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

Prin Phunyaphibarn[*†]
Department of Mathematical Sciences, KAIST
Daejeon, Republic of Korea
prin10517@kaist.ac.kr

Junghyun Lee[*]
Kim Jaechul Graduate School of AI, KAIST
Seoul, Republic of Korea
jh_lee00@kaist.ac.kr

Bohan Wang
University of Science and Technology of China
Hefei, China
bhwangfy@gmail.com

Huishuai Zhang
Microsoft Research Asia
Beijing, China
huzhang@microsoft.com

Chulhee Yun
Kim Jaechul Graduate School of AI, KAIST
Seoul, Republic of Korea
chulhee.yun@kaist.ac.kr

Abstract

Although gradient descent with momentum is widely used in modern deep learning, a concrete understanding of its effects on the training trajectory still remains elusive. In this work, we empirically show that momentum gradient descent with a large learning rate and learning rate warmup displays large catapults, driving the iterates towards flatter minima than those found by gradient descent. We then provide empirical evidence and theoretical intuition that the large catapult is caused by momentum "amplifying" the self-stabilization effect (Damian et al., 2023).

1 Introduction

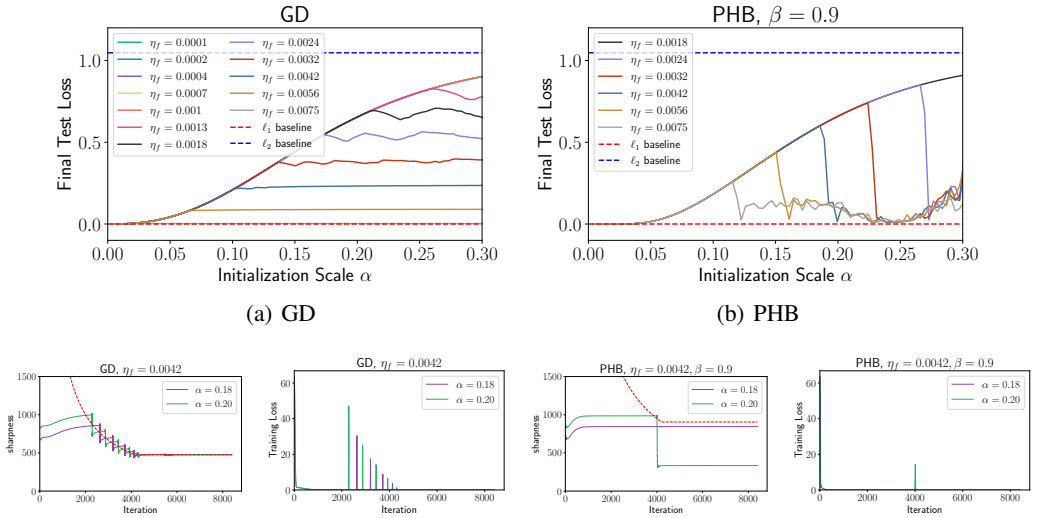
Momentum and learning rate warmups are the most widely used and crucial components for training modern deep neural networks (Goyal et al., 2017; Sutskever et al., 2013). Yet, our understanding of why they work separately, let alone together, is still lacking. Throughout the paper, for a loss function $f(\boldsymbol{w})$, we consider Polyak's heavy-ball momentum (PHB) method given as follows, and from here on, we refer to Polyak's momentum simply as momentum:

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

For momentum, we are not even sure yet why and when it accelerates (stochastic) gradient descent (Fu et al., 2023b; Goh, 2017), and perhaps more importantly, how it changes the implicit bias and thus the resulting model's generalization capability (Ghosh et al., 2023; Jelassi and Li, 2022; Wang et al., 2022a, 2023). For learning rate warmup, some works show that the warmup effectively performs variance reduction (Gotmare et al., 2019; Liu et al., 2020), and more recently, in conjunction with the edge-of-stability phenomena (Cohen et al., 2021, 2022), Gilmer et al. (2022) show that learning rate warmup acts as preconditioning that helps avoid instabilities that arise from using large learning rate. We refer the readers to Appendix A for a more thorough review of related literature.

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

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.



(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. (2022). 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.

In this work, we provide experimental evidence suggesting that momentum combined with learning rate warmup goes through a very steep *sharpness drop*, much more than GD with the same warmup. Thus, this provides an implicit bias effect of driving the iterates towards much flatter minima (Hochreiter and Schmidhuber, 1997), compared to GD. We then discuss why and how such phenomena occur through carefully crafted experimental verifications as well as some theoretical intuitions, inspired by the *self-stabilization* of GD (Damian et al., 2023) and *catapult mechanism* of (S)GD (Lewkowycz et al., 2020; Meltzer and Liu, 2023; Zhu et al., 2023a).

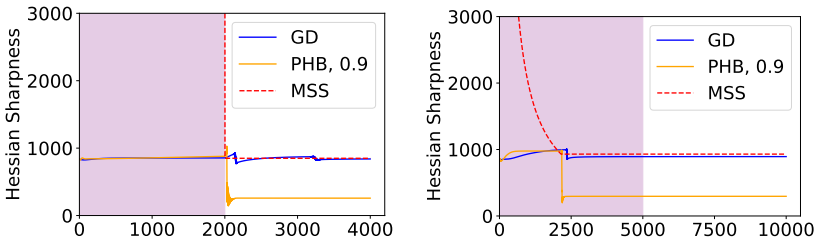
2 Motivating Example: Linear Diagonal Networks

The linear diagonal network (LDN) is known to be one of the simplest non-linear models that display rich, non-trivial algorithmic bias (Woodworth et al., 2020) yet is still tractable. Here, we focus on the depth-2 diagonal linear network given as follows:

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

where $\boldsymbol{x}$ is the input vector, $(\boldsymbol{u}, \boldsymbol{v})$ is the trainable parameter vector, $\boldsymbol{w} := \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}$ is the regression parameter vector. This has been rigorously investigated for gradient flow (Pesme and Flammarion, 2023; Woodworth et al., 2020), stochastic gradient flow (Pesme et al., 2021), and recently, (S)GD with finite step sizes (Even et al., 2023; Nacson et al., 2022). Woodworth et al. (2020) proved that for the sparse regression problem where the true $\boldsymbol{w}$ is assumed to be sparse, the gradient flow for LDN initialized at $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha \cdot \boldsymbol{1}$ converges to the minimum ℓ_1-norm solution when $\alpha \to 0$ and minimum ℓ_2-norm solution when $\alpha \to \infty$. Given that the underlying problem has an intrinsic sparse structure, the minimum ℓ_1-norm solution has a better generalization capability. Later, Nacson et al. (2022) showed that in the finite learning rate regime, gradient descent with a larger learning rate consistently recovers solutions with smaller test loss, even for initializations with large α's, as shown in Figure 1(a).

In this work, we intend to see the implicit bias effect of momentum, and as a motivating example, we first reconduct the sparse regression experiment of Nacson et al. (2022) with momentum. The training set is $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ generated as $\boldsymbol{x}_i \sim \mathcal{N}_d(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{y}_i = \langle \boldsymbol{w}^\star, \boldsymbol{x}_n \rangle$, where $N = 50, \sigma^2 = 5, d = 100, \boldsymbol{\mu} = 5 \cdot \boldsymbol{1}$ and $\boldsymbol{w}^\star = (\delta_{i \le 5}/\sqrt{5})_i$. We initialize $\boldsymbol{u}_0 = \boldsymbol{v}_0 = \alpha \cdot \boldsymbol{1}$, where $\alpha \in [0, 0.3]$ is the initialization scale. For each chosen η_f, we use linear warmup for the first $\eta_f \cdot 10^6$ steps, starting from $\eta_i = 10^{-8}$. For the momentum parameter of PHB, we use $\beta = 0.9$.



