Understanding Polyak's Momentum in Deep Learning May Require Rethinking Non-Convex Optimization

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Polyak's heavy-ball momentum is widely used in deep learning, where it often accelerates training in practice. However, standard smooth non-convex optimization theory, which typically measures convergence by averaged or best-iterate gradient norms, offers only a limited explanation of this advantage. We revisit this gap through worst-case lower bounds. For SGD with heavy-ball momentum (SHB), SignGD with momentum (Signum), and Muon, we show that the lower bounds on averaged gradient norms considered for these methods can exceed the upper bounds for their non-momentum counterparts, even with their respective optimal constant step sizes. These comparisons cover commonly used ranges of the momentum parameter β, become unfavorable to momentum as β increases, and diverge in deterministic settings as $\beta \to 1$. For GD with heavy-ball momentum (HB), we further show that the same separation persists under the best-iterate squared gradient norm. These results indicate that the standard framework can lead to comparisons opposite to the empirical behavior of momentum in deep learning, motivating refinements involving convergence measures, structure, or stochasticity.

1. Introduction

The development of momentum-based acceleration traces back to the foundational work on Polyak's heavy-ball method [25] and Nesterov's accelerated gradient (NAG) [22]. The empirical success of these momentum-based frameworks is particularly evident in deep learning, where they are considered essential for achieving efficient training and superior generalization [32]. In large-scale applications, the heavy-ball type update has become especially prevalent because it is simple to implement and appears across many optimizers, including Adam [18], Signum [2], and Muon [14].

Beyond the quadratic regimes where heavy-ball momentum is classically known to help, smooth non-convex optimization provides a standard framework for studying first-order methods under weaker assumptions. Under L-smoothness, bounded initial gap, and a bounded-variance stochastic first-order oracle, convergence guarantees are commonly stated in terms of gradient norms such as $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q]$ or $\min_{0\le t<T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q]$. We use this framework to compare heavy-ball momentum methods with their non-momentum counterparts.

Viewed through this lens, however, existing convergence analyses often fail to reflect the practical advantages of heavy-ball momentum. For stochastic gradient descent with heavy-ball momentum (SHB) and SignGD with momentum (Signum), known smooth non-convex upper bounds are larger than those for stochastic gradient descent (SGD), even under optimally tuned step sizes, and their dependence on the momentum parameter β can worsen as $\beta \to 1$ [20, 31]. We ask whether these

pessimistic comparisons merely reflect looseness in existing analyses, or whether they can be certified by worst-case lower bounds under the same assumptions and convergence criteria.

1.1. Summary of Our Contributions

Our contributions center on lower bounds for methods with Polyak-style momentum in the stationarity-based smooth non-convex setting. We compare each momentum method with its non-momentum counterpart under optimized constant step sizes, using averaged gradient norm criteria for SHB, Signum, and Muon, and the best-iterate squared gradient norm for deterministic HB.

- In Section 3, we prove a lower bound for SHB under averaged squared gradient norm that exceeds the optimized SGD upper bound over a β-dependent range of stochasticity levels. We also prove a deterministic best-iterate squared-gradient lower bound for HB, yielding an analogous unfavorable comparison for every $\beta \in [7/8, 1)$.
- In Section 4, we prove a lower bound for Signum under averaged gradient ℓ_1-norm that exceeds the optimized SignGD upper bound when $\beta > 0.886$. We also transfer the Signum construction to Muon on diagonal matrix instances.
- In Section 5, we discuss the gap between our theoretical results and the empirical success of momentum in deep learning. We examine what the standard framework may miss, including alternative convergence measures, extra structural assumptions, and a better account of stochasticity.

Our work complements prior positive and negative analyses of momentum, which often depend on additional assumptions. We focus the main text on the lower bound comparisons and defer a broader discussion of related work to Section A.

2. Preliminaries

For vectors, $\|\cdot\|_p$ denotes the ℓ_p-norm, and for matrices, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_*$ denote the Frobenius and nuclear norms. Let $\mathcal{F}_L(\Delta)$ be the class of L-smooth objectives satisfying $f(\mathbf{0}) - f^\star \le \Delta$, where $f^\star = \inf_{\boldsymbol{x}} f(\boldsymbol{x})$. We assume access to an unbiased stochastic first-order oracle $g(\boldsymbol{x}; \xi)$ with variance at most σ^2. Formal definitions and additional notation are collected in Section B.

We consider three first-order methods with Polyak-style momentum: SGD with Polyak's heavy-ball momentum (SHB), SignSGD with momentum (Signum), and Muon.[1] Those algorithms are parameterized by a step size $\eta > 0$ and a momentum parameter $\beta \in [0, 1)$.

For SHB and Signum, given $\boldsymbol{g}_t = g(\boldsymbol{x}_t; \xi_t)$, the momentum variable is updated as $\boldsymbol{m}_{t+1} = \beta \boldsymbol{m}_t + \boldsymbol{g}_t$, with $\boldsymbol{m}_0 = \mathbf{0}$. The corresponding updates are

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \boldsymbol{m}_{t+1} \quad \text{(SHB)}, \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \operatorname{sign}(\boldsymbol{m}_{t+1}) \quad \text{(Signum)}$$

For Muon, given $\boldsymbol{G}_t = g(\boldsymbol{X}_t; \xi_t)$ and $\boldsymbol{M}_0 = \mathbf{0}$, the updates are

$$\boldsymbol{M}_{t+1} = \beta \boldsymbol{M}_t + \boldsymbol{G}_t, \quad (\boldsymbol{U}_{t+1}, \boldsymbol{S}_{t+1}, \boldsymbol{V}_{t+1}) = \mathrm{SVD}(M_{t+1}), \quad \boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta \boldsymbol{U}_{t+1} \boldsymbol{V}_{t+1}^\top.$$

Note that setting $\beta = 0$ recovers SGD, SignSGD, and SpecGD, respectively. In the deterministic setting ($\sigma = 0$), we refer to SignSGD as SignGD.

1. The original Muon applies Newton-Schulz iterations to approximately orthogonalize the momentum matrix $\boldsymbol{M}_t$. In this work, we instead compute the exact SVD of the momentum.

Given a sequence of iterates $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ generated by a first-order method, we evaluate convergence using averaged and best-iterate gradient norms: $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q]$ and $\min_{0\le t<T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q]$, where the expectation is over the oracle randomness. In particular, we use the ℓ_2-norm with $q=2$ for SHB, the ℓ_1-norm with $q=1$ for Signum, and the nuclear norm with $q=1$ for Muon.

3. Lower Bounds for SHB

In this section, after recalling the SGD upper bound, we prove an averaged squared-gradient lower bound for SHB that can exceed the SGD upper bound under optimized step sizes. We then show that an analogous comparison persists for deterministic HB under the best-iterate squared gradient norm.

3.1. Averaged Squared Gradient Norm Lower Bound for SHB

We first recall the SGD upper bound of Ghadimi and Lan [10]. For any $f\in\mathcal{F}_L(\Delta)$ and any unbiased stochastic first-order oracle with variance at most σ^2, the SGD iterates $\boldsymbol{x}_{t+1}=\boldsymbol{x}_t-\eta g(\boldsymbol{x}_t;\xi_t)$ satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right]\le\frac{\Delta}{T\eta(1-\eta L/2)}+\frac{\eta L\sigma^2}{2-\eta L}. \tag{1}$$

Minimizing the right-hand side over $\eta>0$ gives $\frac{\Delta L}{T}\left(1+\sqrt{1+\frac{2\sigma^2T}{\Delta L}}\right)$.

There exists an L-smooth function and a stochastic first-order oracle for which the bound in (1), when stated with $f(\boldsymbol{x}_0)-\mathbb{E}[f(\boldsymbol{x}_T)]$ in place of Δ, holds with *equality*. See Section C.1 for a constructive example.

Remark. Existing optimized upper bound comparisons for SHB are also unfavorable relative to SGD in the large-T stochastic regime; see Section C.1.

Theorem 1 provides a lower bound for SHB, under the averaged squared gradient norm.

Theorem 1 *Suppose* $T\ge\max\left\{25,\frac{1}{1-\beta}\left(5\log\frac{1}{1-\beta}+9\right),39M+10\right\}$, *where* $M:=\frac{\sigma^2}{2\Delta L}$. *When* $0<M<1$, *we additionally assume that* $\beta\ge\frac{M}{4}$. *For any step size* $\eta>0$, *function* $f\in\mathcal{F}_L(\Delta)$, *and an unbiased stochastic oracle* g *with variance at most* σ^2, *let* $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ *be the iterates generated by SHB. Then,*

$$\inf_{\eta>0}\sup_{f\in\mathcal{F}_L(\Delta)}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2]\ge\frac{\Delta L}{T}\left(1-\frac{5.05}{T}\right)\left(\sqrt{\Psi}+\frac{1+\beta^2}{1-\beta^2}-\frac{M}{2}\right) \tag{2}$$

where $\Psi=1-2M\frac{1+\beta}{1-\beta}+4MT$. *Moreover, the right-hand side of (2) is positive.*

We defer the proof of Theorem 1 to Section E.

