max}(\boldsymbol{\theta}^*)$ is the leading eigenvector of $\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)$. Once we calculate Eqn. (16), a "lower bound" on the sharpness displacement can be derived from Theorem 2, although there is no theoretical guarantee that this should indeed hold for the new $\mathcal{L}$. Still, we can numerically calculate it and assess the possibility of extending Theorem 2 to general functions.

However, running GF to convergence at every iteration is the biggest bottleneck of numerical verification. As a proof of concept, we consider a simple function motivated by the LDN architecture: $\mathcal{L}(u, v) = \frac{1}{2}(u^2 - v^2 - 1)^2$. This corresponds to the MSE loss of LDN for a single data point $(\boldsymbol{x}, y) = (\boldsymbol{e}_1, 1)$, and has the nice property that the GF trajectory $(u(t), v(t))$ satisfies $u(t)v(t) = u(0)v(0)$ for all $t \geq 0$ (see Appendix E.6 for the proof). Using this property, we can calculate $\boldsymbol{\theta}_t^*$ for any $\boldsymbol{\theta}_t$, hence calculating the desired lower bound. Starting from a $\boldsymbol{\theta}_0 = (u_0, v_0)$ with sharpness $\frac{2+\epsilon}{\eta}$, we ran GD/PHB for a range of β's, calculated the actual sharpness displacement and its lower bound.

Appendix E.6 contains further details on the experiment. The results are shown in Figure 17(*b*), which shows a similar trend as in the scalar ReLU network. There is a small decrease in sharpness reduction at $\beta = 0.99$, which can be attributed to overshooting; refer to Appendix E.2 for a discussion.

E.6. Experimental Details for Appendix E.5

In Appendix E.5, we discussed extending Theorem 2 to more general loss functions, and presented a numerically computed "lower bound" on the sharpness decrease for a simple LDN loss $\mathcal{L}(u, v) = \frac{1}{2}(u^2 - v^2 - 1)^2$. This appendix provides more details of this process.

For $\mathcal{L}(u, v) = \frac{1}{2}(u^2 - v^2 - 1)^2$, its gradient and Hessian are given as

$$\nabla\mathcal{L}(u, v) = \begin{bmatrix} 2(u^2 - v^2 - 1)u \\ -2(u^2 - v^2 - 1)v \end{bmatrix}, \quad \nabla^2\mathcal{L}(u, v) = \begin{bmatrix} 6u^2 - 2v^2 - 2 & -4uv \\ -4uv & 6v^2 - 2u^2 + 2 \end{bmatrix}. \tag{17}$$

Now consider running gradient flow (GF), whose dynamics is given as

$$\begin{bmatrix} \dot{u}(t) \\ \dot{v}(t) \end{bmatrix} = -\nabla\mathcal{L}(u(t), v(t)),$$

starting from $(u(0), v(0))$. From the gradient values, it can be checked that $u(t)v(t)$ stays constant throughout the GF trajectory:

$$\frac{d}{dt}(u(t)v(t)) = u(t)\dot{v}(t) + \dot{u}(t)v(t) = 0.$$

Hence, the GF trajectory satisfies $u(t)v(t) = u(0)v(0)$ for all t, from which we can exactly calculate the solution of GF.

For simplicity, we focus on the case $u(0) > 0$ and $\mathcal{L}(u(0), v(0)) < 1/2$. In this case, since $\mathcal{L}(u(t), v(t))$ is always non-increasing along the trajectory, the trajectory has to stay in the region $u(t) > 0$ forever. In such a case, the limit of GF is given by solving

$$u(\infty)^2 - \frac{u(0)^2 v(0)^2}{u(\infty)^2} = 1,$$

which amounts to the solution

$$u(\infty) = \sqrt{\frac{1 + \sqrt{1 + 4u(0)^2 v(0)^2}}{2}}, \quad v(\infty) = \frac{u(0)v(0)}{u(\infty)}. \tag{18}$$

Now, consider a global minimum $\boldsymbol{\theta}_* = (u_*, v_*)$ of $\mathcal{L}(u, v)$. The minimum necessarily satisfies $u_*^2 = v_*^2 + 1$. Substituting this to the loss Hessian (17) gives

$$\begin{aligned} \nabla^2\mathcal{L}(u_*, v_*) &= \begin{bmatrix} 4v_*^2 + 4 & -4v_*\sqrt{v_*^2 + 1} \\ -4v_*\sqrt{v_*^2 + 1} & 4v_*^2 \end{bmatrix} \\ &= (8v_*^2 + 4) \begin{bmatrix} \frac{\sqrt{v_*^2+1}}{\sqrt{2v_*^2+1}} \\ -\frac{v}{\sqrt{2v_*^2+1}} \end{bmatrix} \begin{bmatrix} \frac{\sqrt{v_*^2+1}}{\sqrt{2v_*^2+1}} & -\frac{v}{\sqrt{2v_*^2+1}} \end{bmatrix} \end{aligned}$$

from which we can see that

$$S(\boldsymbol{\theta}_*) = 8v_*^2 + 4, \quad \boldsymbol{w}_{\max}(\boldsymbol{\theta}_*) = \begin{bmatrix} \frac{\sqrt{v_*^2+1}}{\sqrt{2v_*^2+1}} \\ -\frac{v}{\sqrt{2v_*^2+1}} \end{bmatrix},$$

which are the key quantities need for the calculation of C_u and C_v (16).

Based on this background, we draw Figure 17 using the following procedure. We start GD/PHB from (u_0, v_0), whose GF solution (u_0^*, v_0^*) has sharpness $\frac{2+\epsilon}{\eta}$. Specifically, we choose $\eta = 0.01$ and $\epsilon = 0.004$, and initialize at $(u_0, v_0) \approx (5.060, 4.950)$. Every time we update the iterates to get $\boldsymbol{\theta}_t = (u_t, v_t)$, we calculate the corresponding GF solution $\boldsymbol{\theta}_t^* = (u_t^*, v_t^*)$ using (18). From there, we calculate the sharpness, and see if $S(\boldsymbol{\theta}_t^*) < \frac{2-\epsilon}{\eta}$; let τ_u be the first time step t that $S(\boldsymbol{\theta}_t^*) < \frac{2-\epsilon}{\eta}$ happens; from this, we can calculate C_u and C_v as defined in (16) until convergence. In our experiments, we observed that $S(\boldsymbol{\theta}_t^*)$ was monotonically non-increasing, so there was no subtlety involved in calculating C_u and C_v. The monotone decrease of sharpness of GF solution is consistent with Kreisler et al. [26].