(a) Step warmup. (b) Terminating the warmup.

Figure 2: 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.

Over the varying initialization scales α and (final) learning rates η_f, we report the final resulting test losses in Figure 1. Note how strikingly different the plot is between (a) GD and (b) PHB. Compared to GD, which displays a rather monotonic relationship between α and test loss (larger α leads to larger test loss until saturation, as reported in Nacson et al. (2022)), it is clear that PHB has a fundamentally different behavior from GD in the finite learning rate regime. For PHB, when α becomes larger than some threshold $\bar{\alpha}(\eta_f)$ (dependent on η_f), the test loss sharply drops close to zero. Note that $\bar{\alpha}(\eta_f)$ appears to be negatively correlated with η_f, that is, $\bar{\alpha}(\eta_f)$ decreases as η_f increases.

Momentum Induces Larger Catapults. Cohen et al. (2021) show that PHB exhibits unstable behaviors when the sharpness, defined as $\lambda := \lambda_{\max}(\nabla^2 f)$ exceeds *maximum stable sharpness* (MSS) $\frac{2(1+\beta)}{\eta}$. Indeed, it is known that when the learning rate is sufficiently large such that the iterates' sharpness goes *above* the MSS, then several interesting phenomena such as edge of stability (EoS) (Cohen et al., 2021) and catapult (Lewkowycz et al., 2020) arise. Throughout this paper, we refer to *catapult* as a sharp increase in loss, followed by a decrease that forms a single spike in the training loss, coupled with a rapid sharpness reduction.

To find the cause for our observed phenomena, we plot sharpness for PHB in Figure 1(e) at 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 "hits" the MSS curve, it goes through a rapid sharpness reduction, coupled with a loss spike (Figure 1(f)). The sharpness reduction of PHB is so drastic that the final sharpness is 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 iterates stay just below the MSS corresponding to the final learning rate at the end; this was also observed in Zhu et al. (2023a). Here, one could make an educated guess that momentum induces much larger catapults that bias the solution towards flatter minima.

One interesting observation here is that the MSS for PHB is higher than that of GD with the same learning rate. This implies that the α that causes a catapult for GD may not for PHB, i.e., momentum does not improve test loss across all settings (e.g., initialization) as shown in Figure 1(b); but, if a catapult does occur, its reduction of sharpness is much more drastic for PHB than GD.

Role of Linear Learning Rate Warmup. Although we use linear learning rate warmup (henceforth referred to as "linear warmup") in our experiments, we emphasize that linear warmup is *not* necessary to induce the catapults. We claim 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 learning rate to increase until 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.

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 2000 iterations and then 0.002355 for the remaining 2000 iterations. Note that under the same learning rate, the MSS

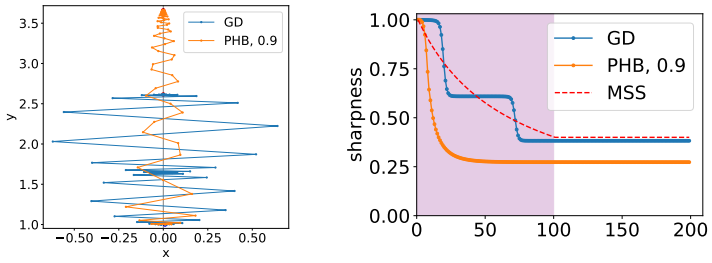


Figure 3: Experiments using $f(x, y) = \frac{x^2}{2y}$ and PHB. Initialization is $(0.01, 1)$, $\eta_i = 2$, $\eta_f = 5$, and $\beta = 0.9$ for PHB. The purple region is the duration of the learning rate warmup.

of GD and PHB differ by a factor of $(1 + \beta)$. To keep the comparison fair, the learning rate used for the PHB experiment is the specified learning rate multiplied by $(1 + \beta)$ so that the MSS match. **Unless stated otherwise, we do this for every plot that compares GD and PHB.** As shown in Figure 2(a), this setting also induces a catapult despite not using linear warmup. However, it should be noted that, unlike linear warmup, 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 2(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.

3 Why Large Catapult? Because Momentum Amplifies Self-Stabilization

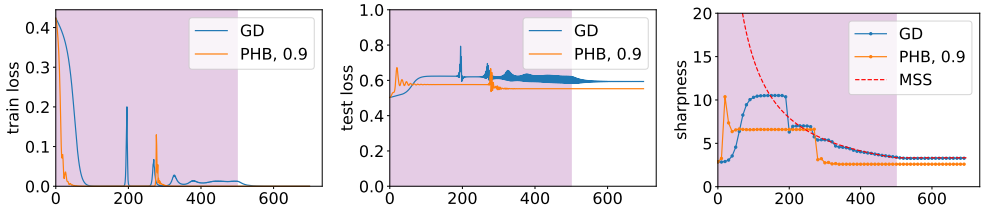
Toy Example. To investigate the mechanism behind this phenomenon, we consider an elementary toy loss function, $f(x, y) = \frac{x^2}{2y}$, restricted to $y > 0$. Note that the (only) manifold of global minima is $x = 0$, and the sharpness decreases as y increases in the $y > 0$ range. Figure 3 provides the trajectory plot for GD and PHB as well as the evolution of sharpness along with the MSS for a specific learning rate setup that induces catapult(s) for *both* GD and PHB. As with the LDN, GD goes through step-wise decreases in sharpness, whereas PHB goes through a single steep sharpness drop. One noteworthy observation is the diverging dynamics in the x-direction (which roughly coincides with the leading eigendirection close to the minima). Note also that the x-direction is orthogonal to both the manifold of global minima and the direction of sharpness decrease (positive y-direction). As such, the sharpness drops occur due to moving in the positive y-direction *while* diverging in the x-direction. In the later stage, the dynamics in the x-direction undergoes damped oscillation due to reduced sharpness at larger y, stabilizing and forcing the dynamics to converge. Appendix C provides additional experimental results and some theoretical intuitions on the effect of momentum.

Remark 3.1. *Here, we provide some justifications on the choice of our toy loss function. The fact that f has a manifold of global minima is consistent with assumptions used in prior works for analyzing gradient-based methods seeking flatter minima (Arora et al., 2022; Damian et al., 2021). Also, the uniqueness of such manifold makes the overall analysis simple, since if there are multiple manifolds of minima, the trajectory of PHB may display chaotic*[3] *behaviors such as overshooting.*

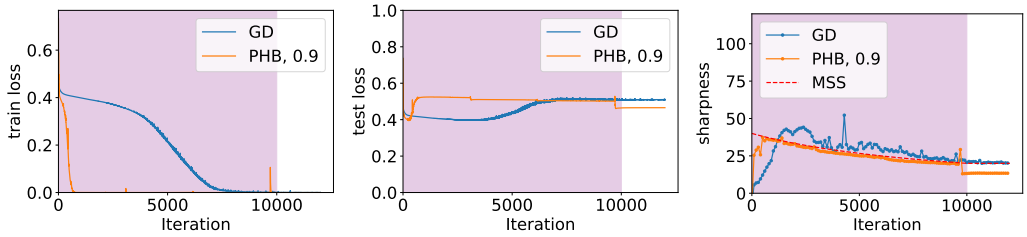
Theoretical Intuitions. The interplay between the movement in the y-direction and the divergence in the x-direction in the above toy example can be explained through *self-stabilization* (Damian et al., 2023). The self-stabilization mechanism consists of four stages: (1) Progressive Sharpening,[4] (2) Blowup, (3) Self-Stabilization, and (4) Return to Stability. In Figure 3, as the MSS decreases below the sharpness, the weights oscillate wildly in the x-direction (stage 2), followed by movement in the y-direction which decreases the gradient in the x-direction (stage 3), allowing the iterates to

[3] Intuitively, each manifold of minima induces a potential field, whose overlap is expected to be quite complex.