Our lower bound instance is simple: up to translation, it is the one-dimensional quadratic $Lx^2/2$ with additive symmetric noise $\pm\sigma$. Its condition number is 1, so the construction lies outside the ill-conditioned quadratic regime underlying Polyak's classical acceleration; rather, it shows that the standard worst-case framework can be driven by geometries where momentum need not help.

Figure 1 compares the SHB lower bound from Theorem 1 with the SGD upper bound from (1) across the stochasticity level $M=\frac{\sigma^2}{2\Delta L}$ and the momentum parameter β. The unfavorable region expands as β increases and shrinks as M increases. The white dashed curve shows the large-T boundary $\frac{1+\beta^2}{1-\beta^2}-\frac{M}{2}=1$, whose derivation is given in Section D.1.

3.2. Best-Iterate Lower Bound for HB

We next show that in the deterministic setting, an analogous separation holds under the best-iterate squared gradient norm. For comparison, Rotaru et al. [27] give the tight GD upper bound under the same metric:

$$\inf_{\eta>0} \sup_{f\in\mathcal{F}_L(\Delta)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 = \frac{6\sqrt{3}\Delta L}{8(T-1)+3\sqrt{3}}. \tag{3}$$

We next establish the corresponding lower bound for HB under the same metric.

Theorem 2 *Fix $\beta \in [7/8, 1)$. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ denote the iterates generated by HB with momentum parameter β and step size $\eta > 0$. Then,*

$$\liminf_{T\to\infty} \left(T \inf_{\eta>0} \sup_{f\in\mathcal{F}_L(\Delta)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2\right) \ge \frac{\Delta L}{4(1-\beta)}.$$

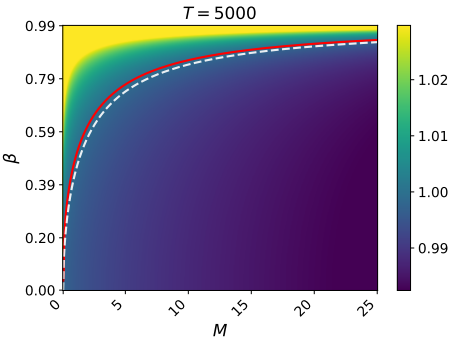

Figure 1: Ratio of the SHB lower bound to the optimized SGD upper bound for $T = 5000$. The red curve marks ratio 1.

The full proof of Theorem 2 is provided in Section F. To complement this asymptotic result, Section I reports finite-horizon Performance Estimation Problem (PEP, Drori and Teboulle [8]) computations. Combining (3) with Theorem 2 yields an unfavorable best-iterate lower/upper-bound comparison for every $\beta \in [7/8, 1)$, with the comparison ratio diverging as $\beta \to 1$; see Section D.2 for the detailed coefficient comparison.

4. Lower Bounds for Signum and Muon

We next consider sign-based and spectral methods with Polyak-style momentum. In the deterministic setting, we prove a worst-case lower bound for Signum and compare it with the optimized upper bound for SignGD. We then transfer the same construction to Muon on diagonal matrix instances.

4.1. Lower Bound for Signum

In this section, we focus on SignGD and Signum in the deterministic setting. We first present a known convergence upper bound for SignGD. For any L-smooth function f, the SignGD iterates with constant step size $\eta > 0$ satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \le \frac{\Delta}{\eta T} + \frac{Ld}{2}\eta, \tag{4}$$

where $\Delta = f(\boldsymbol{x}_0) - f^\star$. The minimum of the right-hand side with respect to $\eta > 0$ is $\sqrt{\frac{2\Delta Ld}{T}}$.

The proof is given in Section C.2. The optimized upper bound in (4) is asymptotically tight; this is also proved in Section C.2.

Remark. Existing optimized upper bound comparisons for Signum are also unfavorable relative to SignGD in the large-momentum regime; see Section C.2.

We next derive a lower bound for Signum under the averaged gradient ℓ_1-norm.

Theorem 3 *Consider the iterates $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ generated by Signum, with $\sigma = 0$. Suppose $T \ge 20 + \frac{1}{1-\beta}$ and $\beta > 0.64$. Then,*

$$\inf_{\eta>0} \sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \ge \sqrt{\frac{3\Delta Ld}{2T}\left(\frac{21}{40} + \frac{35}{128\sqrt{1-\beta}}\right)}.$$

The full proof of Theorem 3 is provided in Section G.3. Combining the optimized SignGD upper bound in (4) with Theorem 3, the Signum lower bound exceeds the optimized SignGD upper bound when $\beta > 0.886$; see Section D.3 for the coefficient comparison.

4.2. Implications for SpecGD vs Muon

The same construction gives a lower bound for Muon. On diagonal matrix instances, the Muon update preserves diagonality and coincides with Signum on the diagonal; moreover, the nuclear norm of the diagonal gradient equals the ℓ_1-norm of the corresponding vector gradient. Therefore, the Signum lower bound in Theorem 3 transfers directly to Muon, with $\|\nabla f(\boldsymbol{X}_t)\|_*$ replacing $\|\nabla f(\boldsymbol{x}_t)\|_1$. The formal reduction and corresponding Muon lower bound are given in Section H.

5. Discussion

Deep Learning Experiments. To contrast our worst-case results with deep learning practice, Section J reports ResNet-18 [13] experiments on CIFAR-10 [19] where SHB can outperform SGD under the averaged squared gradient ℓ_2-norm after tuning the learning rate. Thus, the mismatch points to limitations of the standard smooth non-convex framework rather than the metric itself.

Convergence Metrics. Within the standard smooth non-convex framework, gradient norm-based criteria are hard to avoid: function-value gap is a global-optimality criterion, which is NP-complete to certify in unconstrained non-convex optimization [21], while finite-time guarantees for the last-iterate gradient norm are impossible under the standard assumptions alone [7]. Our results show that these standard stationarity metrics can nevertheless compare momentum unfavorably with its non-momentum counterparts. Thus, the mismatch does not seem to be merely an artifact of averaging over iterates, nor does it appear to be resolved by switching to the standard best-iterate criterion.

Function Class. Another possible source of the mismatch is the function class. Deep learning objectives often violate global L-smoothness [41], but any larger class containing our hard instances inherits the same lower bounds. For the averaged gradient norm lower bounds, these instances are simple isotropic quadratics of the form $\frac{L}{2}\|\boldsymbol{x}\|_2^2$ up to translation, with additive noise for SHB, and have condition number $\kappa = 1$, far from the ill-conditioned quadratic regime where Polyak's heavy-ball method is classically accelerated [25]. Thus, explaining momentum's advantage may require assumptions that distinguish such isotropic geometries from more structured landscapes. As an illustration, Section K gives a Rosenbrock example where curved valley geometry can lead to more favorable comparisons for heavy-ball momentum, consistent with recent views of valley structure in deep learning loss landscapes [29, 30, 37].

Stochasticity. Our SHB comparison also suggests that stochasticity matters: as the noise level increases, the range of β where SHB exhibits the unfavorable comparison narrows. This is consistent with views of heavy-ball momentum as partially filtering gradient noise [5], and suggests that a refined stochastic model may help explain positive effects of momentum. However, stochasticity alone is likely insufficient, as momentum can also be beneficial in full-batch settings; see Section J.

6. Conclusion

We proved worst-case lower bounds showing that standard stationarity-based smooth non-convex analysis can yield comparisons unfavorable to Polyak-style momentum. These separations hold for SHB, Signum, and Muon under averaged (squared) gradient norms, and for deterministic HB under the best-iterate squared gradient norm. This suggests that explaining momentum's practical advantage in deep learning requires refinements beyond the standard framework, such as geometric structure or refined stochastic models.

References

[1] MOSEK ApS. *The MOSEK Python Fusion API manual. Version 11.0.*, 2025. URL https://docs.mosek.com/latest/pythonfusion/index.html.

[2] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pages 560–569. PMLR, 2018.

[3] El Mahdi Chayti and Martin Jaggi. Stochastic difference-of-convex optimization with momentum. *arXiv preprint arXiv:2510.17503*, 2025.

[4] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

[5] Aaron Defazio. Momentum via primal averaging: Theoretical insights and learning rate schedules for non-convex optimization. *arXiv preprint arXiv:2010.00406*, 2020.

[6] Xiaoge Deng, Tao Sun, Dongsheng Li, and Xicheng Lu. Exploring the inefficiency of heavy ball as momentum parameter approaches. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3899–3907, 2024.

[7] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pages 2658–2667. PMLR, 2020.

[8] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

[9] Swetha Ganesh, Rohan Deb, Gugan Thoppe, and Amarjit Budhiraja. Does momentum help in stochastic optimization? A sample complexity analysis. In Robin J. Evans and Ilya Shpitser, editors, *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.

[10] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[11] Baptiste Goujaud, Céline Moucer, François Glineur, Julien M Hendrickx, Adrien B Taylor, and Aymeric Dieuleveut. Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. *Mathematical Programming Computation*, 16(3):337–367, 2024.

[12] Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method: B. goujaud et al. *Mathematical Programming*, pages 1–59, 2025.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[15] E.I. Jury. *Theory and Application of the Z-transform Method*. Wiley, 1964. ISBN 9780471453451. URL https://books.google.co.kr/books?id=NwFRAAAAMAAJ.

[16] Sebastian Kassing and Simon Weissmann. Polyak's heavy ball method achieves accelerated local rate of convergence under polyak-lojasiewicz inequality. *arXiv preprint arXiv:2410.16849*, 2024.

[17] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[20] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d3f5d4de09ea19461dab00590df91e4f-Paper.pdf.

[21] Katta G. Murty and Santosh N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, June 1987. ISSN 1436-4646. doi: 10.1007/bf02592948. URL http://dx.doi.org/10.1007/BF02592948.

[22] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

[23] Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. *arXiv preprint arXiv:2312.14567*, 2023.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[25] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[26] HoHo Rosenbrock. An automatic method for finding the greatest or least value of a function. *The computer journal*, 3(3):175–184, 1960.

[27] Teodor Rotaru, François Glineur, and Panagiotis Patrinos. Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions. *arXiv preprint arXiv:2406.17506*, 2024.

[28] Sharan Sahu, Cameron J Hogan, and Martin T Wells. On the provable suboptimality of momentum sgd in nonstationary stochastic optimization. *arXiv preprint arXiv:2601.12238*, 2026.

[29] Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does sgd really happen in tiny subspaces? In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Minhak Song, Beomhan Baek, Kwangjun Ahn, and Chulhee Yun. Through the river: Understanding the benefit of schedule-free methods for language model training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=CGx4XU9rCA.

[31] Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33077–33099. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sun23l.html.

[32] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *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.

[33] G. Szegő. *Orthogonal Polynomials*. American Math. Soc: Colloquium publ. American Mathematical Society, 1975. ISBN 9780821810231. URL https://books.google.co.kr/books?id=ZOhmnsXlcY0C.

[34] Wei Tao, Sheng Long, Gaowei Wu, and Qing Tao. The role of momentum parameters in the optimal convergence of adaptive polyak's heavy-ball methods. *arXiv preprint arXiv:2102.07314*, 2021.

[35] Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1):307–345, 2017.

[36] Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3JjJezzVkT.

[37] Kaiyue Wen, Zhiyuan Li, Jason S. Wang, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In

The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=m51BgoqvbP.

[38] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, IJCAI-2018, page 2955–2961. International Joint Conferences on Artificial Intelligence Organization, July 2018. doi: 10.24963/ijcai.2018/410. URL http://dx.doi.org/10.24963/ijcai.2018/410.

[39] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization, 2016. URL https://arxiv.org/abs/1604.03257.

[40] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[41] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgnXpVYwS.

Contents

Appendix A. Related Work

Classical results show that heavy-ball momentum achieves acceleration over gradient descent on ill-conditioned quadratic objectives, with an improved dependence on the condition number [25]. Beyond this classical setting, many positive results for heavy-ball momentum typically rely on additional structural assumptions, and the type of benefit varies across settings. For example, heavy-ball momentum has been shown to improve convergence in quadratic objectives with anisotropic gradient noise [23] and to accelerate locally under the Polyak-Łojasiewicz condition around global minima [16]. In stochastic difference-of-convex optimization, momentum can also be necessary for convergence under standard smoothness and bounded-variance assumptions with small batch sizes [3]. For sign-based and normalized methods, momentum has also been shown to improve convergence behavior [4, 31]. In constrained convex optimization with time-varying parameters, momentum has also been shown to play a role in achieving optimal convergence guarantees [34]. These works demonstrate that momentum can be useful under suitable structure, while also suggesting that such benefits depend on assumptions beyond smooth non-convexity alone.

On the other hand, several studies show that momentum does not always improve convergence and can even be unfavorable in some settings. For instance, Goujaud et al. [12] show that HB does not yield an acceleration for the class of smooth strongly-convex functions. In stochastic regimes, Kidambi et al. [17] demonstrate that for certain quadratic objectives, standard momentum methods fail to provide any acceleration over gradient descent. Ganesh et al. [9] highlight that momentum can be insufficient to improve convergence rates in the presence of gradient noise. Wang et al. [36] further note that the advantage of momentum is marginal in regimes where the learning rate is small and gradient noise dominates. More closely related to slowdown phenomena, Deng et al. [6] analyze convex quadratic objectives and show that heavy-ball momentum can slow down SGD as $\beta \to 1$ in a particular step size regime. Furthermore, Sahu et al. [28] demonstrate that in non-stationary environments, momentum can amplify drift-induced tracking error, making it provably suboptimal compared to vanilla SGD.

Together, these works show that heavy-ball momentum can be beneficial under additional structure, but that it does not necessarily provide a uniform improvement across broad problem classes. Our work complements this literature by showing that the standard stationarity-based smooth non-convex setting can certify lower/upper-bound comparisons that are unfavorable to momentum. This highlights a limitation of the standard framework in explaining momentum's empirical benefits.

Appendix B. Additional Preliminaries

Notation. For vectors, we write $\|\cdot\|_p$ for the ℓ_p-norm, and for matrices, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_*$ denote the Frobenius norm and the nuclear norm, respectively. The sign function $\mathrm{sign}(\cdot)$ is applied element-wise. We denote by $f^\star = \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, let $\mathrm{SVD}(\boldsymbol{X}) = (\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V})$ denote the singular value decomposition, where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ have orthonormal columns, $\boldsymbol{S} \in \mathbb{R}^{r\times r}$ is diagonal with nonnegative entries, and $r = \mathrm{rank}(\boldsymbol{X})$. We write $\boldsymbol{0}_{m\times n}$ for the zero matrix in $\mathbb{R}^{m\times n}$ and $\boldsymbol{I}_m$ for the identity matrix in $\mathbb{R}^{m\times m}$. We denote by $\boldsymbol{1}_d \in \mathbb{R}^d$ the vector of all ones, and write $\boldsymbol{1}$ when the dimension is clear from context. For $\boldsymbol{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$, we write $\mathrm{diag}(a_1, \ldots, a_d)$, or equivalently $\mathrm{diag}(\boldsymbol{a})$, for the $d \times d$ diagonal matrix whose (i, i)-th entry is a_i.

B.1. Problem Settings

We consider the unconstrained minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),$$

where the objective function f is smooth and possibly non-convex, and accessible through a (stochastic) first-order oracle.

We first define smoothness and the function class.

Definition 4 (Smoothness) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth for some $L > 0$ if*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. When f is defined over a matrix domain, we replace $\|\cdot\|_2$ by $\|\cdot\|_{\mathrm{F}}$.

Definition 5 (Function class) *For some $\Delta > 0$ and $L > 0$, we define*

$$\mathcal{F}_L(\Delta) := \{f : \mathbb{R}^d \to \mathbb{R} : f(\boldsymbol{0}) - f^\star \le \Delta,\ f \text{ is } L\text{-smooth}\}.$$

The choice of the origin as the initial point is without loss of generality. Indeed, any instance with objective g and initial point $\boldsymbol{x}_0$ satisfying $g(\boldsymbol{x}_0) - g^\star \le \Delta$ can be translated to an instance in $\mathcal{F}_L(\Delta)$ by defining $f(\boldsymbol{x}) := g(\boldsymbol{x} + \boldsymbol{x}_0)$.

We assume access to a stochastic first-order oracle that returns $g(\boldsymbol{x}; \xi)$, when queried at $\boldsymbol{x} \in \mathbb{R}^d$. We make the following assumptions on the stochastic oracle. When $\sigma = 0$, the stochastic oracle reduces to the deterministic first-order oracle.

Assumption 6 (Unbiasedness) *The stochastic oracle is unbiased, i.e., $\mathbb{E}\left[g(\boldsymbol{x}; \xi)\right] = \nabla f(\boldsymbol{x})$.*

Assumption 7 (Bounded variances) *There exists $\sigma \ge 0$ such that $\mathbb{E}\left[\|g(\boldsymbol{x}; \xi) - \nabla f(\boldsymbol{x})\|_2^2\right] \le \sigma^2$.*

Note that when $\sigma = 0$, the stochastic oracle reduces to the deterministic first-order oracle.