[4] Due to the simplicity of our toy model, progressive sharpening (PS) is not observed, which was also the case for simple FCN experiments of Zhu et al. (2023a). Cohen et al. (2021) discuss that PS usually happens for more complex models; indeed, PS is observed in our nonlinear neural network experiments (Section 4).



(a) FCN trained on the Rank2-400-200 dataset, $\eta_i = 0.02$, $\eta_f = 0.6$



(b) ResNet20 trained on a 1k-datapoint subset of CIFAR10, $\eta_i = 0.05$, $\eta_f = 0.1$

Figure 4: Experiments using a 3-layer FCN and a ResNet20, both trained using MSE loss. The purple-shaded region represents the prescribed learning rate warmup period.

stabilize and the sharpness to drop below the MSS (stage 4). Note that stages 2–4 constitute a single catapult. For GD, this process repeats multiple times, resulting in the step-wise decreases in sharpness. However, unlike GD, PHB displays only one steep decrease in sharpness. This phenomenon can be attributed to the role of momentum in stages 3 and 4. For PHB, although the dynamics stabilizes once the sharpness drops below the MSS, momentum prolongs the effect of self-stabilization and continues to propagate the iterates along the y-direction, resulting in a prolonged decrease in sharpness until the momentum term decays to zero. In the process, momentum provides extra stabilization in the x-direction, inducing damped oscillations with respect to the x-axis.[5] In Apppendix B, we follow the self-stabilization analysis of the simple dynamics (Damian et al., 2023) for PHB.

4 Nonlinear Neural Networks

We perform additional experiments on a 3-layer fully connected ReLU network (FCN) with a synthetic dataset and a ResNet20 with a subset of the CIFAR10 dataset; full experimental details are deferred to Appendix D.2. As shown in Figure 4, PHB displays a catapult behavior in nonlinear networks similar to our previous experiments. Notably, the sharpness of GD remains close to the MSS throughout training, going through edge-of-stability (Cohen et al., 2021) and highly unstable training. In contrast, the sharpness of PHB displays a large catapult, resulting in the final sharpness stabilizing well below the MSS, i.e., PHB displays a much more stable training, with the instabilities only appearing during the short catapult phases. Additional ablations and experiments can be found in Appendix D.

Remark 4.1. *We are not claiming that flatter minima necessarily lead to better generalization, and in fact, some "counterexamples" have been recently proposed (Andriushchenko et al., 2023; Dinh et al., 2017; Even et al., 2023). Instead, we solely focus on the effect of momentum and learning rate warmup on how sharpness changes throughout training.*

5 Conclusions and Future Works

In this paper, we empirically observe that PHB with large learning rate and linear learning rate warmup induce large catapults, resulting in a much larger sharpness reduction than that of GD. We then provide empirical evidence on a toy example and theoretical intuition that the large catapult is due to the momentum amplifying the self-stabilization effect (Damian et al., 2023). One immediate and important future direction is to theoretically analyze these newly reported phenomena.

[5] reminiscent of interpreting momentum as a mass-spring-damper system (Muehlebach and Jordan, 2021).

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their helpful comments in improving the paper. 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

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.

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://arxiv.org/abs/2212.07469.

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.

Vladimir Igorevich Arnold. *Catastrophe Theory*. Springer Berlin, Heidelberg, 2 edition, 1986.

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.

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

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.

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.

Alex Damian, Tengyu Ma, and Jason D Lee. Label Noise SGD Provably Prefers Flat Global Minimizers. In *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461. Curran Associates, Inc., 2021. URL https://openreview.net/forum?id=x2TMPhseWAW.

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.

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.

Saber Elaydi. *An Introduction to Difference Equations*. Undergraduate Texts in Mathematics. Springer-Verlag New York, 3 edition, 2005.

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://arxiv.org/abs/2302.08982.

Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):53–98, 1979. doi: https://doi.org/10.1016/0022-0396(79)90152-9. URL https://www.sciencedirect.com/science/article/pii/0022039679901529.

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*, 2023a.

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*, 2023b. URL https://arxiv.org/abs/2306.09000.

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.

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.

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

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.

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.

John Guckenheimer and Philip Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Applied Mathematical Sciences. Springer New York, NY, 1983.

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.

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.

Nikola B. Kovachki and Andrew M. Stuart. Continuous Time Analysis of Momentum Methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021. URL http://jmlr.org/papers/v22/19-466.html.

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.

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.

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.

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.

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.

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://arxiv.org/abs/2304.00488.

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.

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://arxiv.org/abs/2307.04204.

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.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*. Graduate Studies in Mathematics. American Mathematical Society, 2012.

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., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ab3f6bbe121a8f7a0263a9b393000741-Paper-Conference.pdf.

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.

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., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/40bb79c081828bebdc39d65a82367246-Paper-Conference.pdf.

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.

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. *arXiv preprint arXiv:2306.04815*, 2023a. URL https://arxiv.org/abs/2306.04815.

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*, 2023b. URL https://openreview.net/forum?id=p7EagBsMAEO.

Contents

A Further Related Works

A.1 Undertanding the Role of Momentum in Generalization

Closely related works are Ghosh et al. (2023); Jelassi and Li (2022), which also consider the positive effects of momentum for nonlinear neural networks. Jelassi and Li (2022) prove that a binary classification setting exists where PHB provably generalizes better than GD for a one-hidden-layer convolutional network. Ghosh et al. (2023) derive an implicit gradient regularizer (Barrett and Dherin, 2021) for PHB that biases the solution towards flatter minima, and they showed that the regularizer term is strengthened by $\frac{1}{1-\beta}$ times compared to GD.

A.2 Edge of Stability

The first work to systemically report the phenomena of *progressive sharpening (PS)* and the *edge of stability (EoS)* is Cohen et al. (2021), in which the authors consistently observed PS and EoS across various modern neural networks. PS is the phenomenon of the sharpness increasing til MSS, when we run GD(m) with a learning rate below MSS. This is followed by EoS, which is the phenomenon in which, instead of diverging as predicted by classical convex optimization theory, the loss decreases non-monotonically and displays an oscillatory behavior. At the same time, the sharpness also oscillates around the MSS, indicating that as long as the sharpness does not diverge too far from the MSS, the sharpness will force itself to remain close to the MSS. Much works have been devoted to theoretical analyses of PS and EoS (Ahn et al., 2022, 2023; Arora et al., 2022; Damian et al., 2023; Song and Yun, 2023; Wang et al., 2022b; Zhu et al., 2023b) as well as their systematic empirical analyses (Cohen et al., 2021, 2022; Gilmer et al., 2022). There are also works explaining the effect of optimization tricks by considering their interaction with EoS (Fu et al., 2023a; Gilmer et al., 2022).

A.3 Catapults in (S)GD

Lewkowycz et al. (2020) show that GD with large learning rates can induce catapults which results in momentary spikes in training loss but also results in lower sharpness. The theoretical analysis in Lewkowycz et al. (2020) focuses on the network function $f = d^{-1/2}\boldsymbol{u}^T\boldsymbol{v}$, where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$. Meltzer and Liu (2023) extend this work by theoretically deriving sufficient conditions for catapults in quadratic nets. Zhu et al. (2023a) study the loss spikes that occur during the catapults by decomposing the loss into components corresponding to different eigenspaces of the NTK. They observe that the loss spikes occur in the top eigenspace of the NTK while the loss components corresponding to lower eigenspaces are not affected, resulting in the loss quickly decreasing after the catapults. They also show that catapults improve generalization through feature learning and quantitatively measure this using average gradient outer product (AGOP) alignment and that learning rate warmup can induce multiple catapults. 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 (in the presence of learning rate warmup).

B Self-Stabilization of PHB for Simplified Dynamics

We follow the theoretical intuition of the simplified dynamics of self-stabilization as described in Damian et al. (2023). To parallel their argument, we closely follow their notation and assumption. Let $L(\boldsymbol{\theta})$ be the loss function, $S(\boldsymbol{\theta}) := \lambda_{\max}$ be the sharpness, and $\boldsymbol{e}(\boldsymbol{\theta})$ be the corresponding leading eigenvector. We consider a reference point $\boldsymbol{\theta}_\star$ and define the following quantities at $\boldsymbol{\theta}_\star$:

$$\nabla L := \nabla L(\boldsymbol{\theta}_\star), \nabla^2 L := \nabla^2 L(\boldsymbol{\theta}_\star), \boldsymbol{e} := \boldsymbol{e}(\boldsymbol{\theta}_\star), \nabla S := \nabla S(\boldsymbol{\theta}_\star), \gamma := \|\nabla S\|^2.$$

We also define the progressive sharpening coefficient at $\boldsymbol{\theta}_\star$ as $\alpha := -\nabla L \cdot \nabla S$. For simplicity, we assume that $\gamma > 0$ throughout this section.

B.1 Negative Feedback-type Recurrence for PHB

We start by rederiving the negative feedback-type recurrence relation for GD *with momentum*. Here, we do not consider the learning rate warmup. This is because, as discussed in Section 3, warmup's role

is to enable stable training with a large learning rate η_f and deter fast early convergence. Theoretically analyzing the exact effect of warmup in the early stage is left for future works.

Define $a_t = \boldsymbol{e} \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star)$ to be the movement in the unstable direction and $b_t = \nabla S \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star)$ to be the change in sharpness.

We do not assume anything on α, i.e., there are three cases: the progressive sharpening happens ($\alpha > 0$), the gradient of sharpness is orthogonal to that of the loss ($\alpha = 0$), or progressive flattening ($\alpha < 0$). Out of these cases, we especially focus on $\alpha = 0$, as our toy example did not display any change in sharpness during the intervals before the sharpness hits the MSS. We remark that Damian et al. (2023) consider the case of $\alpha > 0$ for GD.

Stage 1: Progressive Sharpening (if $\alpha > 0$) First, we have the following:

$$\begin{aligned} b_{t+1} - b_t &= \nabla S \cdot (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) \\ &= \nabla S \cdot (-\eta \nabla L(\boldsymbol{\theta}_t) + \beta(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})) \\ &\approx \nabla S \cdot (-\eta \nabla L + \beta(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})) \\ &= \eta\alpha + \beta(b_t - b_{t-1}). \end{aligned}$$

Solving this recurrence relation gives

$$b_t = \frac{\eta\alpha}{1-\beta} t + c_1 \beta^t + c_2$$

for some initialization (b_1, b_2) dependent constants c_1, c_2. Note that given $\alpha > 0$, this corresponds to progressive sharpening, and compared to GD, the rate of sharpening is accelerated by $\frac{1}{1-\beta}$ times.

Stage 2: Blowup Next, we consider the approximation of $S(\boldsymbol{\theta}_t) \approx \frac{2(1+\beta)}{\eta} + b_t$. Then,

$$\begin{aligned} a_{t+1} &= \boldsymbol{e} \cdot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star - \eta \nabla L(\boldsymbol{\theta}_t) + \beta(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})) \\ &= a_t - \eta \boldsymbol{e} \cdot \nabla L(\boldsymbol{\theta}_t) + \beta(a_t - a_{t-1}) \\ &\approx a_t - \eta S(\boldsymbol{\theta}_t) + \beta(a_t - a_{t-1}) \\ &\approx a_t - \eta \left(\frac{2(1+\beta)}{\eta} + b_t \right) a_t + \beta(a_t - a_{t-1}) \\ &= -(1 + \beta + \eta b_t) a_t - \beta a_{t-1}. \end{aligned}$$

Stage 3: Self-Stabilization Via similar reasoning, we can derive a similar extension for this stage as follows:

$$b_{t+1} - b_t \approx \eta \left(\alpha - \frac{\gamma}{2} a_t^2 \right) + \beta(b_t - b_{t-1}).$$

Stage 4: Return to Stability When $b_t < 0$ due to the self-stabilization effect, $|a_t|$ shrinks exponentially (at a faster rate thanks to momentum), and we return to Stage 1 again.

Putting them all together All in all, we are left with the following negative feedback-type recurrence for PHB:

$$a_{t+1} \approx -(1 + \beta + \eta b_t) a_t - \beta a_{t-1} \tag{3}$$

$$b_{t+1} - b_t \approx \eta \left(\alpha - \frac{\gamma}{2} a_t^2 \right) + \beta(b_t - b_{t-1}). \tag{4}$$

Note that when $\beta = 0$, we recover the recurrence as derived in Damian et al. (2023). Now, under the approximation that $a_t \approx -a_{t-1}$, we analyze a coupled[6] system of ODEs that roughly describe the dynamics of $(|a_t|, b_t)$.

[6] in the informal sense

B.2 First-order ODE Approximation

B.2.1 Theoretical Analysis

We consider the following system of first-order ODE approximation:

$$\dot{A}(t) = \frac{1}{1-\beta} A(t)B(t), \tag{5}$$

$$\dot{B}(t) = \frac{1}{1-\beta}\left(\alpha - \frac{\gamma}{2}A(t)^2\right). \tag{6}$$

Note that extra $\frac{1}{1-\beta}$ factor due to momentum; see Kovachki and Stuart (2021) for more detailed discussions on momentum.

Depending on the range of α, the equillibrium point(s) differ:

- $\alpha > 0$: unique equillibrium point $(A_\star, B_\star) = (\delta, 0)$, where $\delta := \sqrt{\frac{2\alpha}{\gamma}}$,
- $\alpha = 0$: a manifold of equilibria $A_\star = 0$,
- $\alpha < 0$: no equillibrium point

In other words, the system displays a bifurcation depending on α (often referred to as control variable), which can be studied via the *catastrophe theory* (Arnold, 1986); this is left for future work. In this work, we mainly stick to "elementary" tools from dynamical system theory; see Guckenheimer and Holmes (1983); Teschl (2012) for an overview/introduction.