B.2. Worst-Case Performance

To compare optimization algorithms independently of specific parameter choices, we adopt a worst-case performance criterion based on a minimax formulation. Given an algorithm class $\mathcal{A}$ and a function class $\mathcal{F}$, we measure performance by

$$\inf_{A\in\mathcal{A}} \sup_{f\in\mathcal{F}} \mathbb{M}(f, A),$$

where $\mathbb{M}(f, A)$ denotes a convergence metric for algorithm A on function f, such as

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q], \qquad \min_{0\le t\le T-1} \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q].$$

The infimum and supremum represent optimal algorithmic parameter tuning and the worst-case instance within the function class, respectively.

In this work, $\mathcal{A}$ consists of heavy-ball momentum-based first-order methods (*e.g.*, SHB, Signum, or Muon) parameterized by the step size, while the momentum parameter β is treated as fixed. The function class $\mathcal{F}$ is given by $\mathcal{F}_L(\Delta)$. Our comparison focuses on contrasting the cases $\beta = 0$ and $\beta > 0$, corresponding to algorithms without and with momentum, respectively.

Appendix C. Existing SHB/Signum Upper Bound Comparisons

C.1. SGD/SHB Upper Bound

SGD Upper Bound. As shown in Section 3.1, the SGD upper bound of Ghadimi and Lan [10] is given by

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right] \leq \frac{\Delta}{T\eta(1-\eta L/2)} + \frac{\eta L\sigma^2}{2-\eta L},$$

and the minimum of the upper bound over $\eta > 0$ is

$$\text{SGD UB} := \frac{\Delta L}{T}\left(1+\sqrt{1+\frac{2\sigma^2 T}{\Delta L}}\right).$$

Moreover, there exists an L-smooth function and a stochastic first-order oracle for which the bound in (1), when stated with $f(\boldsymbol{x}_0) - \mathbb{E}[f(\boldsymbol{x}_T)]$ in place of Δ, holds with *equality*. We prove this below.

Proof

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $f(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x}\|_2^2$ and consider the stochastic oracle $g(\boldsymbol{x};\xi) = L\boldsymbol{x} + \boldsymbol{\xi}$, where

$$\mathbb{P}\left(\boldsymbol{\xi} = +\frac{\sigma}{\sqrt{d}}\mathbf{1}_d\right) = \mathbb{P}\left(\boldsymbol{\xi} = -\frac{\sigma}{\sqrt{d}}\mathbf{1}_d\right) = \frac{1}{2}.$$

Then, Assumptions 6 and 7 are satisfied.

We also have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &= f(\boldsymbol{x}_t - \eta\boldsymbol{g}_t) \\ &= f(\boldsymbol{x}_t) - \eta\nabla f(\boldsymbol{x}_t)^\top \boldsymbol{g}_t + \frac{\eta^2 L}{2}\|\boldsymbol{g}_t\|_2^2. \end{aligned} \tag{5}$$

Moreover, it holds that

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)\|_2^2] &= \mathbb{E}[\|\boldsymbol{g}_t\|_2^2] + \|\nabla f(\boldsymbol{x}_t)\|_2^2 - 2\nabla f(\boldsymbol{x}_t)^\top \mathbb{E}[\boldsymbol{g}_t] \\ &= \mathbb{E}[\|\boldsymbol{g}_t\|_2^2] - \|\nabla f(\boldsymbol{x}_t)\|_2^2 \end{aligned}$$

and

$$\mathbb{E}[\|\boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)\|_2^2] = \mathbb{E}[\|\boldsymbol{\xi}\|_2^2] = \sigma^2,$$

which implies that $\mathbb{E}[\|\boldsymbol{g}_t\|_2^2] = \sigma^2 + \mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2]$.

Taking expectation and rearranging (5) yields

$$\eta\left(1-\frac{\eta L}{2}\right)\|\nabla f(\boldsymbol{x}_t)\|_2^2 = f(\boldsymbol{x}_t) - \mathbb{E}[f(\boldsymbol{x}_{t+1}) \mid \boldsymbol{x}_t] + \frac{\eta^2 L\sigma^2}{2}.$$

Therefore,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right] = \frac{f(\boldsymbol{x}_0) - \mathbb{E}[f(\boldsymbol{x}_T)]}{T\eta(1-\frac{\eta L}{2})} + \frac{\eta L\sigma^2}{2-\eta L}.$$

■

SHB Upper Bound. We recall a representative convergence guarantee for the SHB method under the same assumptions. The following result of Liu et al. [20] states an upper bound for SHB.

Proposition 8 (Liu et al. [20], Theorem 1) *Let $f \in \mathcal{F}_L(\Delta)$ and $0 < \eta \leq \frac{1}{L(4-\beta+\beta^2)}$.*[2] *Suppose that the stochastic first-order oracle g satisfies Assumptions 6 and 7. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ denote the iterates generated by SHB. Then,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] \leq \frac{2\Delta}{T\eta(1-\beta)} + \left(\frac{\beta+5\beta^2}{8(1+\beta)}+1\right)(1-\beta)\eta L\sigma^2. \tag{6}$$

Remark. Note that Liu et al. [20] adopt the update $\boldsymbol{m}_{t+1} = \beta\boldsymbol{m}_t + (1-\beta)\boldsymbol{g}_t$. Consequently, their step size parameter (denoted as α in their work) corresponds to $\eta(1-\beta)$ in our notation.

The upper bound (6) is minimized at

$$\eta = \begin{cases} \frac{4}{\sigma(1-\beta)}\sqrt{\frac{\Delta(1+\beta)}{LT(5\beta^2+9\beta+8)}}, & \text{if } \frac{4}{1-\beta}\sqrt{\frac{\Delta(1+\beta)}{LT(5\beta^2+9\beta+8)}} \leq \frac{\sigma}{L(4-\beta+\beta^2)}, \\ \frac{1}{L(4-\beta+\beta^2)}, & \text{otherwise,} \end{cases}$$

and its minimum is

$$\begin{cases} \sigma\sqrt{\frac{5\beta^2+9\beta+8}{\beta+1}\frac{\Delta L}{T}}, & \text{if } \frac{4}{1-\beta}\sqrt{\frac{\Delta(1+\beta)}{LT(5\beta^2+9\beta+8)}} \leq \frac{\sigma}{L(4-\beta+\beta^2)}, \\ \frac{2\Delta L(\beta^2-\beta+4)}{T(1-\beta)} + \sigma^2\frac{(1-\beta)\left(5\beta^2+9\beta+8\right)}{8(\beta+1)(\beta^2-\beta+4)}, & \text{otherwise.} \end{cases}$$

Thus, if

$$T > \frac{16\Delta L}{\sigma^2(1-\beta)^2}\frac{(1+\beta)(4-\beta+\beta^2)}{5\beta^2+9\beta+8},$$

then the minimized upper bound becomes

$$\text{SHB UB} := \sigma\sqrt{\frac{5\beta^2+9\beta+8}{\beta+1}\frac{\Delta L}{T}}.$$

Comparison. The ratio between SHB upper bound and SGD upper bound is

$$\frac{\text{SHB UB}}{\text{SGD UB}} = \frac{\sigma\sqrt{\frac{5\beta^2+9\beta+8}{\beta+1}}}{\sqrt{\frac{\Delta L}{T}}+\sqrt{\frac{\Delta L}{T}+2\sigma^2}} \to \sqrt{\frac{5\beta^2+9\beta+8}{2(\beta+1)}}$$

as $T \to \infty$.

For all $\beta \in [0,1)$, this ratio remains strictly greater than 1. In particular, the ratio equals 2 at $\beta = 0$ and monotonically approaches $\sqrt{5.5} \approx 2.34$ as $\beta \to 1$. This indicates that the upper bound for SHB by Liu et al. [20] is consistently larger than that of SGD by a constant factor. This observation is consistent with other existing analyses [38–40].

The discrepancy becomes even more significant in the deterministic setting. In this case, the optimized upper bound for SGD is $\frac{2\Delta L}{T}$, whereas the bound for SHB in Proposition 8 reduces to $\frac{2\Delta L}{T}\frac{\beta^2-\beta+4}{1-\beta}$. Comparing these deterministic bounds, the gap between the two methods scales as $\mathcal{O}((1-\beta)^{-1})$.

2. While the original condition on the step size in Liu et al. [20] involves a minimum of two terms, the term $\frac{1}{L(4-\beta+\beta^2)}$ is the stronger constraint for all $\beta \in [0,1)$, which allows us to simplify the step size requirement to $\eta \leq \frac{1}{L(4-\beta+\beta^2)}$.

C.2. Signum Upper Bound

SignGD Upper Bound. As shown in Section 4.1, the SignGD upper bound is given by

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \leq \frac{\Delta}{\eta T} + \frac{Ld}{2}\eta,$$

where $\Delta = f(\boldsymbol{x}_0) - f^\star$, and the minimum of the upper bound over $\eta > 0$ is

$$\text{SignGD UB} := \sqrt{\frac{2\Delta Ld}{T}}.$$

We provide the proof of the upper bound below.

Proof By the descent lemma, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Substituting $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\,\mathrm{sign}(\nabla f(\boldsymbol{x}_t))$ yields

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) - \eta\nabla f(\boldsymbol{x}_t)^\top \mathrm{sign}(\nabla f(\boldsymbol{x}_t)) + \frac{L}{2}\eta^2\|\,\mathrm{sign}(\nabla f(\boldsymbol{x}_t))\|_2^2 \\ &\leq f(\boldsymbol{x}_t) - \eta\|\nabla f(\boldsymbol{x}_t)\|_1 + \frac{\eta^2 Ld}{2}. \end{aligned}$$

Rearranging and summing over $t = 0, \ldots, T-1$:

$$\eta\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \leq \sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1})) + \frac{\eta^2 LdT}{2}.$$

Dividing by ηT:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \leq \frac{\Delta}{\eta T} + \frac{Ld}{2}\eta.$$

By minimizing the right-hand side, we have

$$\frac{\Delta}{\eta T} + \frac{Ld}{2}\eta \geq \sqrt{\frac{2\Delta Ld}{T}}.$$

■

We also prove the asymptotic tightness of the upper bound.

Proof Assume $T > 2$. By Lemma 42, we have

$$\sup_{f\in\mathcal{F}_L(\Delta)}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\tilde{\boldsymbol{x}}_t)\|_1 \geq \Lambda_1(\eta),$$

where

$$\Lambda_1(\eta) := \begin{cases} \sqrt{2\Delta Ld} - \frac{Ld}{2}(T-1)\eta, & \text{if } \eta \leq \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}, \\ \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(1 - \frac{1}{T}\right)\eta, & \text{if } \eta > \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}. \end{cases}$$

The first branch is decreasing in η, so its minimum over its regime is attained at

$$\eta_\star = \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}$$

and equals

$$\frac{1}{2}\sqrt{2\Delta Ld}.$$

For the second branch, we have

$$\min_{\eta>0}\left\{\frac{\Delta}{\eta T} + \frac{Ld}{2}\left(1 - \frac{1}{T}\right)\eta\right\} = \sqrt{\frac{2\Delta Ld}{T}}\sqrt{1 - \frac{1}{T}},$$

and the minimum is attained at

$$\eta_\star = \sqrt{\frac{2\Delta}{Ld(T-1)}}.$$

Thus, we have

$$\min_{\eta>0}\Lambda_1(\eta) = \sqrt{\frac{2\Delta Ld}{T}}\sqrt{1 - \frac{1}{T}}$$

because $\sqrt{\frac{2\Delta}{Ld}\frac{1}{T-1}} > \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}$ for all $T > 2$. Thus, the ratio between the minimized upper bound and the corresponding lower bound (each optimized with respect to η) is $\geq \sqrt{1 - \frac{1}{T}}$. ■

Signum Upper Bound. We recall a convergence guarantee for Signum, via the proof framework in Sun et al. [31].

Proposition 9 *Suppose that f is L-smooth. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ denote the iterates generated by Signum, with $\sigma = 0$. Then, with $\Delta = f(\boldsymbol{x}_0) - f^\star$,*

$$\frac{1}{T}\sum\nolimits_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \leq \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(\frac{1+3\beta}{1-\beta}\right)\eta + \frac{2\sqrt{2\Delta Ld}}{T}\frac{\beta}{1-\beta}, \tag{7}$$

The minimum of the right-hand side with respect to $\eta > 0$ is $\sqrt{\frac{2\Delta Ld}{T}\frac{1+3\beta}{1-\beta}} + \frac{2\sqrt{2\Delta Ld}}{T}\frac{\beta}{1-\beta}$.

Proof By the descent lemma, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Substituting $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \operatorname{sign}(\boldsymbol{m}_{t+1})$ yields

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) - \eta \nabla f(\boldsymbol{x}_t)^\top \operatorname{sign}(\boldsymbol{m}_{t+1}) + \frac{L}{2}\eta^2 \| \operatorname{sign}(\boldsymbol{m}_{t+1})\|_2^2 \\ &\leq f(\boldsymbol{x}_t) - \eta \nabla f(\boldsymbol{x}_t)^\top \operatorname{sign}(\nabla f(\boldsymbol{x}_t)) \\ &\quad + \eta \nabla f(\boldsymbol{x}_t)^\top (\operatorname{sign}(\nabla f(\boldsymbol{x}_t)) - \operatorname{sign}(\boldsymbol{m}_{t+1})) + \frac{\eta^2 L d}{2} \\ &= f(\boldsymbol{x}_t) - \eta \|\nabla f(\boldsymbol{x}_t)\|_1 + \eta \nabla f(\boldsymbol{x}_t)^\top (\operatorname{sign}(\nabla f(\boldsymbol{x}_t)) - \operatorname{sign}(\boldsymbol{m}_{t+1})) + \frac{\eta^2 L d}{2}. \end{aligned} \tag{8}$$

Now, we show that $x(\operatorname{sign}(x) - \operatorname{sign}(y)) \leq 2|x - y|$ for any $x, y \in \mathbb{R}$. If $\operatorname{sign}(x) = \operatorname{sign}(y)$, then $\operatorname{sign}(x) - \operatorname{sign}(y) = 0$ and the inequality holds. Otherwise, we have $xy \leq 0$, which implies $|x - y| \geq |x|$. Thus, we have

$$\nabla f(\boldsymbol{x}_t)^\top (\operatorname{sign}(\nabla f(\boldsymbol{x}_t)) - \operatorname{sign}(\boldsymbol{m}_{t+1})) \leq 2 \|\nabla f(\boldsymbol{x}_t) - \boldsymbol{m}_{t+1}\|_1.$$

For $t \geq 0$, let $\boldsymbol{z}_t := (1-\beta)\boldsymbol{m}_{t+1} - \nabla f(\boldsymbol{x}_t)$ and for $t \geq 1$, let $\boldsymbol{s}_t := \nabla f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t)$. Then,

$$\begin{aligned} \boldsymbol{z}_t &= (1-\beta)\boldsymbol{m}_{t+1} - \nabla f(\boldsymbol{x}_t) \\ &= \beta \left((1-\beta)\boldsymbol{m}_t - \nabla f(\boldsymbol{x}_{t-1}) + \boldsymbol{s}_t \right) \\ &= \beta \boldsymbol{z}_{t-1} + \beta \boldsymbol{s}_t, \end{aligned}$$

and consequently $\boldsymbol{z}_t = \beta^t \boldsymbol{z}_0 + \sum_{i=1}^t \beta^{t+1-i} \boldsymbol{s}_i$. By the L-smoothness, we have

$$\begin{aligned} \|\boldsymbol{z}_t\|_2 &\leq \beta^t \|\boldsymbol{z}_0\|_2 + \sum_{i=1}^t \beta^{t+1-i} \|\boldsymbol{s}_i\|_2 \\ &\leq \beta^t \| - \beta \nabla f(\boldsymbol{x}_0)\|_2 + \eta L \sqrt{d} \frac{\beta}{1-\beta}. \end{aligned}$$

Summing (8) from $t = 0, \ldots, T-1$ yields

$$f(\boldsymbol{x}_T) \leq f(\boldsymbol{x}_0) - \eta \sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 + 2\sqrt{d}\eta \sum_{t=0}^{T-1} \|\boldsymbol{z}_t\|_2 + \frac{\eta^2 L d T}{2}.$$

Dividing by ηT, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \leq \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(\frac{1+3\beta}{1-\beta}\right)\eta + \frac{2\sqrt{d}}{T}\frac{\beta}{1-\beta}\|\nabla f(\boldsymbol{x}_0)\|_2.$$

By the L-smoothness, we have

$$f(\boldsymbol{x}_0) - f^\star \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_0)\|_2^2,$$

so we have $\|\nabla f(\boldsymbol{x}_0)\|_2 \leq \sqrt{2\Delta L}$. Thus,

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \leq \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(\frac{1+3\beta}{1-\beta}\right)\eta + \frac{2\sqrt{2\Delta L d}}{T}\frac{\beta}{1-\beta}.$$

Minimizing the right-hand side gives

$$\sqrt{\frac{2\Delta L d}{T}\left(\frac{1+3\beta}{1-\beta}\right)} + \frac{2\sqrt{2\Delta L d}}{T}\frac{\beta}{1-\beta}.$$

■

Comparison. Under optimally tuned step sizes for each method, the leading $T^{-1/2}$ term in Proposition 9 is larger than the optimized SignGD upper bound in (4) by the factor $\sqrt{\frac{1+3\beta}{1-\beta}}$. Thus, the known upper bound comparison becomes increasingly unfavorable to Signum in the large momentum regime.