We first consider the stability of the equilibrium point(s):

Proposition B.1. *For $\alpha > 0$, $(\delta, 0)$ is a saddle point, thus an unstable equilibrium. For $\alpha = 0$, $(0, B_\star)$ is a globally asymptotically stable equilibrium if and only if $B_\star \le 0$.*

Proof. Consider the Jacobian of the ODE system:

$$\boldsymbol{J}(A, B) = \frac{1}{1-\beta}\begin{bmatrix} B & A \\ -\gamma A & 0 \end{bmatrix}.$$

For simplicity, let us ignore the $\frac{1}{1-\beta}$ factor.

When $\alpha > 0$, $\boldsymbol{J}(\delta, 0)$ has two non-zero eigenvalues, $\delta > 0$ and $-\gamma\delta < 0$, i.e., $(\delta, 0)$ is a hyperbolic equilibrium point that is unstable in the linearized dynamics. By Hartman-Grobman theorem (see e.g., Theorem 9.9 of Teschl (2012)), the linearization provides "qualitatively same behavior"[7], and thus $(\delta, 0)$ is an unstable equilibrium.

When $\alpha = 0$, $\boldsymbol{J}(0, B_\star)$ has two eigenvalues, 0 and $B_\star$. As $(0, B_\star)$ is unstable if $B_\star > 0$, let us consider the case $B_\star \le 0$. Note that as we have a zero eigenvalue, linearization analysis is inapplicable, and thus we need to look at its center manifold. It is quite easy to see that the center manifold of the system at $(0, B_\star)$ is $\mathcal{C} := \{(A, B) : A = 0\}$, and the restricted dynamics of the system on $\mathcal{C}$ is globally asymptotically stable, which implies the same to the original system by the Center Manifold Theorem (see e.g., Theorem 3.2.1 and 3.2.2 of Guckenheimer and Holmes (1983)). □

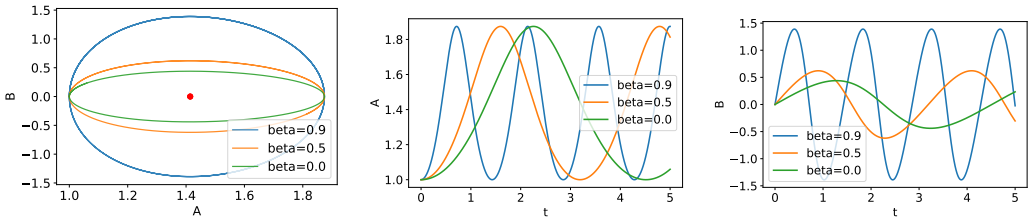
We also have the following conservation law:

Proposition B.2. *Let $h(z) = z - 1 - \log z$. Then,*

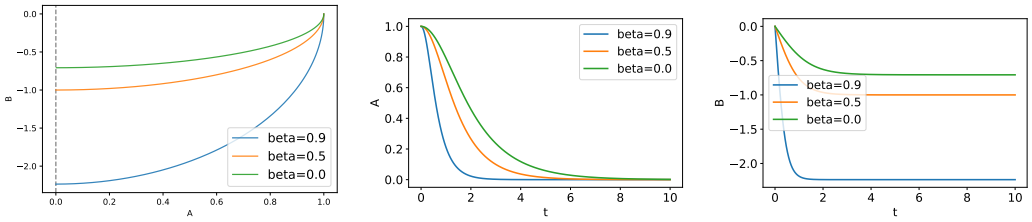
$$E_t(A, B; \alpha, \beta, \gamma) = \begin{cases} h\left(\frac{A^2}{\delta^2}\right) + \frac{(1-\beta)B^2}{\alpha}, & \alpha > 0, \\ A^2 + \frac{2(1-\beta)}{\gamma}B^2, & \alpha = 0. \end{cases} \tag{7}$$

is conserved, i.e., $\partial_t E_t = 0$.

[7] This is formalized using the notion of topological conjugacy.



(a) Simulation results when $\alpha = 1$.



(b) Simulation results when $\alpha = 0$.

Figure 5: (from left to right) A-B phase portrait, $A(t)$ vs. t, and $B(t)$ vs. t. For the phase portrait at $\alpha = 1$, the red dot represents the equilibrium point $(\delta, 0)$, where in our case, $\delta = \sqrt{2\alpha} > 0$. At $\alpha = 0$, the grey dotted line is the x-axis, which is the center manifold.

Proof. Note that $\frac{d}{dt}A^2 = 2A\dot{A} = 2A^2B$ and $\frac{d}{dt}\log A = B$. Then,

$$2B\dot{B} = \frac{2\alpha}{1-\beta}B - \frac{\gamma}{1-\beta}A^2B = \frac{2\alpha}{1-\beta}B - \frac{\gamma}{2(1-\beta)}(\dot{A^2}).$$

Rearranging, we have the following: for $\alpha > 0$,

$$\frac{d}{dt}\left(\frac{(1-\beta)B^2}{\alpha} + \frac{A^2}{\delta^2} - 1 - \log\frac{A^2}{\delta^2}\right) = 0,$$

and for $\alpha = 0$,

$$\frac{d}{dt}\left(A^2 + \frac{2(1-\beta)}{\gamma}B^2\right) = 0.$$

□

Corollary B.1. *For all t, the following holds:*

(i) $\alpha > 0$*:*

$$|A(t)| \lesssim \delta\sqrt{\frac{1-\beta}{\alpha}}|B(0)| + |A(0)|, \quad |B(t)| \lesssim \sqrt{\frac{\alpha}{1-\beta}}\frac{|A(0)|}{\delta} + |B(0)|, \tag{8}$$

(ii) $\alpha = 0$*:*

$$|A(t)| \leq |A(0)| + \sqrt{\frac{2(1-\beta)}{\gamma}}|B(0)|, \quad |B(t)| \leq \sqrt{\frac{\gamma}{2(1-\beta)}}|A(0)| + |B(0)|. \tag{9}$$

Proof. Trivial from triangle inequality and the energy conservation. □

B.2.2 Simulation Results

As done in Damian et al. (2023), we also provide simulation results. Here, we fix $\gamma = 1$ and the initialization to $(A(0), B(0)) = (1, 0)$. Figure 5 shows the result for both $\alpha = 1$ and $\alpha = 0$.

When $\alpha = 1$, the phase portrait orbits around the unstable equilibrium $(\delta, 0)$, just as reported in Damian et al. (2023). Moreover, with larger β, the fluctuation in A stays the same while the fluctuation in B is amplified by approximately $\frac{1}{\sqrt{1-\beta}}$; also, looking at the plots of $A(t)$ and $B(t)$, momentum reduces the period, causing more oscillations. Recalling that B is the change in sharpness, this suggests that momentum does indeed amplify the catapult, resulting in a larger sharpness reduction.

Things are quite different at $\alpha = 0$. Here, as predicted by our theory, note how all trajectories converge to the globally asymptotically stable part of a center manifold $\{(0, B) : B \leq 0\}$, and that with larger β, the final converged point has larger $B_\star$ in magnitude. Such convergent behavior might be connected to fast convergence behavior that we occasionally observe (see e.g., $\alpha = 0.18$ plot of Figure 1 in Section 2).