Appendix D. Lower/Upper-Bound Comparisons

D.1. SGD vs SHB under Averaged Squared Gradient ℓ_2-Norm

To compare the SGD upper bound and the SHB lower bound, we note that

$$\text{SGD UB} = \frac{\Delta L}{T}\left(1+\sqrt{1+4MT}\right), \quad \text{SHB LB} = \frac{\Delta L}{T}\left(1-\frac{5.05}{T}\right)\left(\frac{1+\beta^2}{1-\beta^2}-\frac{M}{2}+\sqrt{\Psi}\right),$$

where $\Psi = 1 - 2M\frac{1+\beta}{1-\beta} + 4MT$. Thus, for sufficiently large T, we have $1 - \frac{5.05}{T} \approx 1$ and $\Psi \approx 1 + 4MT$. In this regime, the comparison between SHB and SGD is therefore governed by whether $\frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}$ exceeds the constant term 1 in the SGD bound. We characterize the resulting asymptotic boundary between SHB and SGD for large T in the following corollary.

Corollary 10 *Let $M = \frac{\sigma^2}{2\Delta L} \geq 0$ be fixed and let $\beta \in [0,1)$ satisfy $\frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} > 1$. Then, there exists $T_0 > 0$ such that for any $T > T_0$, the lower bound in Theorem 1 on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|_2^2\right]$ for SHB exceeds the corresponding upper bound for SGD.*

Proof We first show that the additional assumption $\beta \geq \frac{M}{4}$ for $0 < M < 1$ in Theorem 1 holds when $\frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} > 1$, which is equivalent to $\beta > \sqrt{\frac{M}{M+4}}$. Since $\sqrt{\frac{M}{M+4}} > \frac{M}{4}$ if $0 < M < 1$, we have $\beta > \sqrt{\frac{M}{M+4}}$ implies $\beta \geq \frac{M}{4}$ when $0 < M < 1$.

We assume that

$$T \geq \max\left\{25, \frac{1}{1-\beta}\left(5\log\frac{1}{1-\beta}+9\right), 39M+10\right\},$$

as in Theorem 1. Let

$$\delta := \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} - 1.$$

Since

$$\text{SHB LB} - \text{SGD UB} = \delta + \left(\sqrt{\Psi}-\sqrt{1+4MT}\right) - \frac{5.05}{T}\left(1+\delta+\sqrt{\Psi}\right),$$

it suffices to show that

$$X := \sqrt{1+4MT} - \sqrt{\Psi} + \frac{5.05}{T}\left(1+\delta+\sqrt{\Psi}\right) < \frac{\delta}{2}.$$

First, we have

$$\begin{aligned}\sqrt{1+4MT} - \sqrt{\Psi} &= \frac{2M\frac{1+\beta}{1-\beta}}{\sqrt{1+4MT}+\sqrt{\Psi}} \\ &\leq M\frac{1+\beta}{1-\beta}\frac{1}{\sqrt{\Psi}},\end{aligned}$$

because $\Psi \leq 1 + 4MT$. If $T \geq \frac{1+\beta}{1-\beta}$, we have $\Psi \geq 2MT$. Thus,

$$\sqrt{1+4MT} - \sqrt{\Psi} \leq \frac{1+\beta}{1-\beta}\sqrt{\frac{M}{2}}\frac{1}{\sqrt{T}}.$$

Second, using the fact that $\sqrt{\Psi} \leq \sqrt{1+4MT} \leq 1+2\sqrt{MT}$, we have

$$\frac{5.05}{T}\left(1+\delta+\sqrt{\Psi}\right) \leq \frac{5.05(2+\delta)}{T} + \frac{10.1\sqrt{M}}{\sqrt{T}}.$$

Thus,

$$X \leq \frac{1+\beta}{1-\beta}\sqrt{\frac{M}{2}}\frac{1}{\sqrt{T}} + \frac{5.05(2+\delta)}{T} + \frac{10.1\sqrt{M}}{\sqrt{T}}.$$

Since

$$\begin{aligned}
\frac{1+\beta}{1-\beta}\sqrt{\frac{M}{2}}\frac{1}{\sqrt{T}} \leq \frac{\delta}{4} &\implies T > \frac{8M}{\delta^2}\left(\frac{1+\beta}{1-\beta}\right)^2 \\
\frac{5.05(2+\delta)}{T} \leq \frac{\delta}{8} &\implies T > \frac{40.4(2+\delta)}{\delta} \\
\frac{10.1\sqrt{M}}{\sqrt{T}} \leq \frac{\delta}{8} &\implies T > \frac{(80.8)^2 M}{\delta^2},
\end{aligned}$$

if

$$T > \max\left\{\frac{1+\beta}{1-\beta}, \frac{8M}{\delta^2}\left(\frac{1+\beta}{1-\beta}\right)^2, \frac{(80.8)^2 M}{\delta^2}, \frac{40.4(2+\delta)}{\delta}\right\},$$

$X < \frac{\delta}{2}$ holds.

Finally, if

$$T_0 = \left\{25, \frac{1}{1-\beta}\left(5\log\frac{1}{1-\beta}+9\right), 39M+10, \frac{8M}{\delta^2}\left(\frac{1+\beta}{1-\beta}\right)^2, \frac{(80.8)^2 M}{\delta^2}, \frac{40.4(2+\delta)}{\delta}\right\},$$

then $T > T_0$ implies SHB LB $>$ SGD UB, as desired.

■

D.2. GD vs HB under Best-Iterate Squared Gradient Norm

We recall that the GD upper bound in (3) is

$$\frac{6\sqrt{3}\Delta L}{8(T-1)+3\sqrt{3}},$$

while the HB lower bound in Theorem 2 is

$$\frac{\Delta L}{4(1-\beta)}.$$

Combining the GD upper bound (3) and the HB lower bound in Theorem 2, gradient descent admits a best-iterate upper bound with leading T^{-1} coefficient at most $\frac{3\sqrt{3}}{4}\Delta L$, whereas HB has a best-iterate lower bound with coefficient at least $\frac{\Delta L}{4(1-\beta)}$, both under their respective optimal step sizes. Hence the lower/upper bound comparison ratio is asymptotically at least $\frac{1}{3\sqrt{3}(1-\beta)}$. For every $\beta \in [7/8, 1)$, this ratio is at least $\frac{8}{3\sqrt{3}} > 1$, and it diverges as $\beta \to 1$. Thus, the pessimistic lower/upper-bound comparison for heavy-ball momentum is not solely an artifact of using the averaged gradient norm.

D.3. SignGD vs Signum under Averaged Gradient ℓ_1-Norm

We compare the SignGD upper bound and the Signum lower bound under the optimal step sizes. We have

$$\text{SignGD UB} = \sqrt{\frac{2\Delta L d}{T}}, \quad \text{Signum LB} = \sqrt{\frac{2\Delta L d}{T}}\sqrt{\frac{63}{160} + \frac{105}{512\sqrt{1-\beta}}}.$$

Therefore, if $\beta > 0.886$, the Signum lower bound exceeds the corresponding SignGD upper bound, and the ratio between the two bounds scales as $\mathcal{O}((1-\beta)^{-1/4})$.

Remark. The exponent $1/4$ in the lower bound comparison does not close the gap with the $1/2$ dependence suggested by the leading term in Proposition 9. While closing this exponent gap is an intriguing open problem, we expect it to be technically nontrivial because Signum is highly sensitive to the step size; see Section G.4 for more discussion.

Appendix E. Proof of Theorem 1

E.1. Proof Sketch

The proof proceeds by reducing the lower-bound construction to the analysis of a scalar linear recurrence. We first restrict the supremum to translated one-dimensional quadratics. By translation invariance, it is enough to consider the quadratic $f(x) = Lx^2/2$ with an initial point satisfying $x_0^2 \leq 2\Delta/L$. We then use the stochastic oracle $g(x;\xi) = Lx + \xi$, where the noise variables are independent and have mean 0 and variance σ^2. The SHB iterates satisfy a second-order linear recurrence, and the iterate x_k decomposes into a deterministic heavy-ball part y_k and a noise-response part governed by coefficients d_k. Both y_k and d_k satisfy the same homogeneous recurrence. After normalization and maximizing over the allowed initial gap, the lower-bound objective becomes

$$2\Delta L \left(\frac{Y_T}{T} + (\eta L)^2 M \left(D_T - \frac{E_T}{T} \right) \right),$$

where $M = \frac{\sigma^2}{2\Delta L}$ and the quantities Y_T, D_T, and E_T are partial sums associated with the normalized deterministic trajectory and the noise sequence.