Overall, $\alpha = 0$ induces a somewhat degenerate dynamics, while $\alpha = 1$ (or generally $\alpha > 0$) induces a limit cycle of Damian et al. (2023), amplified by some factors of β. However, these dynamics do not seem to explain the oscillatory behavior of PHB during/after a catapult.

C Theoretical Intuitions on Late Phase Dynamics of the Toy Example

For vanilla PHB, we have the following discrete dynamical system:

$$x_{t+1} = x_t - \eta \frac{x_t}{y_t} + \beta(x_t - x_{t-1}) = \left(1 - \frac{\eta}{y_t} + \beta\right) x_t - \beta x_{t-1}, \tag{10}$$

$$y_{t+1} = y_t + \eta \frac{x_t^2}{2y_t^2} + \beta(y_t - y_{t-1}), \tag{11}$$

where for simplicity we've set η to be constant. One immediate observation here is that the sequence $\{y_t\}$ is monotone increasing, and the rate of increase roughly gets faster with higher β.

Now suppose that there exists some time step t_0 such that $y_t \gg \sqrt{\eta} x_t$ for all $t \geq t_0$. Then, considering above as a slow-fast-type dynamical system[8], we can consider the following "approximation":

$$x_{t+1} = \left(1 - \frac{\eta}{y_t} + \beta\right) x_t - \beta x_{t-1}, \tag{12}$$

$$y_{t+1} = y_t + \beta(y_t - y_{t-1}). \tag{13}$$

By expanding, we can get a closed form for y_t:

$$y_t = \begin{cases} y_{t_0} + (y_{t_0+1} - y_{t_0}) \frac{\beta^{t_0-1} - \beta^{t-1}}{1-\beta}, & \beta > 0, \\ y_{t_0}, & \beta = 0 \end{cases} \tag{14}$$

For both cases, asymptotically $\{y_t\}$ is close to a constant sequence, say $y_t = y$. Then, we are left with the following second-order homogeneous difference equation:

$$x_{t+1} - \left(1 - \frac{\eta}{y} + \beta\right) x_t + \beta x_{t-1} = 0. \tag{15}$$

When $\beta > 0$, by Theorem 2.37 of Elaydi (2005), x_t is asymptotically stable if and only if $\frac{1}{y} < \frac{2(1+\beta)}{\eta}$, which is true. Also, by Theorem 2.35 of Elaydi (2005), x_t goes through a damped oscillation at the absolute scale of β^t if and only if $\frac{1+\beta}{\eta} < \frac{1}{y}$. This is consistent with our empirical observation.

[8] In the classical slow-fast coupled ODEs, we usually have an infinitesimal time scale ε multiplied to one of the two coupled dynamics, on which the geometric singular perturbation theory (Fenichel, 1979) has been extensively developed. Here, we use the terminology as the behavior is somewhat similar, but we do not claim that our dynamical system is a slow-fast dynamical system in the strict sense.

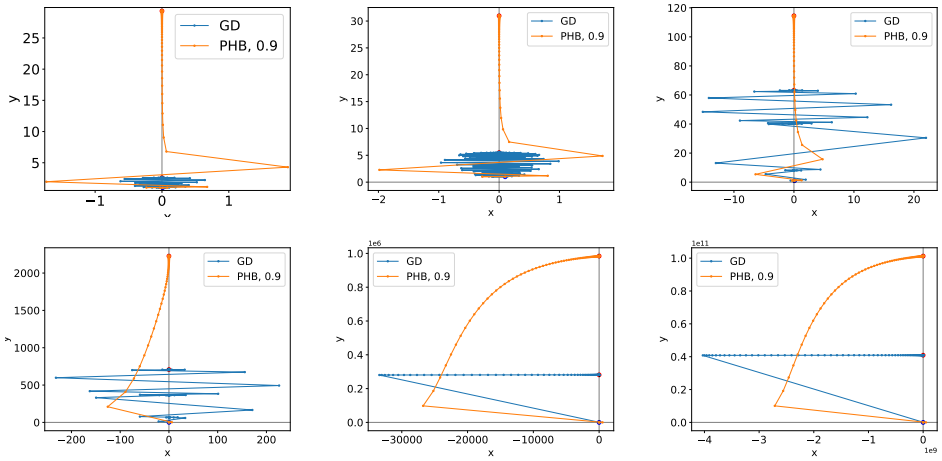


Figure 6: Trajectories of GD and PHB for the toy example with momentum turned off for the x-direction. Initialization is $(0.01, 1)$, $\eta_i = 2$, and $\beta = 0.9$ for PHB.

When $\beta = 0$, we have that

$$x_{t+1} - \left(1 - \frac{\eta}{y}\right) x_t = 0, \tag{16}$$

which amounts to

$$x_t = x_{t_0} \left(1 - \frac{\eta}{y}\right)^{t - t_0}, \tag{17}$$

i.e., x_t converges to 0 geometrically fast, without any oscillation. This is consistent with our empirical observation when considering PHB with momentum turned off for the x-direction.

Our underlying assumption that there exists some time step t_0 such that $y_t \gg \max(\eta, x_t)$ for all $t \geq t_0$ is inspired by our empirical observations. A rigorous theoretical characterization of PHB for even such simple loss functions is expected to be quite hard but, at the same time, fruitful, and we leave it to future work.

C.1 Simulation Studies

To see that our simplified analysis captures the behavior, we consider the scenario in which momentum is only applied to the y-direction and the x-direction goes through vanilla GD updates. The results are shown in Figure 6, for $\eta_f \in \{5, 10^1, 10^2, 10^3, 10^5, 10^{10}\}$. The result shows that turning off the momentum in the x-axis leads to instability in the training trajectory: the fluctuation in the x-axis increases significantly, and no damped oscillation is observed. This matches the presented theoretical intuition and our hypothesis that the momentum amplified the self-stabilization effect, i.e., more stable.

Some observations to look out for:

- With higher η_f, the slow-fast-type dynamics are more apparent; note how, at certain intervals, the x (or y) coordinates of the iterates stay roughly the same!
- Our theoretical intuition partially explains some of the behaviors. For instance, for $\eta_f \in \{10^5, 10^{10}\}$, PHB has a gradual (yet still quite large compared to GD) increase in both x and y in magnitude, then suddenly changes the course to the predicted slow-fast dynamics phase where y stays roughly constant.

Indeed, our rough theoretical arguments do not explain all the observed behaviors. For instance, for PHB with momentum turned off for the x-direction, the trajectory has a single large catapult at the beginning, then goes through a phase in which x stays roughly constant while y increases significantly, then enters the slow-fast dynamics phase. Our current "theory" also does not cover the middle phase, which displays quite an interesting behavior.

D Additional Experimental Results

D.1 Toy Example

Here, we provide additional results for the toy example, $f(x, y) = \frac{x^2}{2y}$, for varying the final learning rate $\eta_f \in \{10^1, 10^2, 10^3, 10^5, 10^{10}\}$. We set the initialization to be $(0.01, 1)$, $\eta_i = 2$, and $\beta = 0.9$ for PHB. We report the results in Figures 7 and 8.