The main technical task is therefore to lower bound this normalized expression over the step size. We parameterize the recurrence by $z = \frac{1-\eta L+\beta}{2\sqrt{\beta}}$. The stable step-size range corresponds to $|z| < B$, where $B = \frac{1+\beta}{2\sqrt{\beta}}$. This parameterization separates the analysis according to the nature of the characteristic roots: positive real roots, complex roots, and negative real roots. In these regions, we derive region-specific upper bounds on the tail sums and combine them with closed forms for the infinite sums Y_∞, D_∞, and E_∞.

The proof then has three parts. First, in the positive-real-root region $z \in [1, B)$, we control the tails using log-concavity of the Chebyshev polynomial representations of y_k^2 and d_k^2 as functions of z. Evaluating the logarithmic derivatives at the endpoint $z = B$ gives exponential tail bounds. Second, in the complex-root region $z \in [-1, 1]$, we use the uniform bound $|U_k(z)| \leq k + 1$ for Chebyshev polynomials of the second kind, which yields polynomial factors multiplied by β^T. Third, in the negative-real-root region $z \in (-B, -1]$, monotonicity in z shows that these larger step sizes cannot improve the lower-bound objective; hence this region can be ruled out by comparison with the boundary behavior.

It remains to identify where the minimum over step sizes can occur. We use the change of variables $u = \frac{B-z}{B+z}$, under which the infinite-sum surrogate becomes a rational function of the form

$$F(u) = \frac{c_{-1}}{u} + c_0 + c_1 u + c_2 u^2.$$

Very small values of u correspond to very small step sizes, and we exclude this range by a direct lower bound. On the remaining part of the positive-real-root and complex-root regions, the tail terms are shown to be smaller than a $5/T$ of the infinite-sum surrogate. The minimization of F is then compared to the simpler function $G(u) = \frac{c_{-1}}{u} + c_0 + c_1 u$, whose minimum is explicit. The difference between the minima of F and G is controlled by a perturbation estimate.

Combining these estimates yields the desired uniform lower bound over all stable step sizes. The constants in the lower bound arise from making the tail estimates, the small-step exclusion, and the perturbation comparison simultaneous and uniform over β, M, and T. The final bound has the form

$$\frac{1}{2T} \left(1 - \frac{5.05}{T} \right) \left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} \right),$$

where $\Psi = 1 - 2M\frac{1+\beta}{1-\beta} + 4MT$. Multiplying back by the normalization factor $2\Delta L$ completes the proof of Theorem 1.

E.2. Setup and Preliminary Reductions

Let $\mathcal{Q}_L(\Delta) \subset \mathcal{F}_L(\Delta)$ be a class of quadratic functions:

$$\mathcal{Q}_L(\Delta) := \left\{ \frac{L}{2}(x+\alpha)^2 : \alpha^2 \leq \frac{2\Delta}{L} \right\}.$$

Then,

$$\sup_{f \in \mathcal{F}_L(\Delta)} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \geq \sup_{f \in \mathcal{Q}_L(\Delta)} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right].$$

Suppose $f(x) = \frac{L}{2}x^2$ and $g(x) = \frac{L}{2}(x+\alpha)^2$. Let $\{x_t\}_{t=0}^{T-1}$ be generated by SHB on f with $x_0 = \alpha$ and $\{\tilde{x}_t\}_{t=0}^{T-1}$ be generated by SHB on g with $\tilde{x}_0 = 0$. Then,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla g(\tilde{x}_t)\|_2^2\right].$$

Thus, without loss of generality, we may assume that $x_0^2 \leq \frac{2\Delta}{L}$.

Consider the quadratic function $f(x) = \frac{L}{2}x^2$ and the stochastic oracle $g(x;\xi) = Lx + \xi$, where

$$\mathbb{P}(\xi = +\sigma) = \mathbb{P}(\xi = -\sigma) = \frac{1}{2}.$$

Then the oracle satisfies Assumptions 6 and 7 with variance σ^2. Moreover, the random variables $\{\xi_t\}_{t\geq 0}$ are assumed to be independent across iterations.

Then,

$$x_{k+1} = (1 - \eta L + \beta)x_k - \beta x_{k-1} + \eta\xi_k$$

with $x_{-1} = x_0$.

Now, we define y_k as

$$y_{-1} = y_0 = x_0, \quad y_{k+1} = (1 - \eta L + \beta)y_k - \beta y_{k-1},$$

which represents the *deterministic* heavy-ball method iterations.

Recurrence relation. We then prove the following lemma, which shows the recurrence relation of $\{x_k\}$:

Lemma 11

$$x_k = y_k + \eta \sum_{i=0}^{k-1} d_{k-i}\xi_k, \tag{9}$$

where $d_0 = 0, d_1 = 1$ *and* $d_n = (1 - \eta L + \beta)d_{n-1} - \beta d_{n-2}$ *for* $n \geq 2$.

Proof

We use induction.

Base case.

$$\begin{aligned} x_1 &= (1-\eta L+\beta)x_0 - \beta x_{-1} + \eta\xi_0 \\ &= (1-\eta L)x_0 + \eta\xi_0 \\ &= (1-\eta L)y_0 + \eta\xi_0 \\ &= y_1 + \eta d_1 \xi_0. \end{aligned}$$

Inductive step. We now assume that (9) holds for all $k \leq n$. Then,

$$\begin{aligned} x_{n+1} &= (1-\eta L+\beta)y_k + \eta(1-\eta L+\beta)\sum_{i=0}^{n-1} d_{n-i}\xi_i + \eta\xi_n - \beta\eta\sum_{i=0}^{n-2} d_{n-1-i}\xi_i \\ &= y_{n+1} + \eta\left((1-\eta L+\beta)\sum_{i=0}^{n-1} d_{n-i}\xi_i + \xi_n - \beta\sum_{i=0}^{n-2} d_{n-1-i}\xi_i\right) \\ &= y_{n+1} + \eta\left(\sum_{i=0}^{n-1}((1-\eta L+\beta)d_{n-i} - \beta d_{n-1-i})\xi_i + (1-\eta L+\beta)d_1\xi_{n-1} + \xi_n\right) \\ &= y_{n+1} + \eta\left(\sum_{i=0}^{n-1} d_{n+1-i}\xi_i + d_2\xi_{n-1} + \xi_n\right) \\ &= y_{n+1} + \eta\left(\sum_{i=0}^{n} d_{n+1-i}\xi_i\right). \end{aligned}$$

Therefore, by the induction, (9) is proved. ■

By the independence of ξ_i, we have

$$\begin{aligned} \mathbb{E}|x_k|^2 &= |y_k|^2 + \eta^2\sigma^2\sum_{i=0}^{k-1}(d_{k-i})^2 \\ &= |y_k|^2 + \eta^2\sigma^2\sum_{j=1}^{k}|d_j|^2 \end{aligned}$$

and consequently

$$\begin{aligned} \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] &= \frac{L^2}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[|x_k|^2\right] \\ &= \frac{L^2}{T}\sum_{k=0}^{T-1}|y_k|^2 + \frac{L^2\eta^2\sigma^2}{T}\sum_{k=0}^{T-1}\sum_{j=1}^{k}|d_j|^2 \\ &= \frac{L^2}{T}\sum_{k=0}^{T-1}|y_k|^2 + \frac{L^2\eta^2\sigma^2}{T}\sum_{j=1}^{T-1}(T-j)|d_j|^2 \\ &= \frac{L^2}{T}\sum_{k=0}^{T-1}|y_k|^2 + \frac{L^2\eta^2\sigma^2}{T}\sum_{j=0}^{T-1}(T-j)|d_j|^2 \quad (\because d_0 = 0). \end{aligned}$$

Normalization. For convenience, let

$$\tilde{x}_k := \frac{x_k}{x_0}, \quad \tilde{y}_k := \frac{y_k}{y_0}.$$

Then,

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] &= \frac{(Lx_0)^2}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[|\tilde{x}_k|^2\right] \\ &= \frac{(Lx_0)^2}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 + \frac{(\eta L)^2\sigma^2}{T}\sum_{j=0}^{T-1}(T-j)|d_j|^2.\end{aligned}$$

Since taking the supremum over $f \in \mathcal{Q}_L(\Delta)$ is equivalent to taking the supremum over x_0 subject to $x_0^2 \le \frac{2\Delta}{L}$, we have

$$\begin{aligned}\sup_{f\in\mathcal{Q}_L(\Delta)}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] &= \sup_{x_0^2\le\frac{2\Delta}{L}}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \\ &= 2\Delta L\left(\frac{1}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 + \frac{(\eta L)^2 M}{T}\sum_{j=0}^{T-1}(T-j)|d_j|^2\right),\end{aligned}$$

where $M := \frac{\sigma^2}{2\Delta L}$.