Remark D.1. *The reason that we could consider such large learning rates for our specific toy model is that the gradient at (x, y) is $\left(\frac{x}{y}, -\frac{x^2}{2y^2}\right)$, and thus at higher y, the magnitude of the gradient field gets smaller. This deters the iterates from ever diverging, and we believe that in our example, the iterates always converge regardless of the initialization and learning rate. However, as this is often not the case, we expect that the self-stabilization effect described in Appendix B is critical in stabilizing the dynamics of PHB for realistic scenarios.*

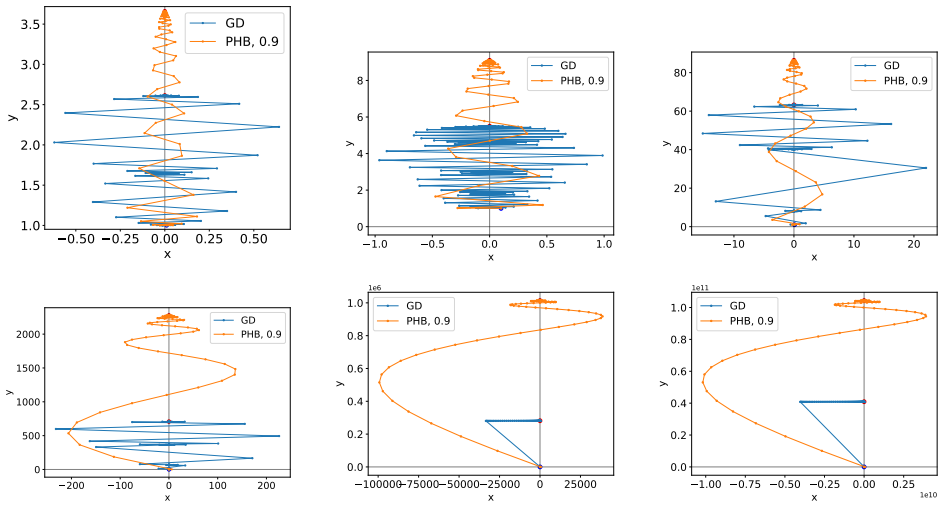


Figure 7: Trajectories of GD and PHB for the toy example.

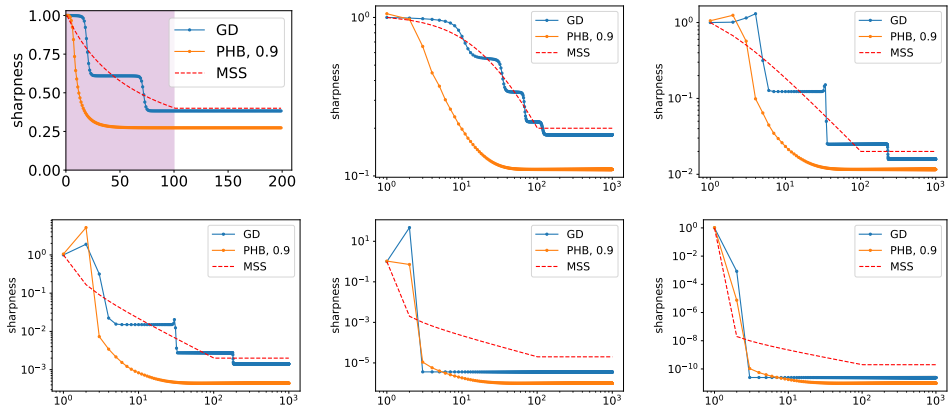


Figure 8: Sharpness plots of GD and PHB for the toy example.

D.2 Nonlinear Neural Networks

D.2.1 Deferred Settings

For the first nonlinear experiment, we follow the setting of Zhu et al. (2023a). 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 the second nonlinear experiment, we use a ResNet20 with a 1k-datapoint, 10-class subset of CIFAR10 (Figure 9 shows the result for a 128-datapoint, 2-class subset of CIFAR10.) For both experiments, we used a momentum rate of $\beta = 0.9$ and MSE loss.

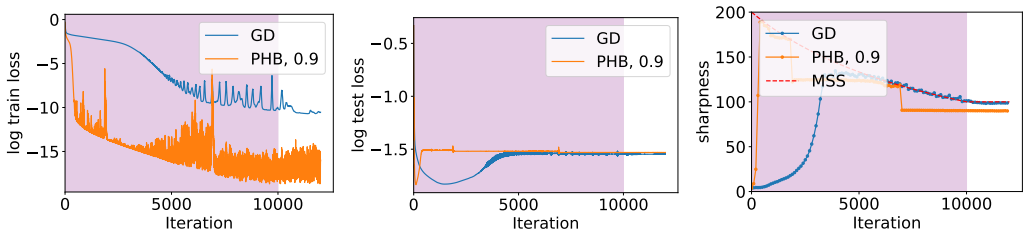
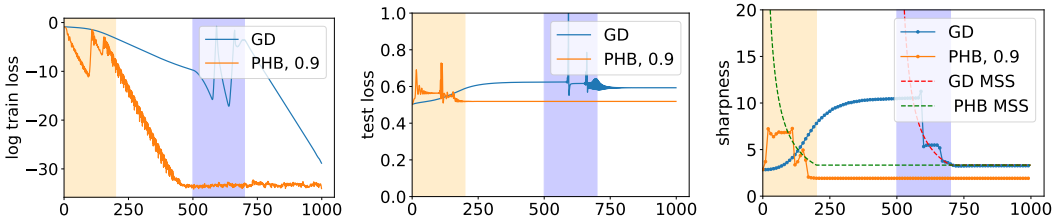


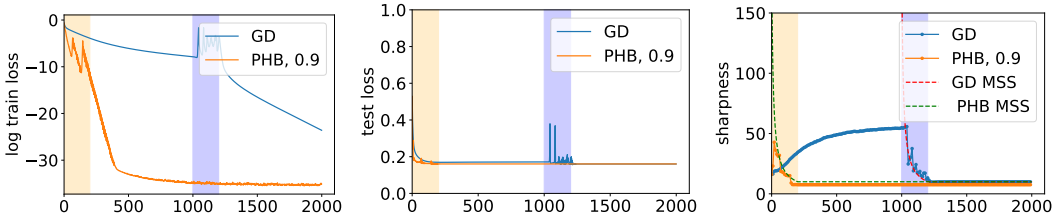
Figure 9: ResNet20 trained on a 128-datapoint, 2-class subset of CIFAR10 with MSE loss.

D.2.2 Ablation for Nonlinear Neural Networks: Pre-catapult Training Loss

We conduct ablation studies to show that the catapult effect is an inherent characteristic of momentum itself and not simply a byproduct of confounding variables such as the pre-catapult training loss. In Fig. 10, we conduct the same experiments as in Section 4 but where the warmup period does not begin from the first iteration and may not necessarily be equal for GD and PHB. Looking at the log training loss plot, we note that the catapult occurs either when GD and PHB achieve similar training loss or when PHB has higher training loss than GD. Thus it seems that the pre-catapult loss level does not matter when inducing catapults.

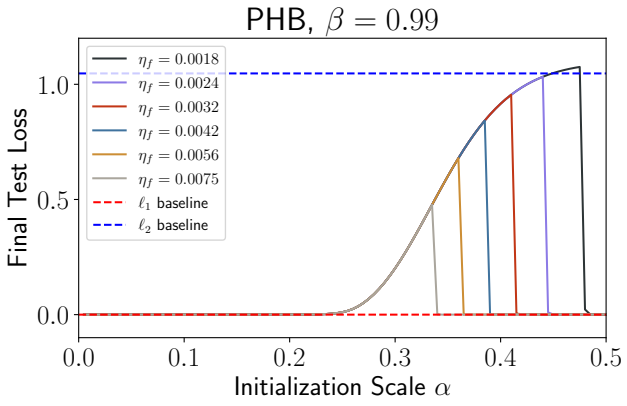


(a) Rank 2, width=64, $\eta_i = 0.02$, $\eta_f = 0.6$

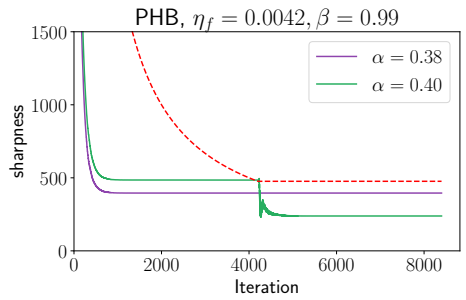


(b) CIFAR10-128, width=256, $\eta_i = 0.01$, $\eta_f = 0.2$

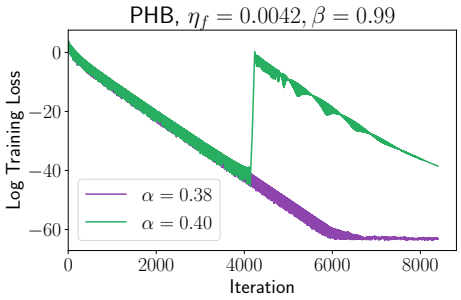
Figure 10: Experiments using a fully connected 3-layer ReLU network with constant width trained using MSE loss. The region represents the learning rate warmup period. We use different warmup periods for GD and PHB to ensure that PHB does not achieve significantly lower training loss than GD when the first catapult occurs. Note that we use a log scale for the training loss plots



(a) PHB



(b) PHB Sharpness



(c) PHB Log Train Loss

Figure 11: Again, note that in (a), "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.

D.3 Additional Results for $\beta = 0.99$

We redo every experiment 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.

D.3.1 Linear Diagonal Networks

Here, we redo the experiments of Section 2 with $\beta = 0.99$, where the results are reported in Figure 11. 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's 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), where there were some instabilities over α's.

D.3.2 Toy Example

Here, we provide additional results for the toy example with $\beta = 0.99$ for vanilla GD and vanilla PHB. The results are shown in Figures 12 and 13.

D.3.3 Nonlinear Neural Networks

For nonlinear networks, momentum with $\beta = 0.99$ has very unstable training dynamics when trained on a small dataset. For the CIFAR10 experiment, we train on a small subset of CIFAR10 with 2 classes and *2000* training images. For the synthetic rank-2 dataset, we use a Rank2-400-4000 dataset. Results are shown in Figure 14

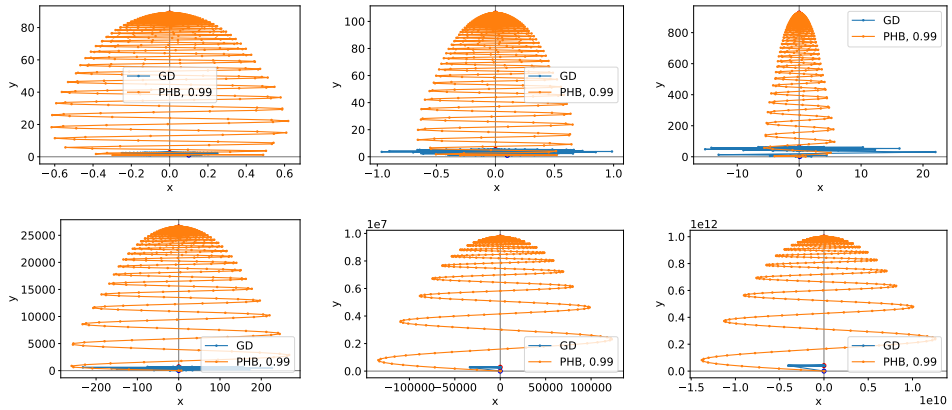


Figure 12: Trajectories of GD and PHB for the toy example.

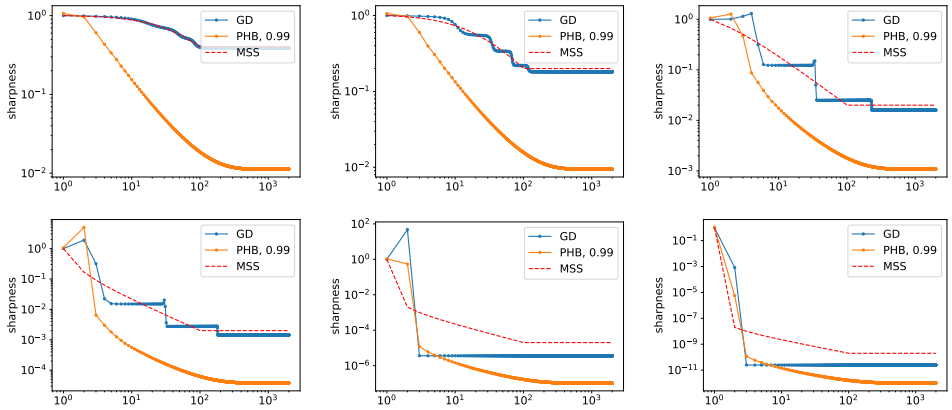
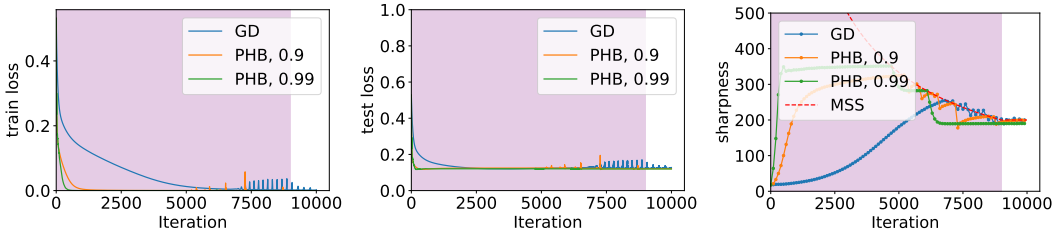
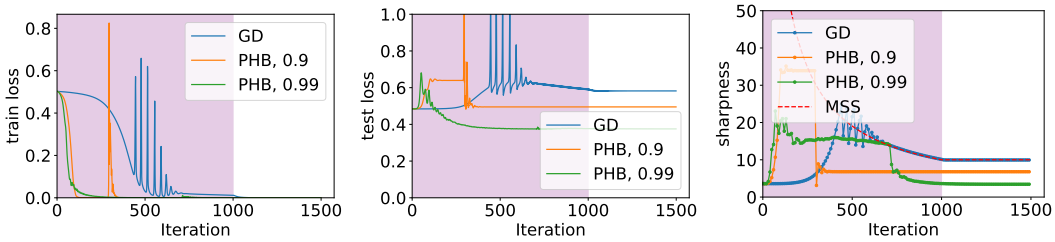


Figure 13: Sharpness plots of GD and PHB for the toy example.



(a) CIFAR10-2k, width=256, MSE loss, $\eta_i = 0.001$, $\eta_f = 0.01$



(b) Rank2-400-4000, width=128, MSE loss, $\eta_i = 0.01$, $\eta_f = 0.2$

Figure 14: Nonlinear Neural Network Experiments with $\beta \in \{0.0, 0.9, 0.99\}$ and larger training set