Thus, after normalization, the behavior is characterized by the deterministic heavy-ball trajectory $\{\tilde{y}_k\}$, together with the noise contribution involving $\{d_k\}$ and scaled by ηL, β, and M.

Convergence condition. Both y_k and d_k satisfy the following recurrence: $c_{k+1} = (1 - \eta L + \beta)c_k - \beta c_{k-1}$. Thus, by analyzing the stability condition of the linear recurrence relation, we can determine the convergence of x_k. The characteristic polynomial is given by

$$\lambda^2 - (1 - \eta L + \beta)\lambda + \beta = 0,$$

with roots

$$\lambda_{1,2} = \frac{1 - \eta L + \beta \pm \sqrt{(1 - \eta L + \beta)^2 - 4\beta}}{2}.$$

The stability of the recurrence is governed by the magnitudes of λ_1 and λ_2. In particular, $c_k \to 0$ as $k \to \infty$ for any initial conditions if and only if

$$|\lambda_1| < 1 \quad \text{and} \quad |\lambda_2| < 1.$$

By the Jury stability criterion [15], the characteristic roots of $\tilde{y}_k$ (and d_k) are in the unit disk (in $\mathbb{C}$) if and only if $|\beta| > 0$ and $1 \pm (1 - \eta L + \beta) + \beta > 0$. The first condition is satisfied because $\beta \in [0, 1)$. The second one is equivalent to $0 < \eta L < 2(1 + \beta)$. Therefore, we will assume that $0 < \eta L < 2(1 + \beta)$.

E.3. Infinite and Tail Sums

Let

$$Y_T = \sum_{k=0}^{T-1} \tilde{y}_k^2, \qquad Y_\infty = \sum_{k=0}^{\infty} \tilde{y}_k^2$$
$$D_T = \sum_{k=0}^{T-1} d_k^2, \qquad D_\infty = \sum_{k=0}^{\infty} d_k^2$$
$$E_T = \sum_{k=0}^{T-1} k d_k^2, \quad E_\infty = \sum_{k=0}^{\infty} k d_k^2.$$

Our goal is to obtain a lower bound for

$$\frac{Y_T}{T} + (\eta L)^2 M \left(D_T - \frac{E_T}{T} \right), \tag{10}$$

minimized over η.

We first prove the following lemma:

Lemma 12 *Define*

$$S := \sum_{k=0}^{\infty} a_k^2, \quad U := \sum_{k=1}^{\infty} a_k a_{k-1}.$$

If a_k satisfies the following recurrence relation:

$$a_{k+1} = 2\sqrt{\beta} z a_k - \beta a_{k-1},$$

then S and U are finite. Moreover,

$$S = \frac{(1+\beta-4\beta(1-\beta)z^2)a_0^2 + (1+\beta)a_1^2 - 4\beta^{3/2} z\, a_0 a_1}{(1-\beta)\left((1+\beta)^2 - 4\beta z^2\right)}$$
$$U = \frac{2\sqrt{\beta} z \left((1+\beta)a_0^2 + a_1^2 - 2\sqrt{\beta} z\, a_0 a_1\right)}{(1-\beta)\left((1+\beta)^2 - 4\beta z^2\right)}.$$

Proof The characteristic equation of the recurrence is $\lambda^2 - 2\sqrt{\beta} z \lambda + \beta = 0$. The condition $|z| < B$ is precisely the region where the magnitudes of both roots, $|\lambda_1|, |\lambda_2|$, are strictly less than 1. Let $\rho = \max(|\lambda_1|, |\lambda_2|) < 1$.

The general solution is $a_k = C_1 \lambda_1^k + C_2 \lambda_2^k$ ($\lambda_1 \neq \lambda_2$) or $(C_1 + kC_2)\lambda_1^k$ ($\lambda_1 = \lambda_2$) for some constants C_1 and C_2, so $|a_k| \leq M_0(k+1)\rho^k$ for some constant $M_0 > 0$. Thus,

$$a_k^2 \leq M_0^2 (k+1)^2 (\rho^2)^k \quad \text{and} \quad |a_k a_{k-1}| \leq M_0^2 k(k+1)\rho^{2k+1}.$$

Since both bounding series converge, both $\sum a_k^2$ and $\sum a_k a_{k-1}$ converge.

From the recurrence relation, we have

$$a_{k+1}^2 = 4\beta z a_k^2 + \beta^2 a_{k-1}^2 - 4\beta^{3/2} z a_k a_{k-1}$$
$$a_{k+1} a_k = 2\sqrt{\beta} z a_k^2 - \beta a_k a_{k-1}.$$

Summing the first identity over $k = 1, 2, \ldots, T$ gives

$$\sum_{k=1}^{T} a_{k+1}^2 = 4\beta z^2 \sum_{k=1}^{T} a_k^2 + \beta^2 \sum_{k=1}^{T} a_{k-1}^2 - 4\beta^{3/2} z \sum_{k=1}^{T} a_k a_{k-1}.$$

Letting $T \to \infty$, we obtain

$$S - a_0^2 - a_1^2 = 4\beta z^2 (S - a_0^2) + \beta^2 S - 4\beta^{3/2} z U,$$

because $|S|,\ |U| < \infty$. Hence

$$(1 - \beta^2 - 4\beta z^2) S = (1 - 4\beta z^2) a_0^2 + a_1^2 - 4\beta^{3/2} z U.$$

This is the first linear relation between S and U.

Similarly, summing the second identity over $k = 1, 2, \ldots, T$ gives

$$\sum_{k=1}^{T} a_{k+1} a_k = 2\sqrt{\beta} z \sum_{k=1}^{T} a_k^2 - \beta \sum_{k=1}^{T} a_k a_{k-1}.$$

Letting $T \to \infty$ we obtain

$$U = 2\sqrt{\beta} z (S - a_0^2) - \beta U,$$

hence

$$(1 + \beta) U = 2\sqrt{\beta} z (S - a_0^2).$$

Solving the two linear equations for S and U yields

$$S = \frac{(1 + \beta - 4\beta(1 - \beta) z^2) a_0^2 + (1 + \beta) a_1^2 - 4\beta^{3/2} z\, a_0 a_1}{(1 - \beta)\left((1 + \beta)^2 - 4\beta z^2\right)}$$
$$U = \frac{2\sqrt{\beta} z \left((1 + \beta) a_0^2 + a_1^2 - 2\sqrt{\beta} z a_0 a_1\right)}{(1 - \beta)\left((1 + \beta)^2 - 4\beta z^2\right)}.$$

■

Closed forms of infinite sums. Finally, we find the closed forms of Y_∞, D_∞ and E_∞. Since

$$Y_\infty = \sum_{k=-1}^{\infty} \tilde{y}_k^2 - \tilde{y}_{-1}^2, \quad \tilde{y}_{-1} = \tilde{y}_0 = 1,$$

we have

$$Y_\infty = \frac{2(1 + \beta) - 4\beta z^2 (1 - \beta) - 4\beta^{3/2} z}{(1 - \beta)\left((1 + \beta)^2 - 4\beta z^2\right)} - 1$$
$$= \frac{(1 + \beta)(1 + \beta^2) - 4\beta\sqrt{\beta} z}{(1 - \beta)((1 + \beta)^2 - 4\beta z^2)}.$$

Similarly, since

$$D_\infty = \sum_{k=0}^{\infty} d_k^2, \quad d_0 = 0, \quad d_1 = 1,$$

we have

$$D_\infty = \frac{1+\beta}{(1-\beta)\left((1+\beta)^2 - 4\beta z^2\right)}.$$

We now evaluate $E_\infty = \sum_{k=0}^{\infty} k d_k^2$. We begin with the identity

$$E_\infty = \sum_{k=0}^{\infty} \sum_{i=k}^{\infty} d_i^2 - D_\infty.$$

The inner sum $S_k := \sum_{i=k}^{\infty} d_i^2$ is an application of the lemma to a sequence satisfying the same recurrence but with initial conditions d_k and d_{k+1}. Let $C = (1-\beta)\left((1+\beta)^2 - 4\beta z^2\right)$. Applying the formula for S from the lemma gives

$$S_k = \frac{1}{C}\left((1+\beta-4\beta(1-\beta)z^2)d_k^2 + (1+\beta)d_{k+1}^2 - 4\beta^{3/2} z\, d_k d_{k+1}\right).$$

Summing S_k from $k=0$ to ∞, we use $d_0 = 0$ and the definitions of $D_\infty = \sum_{k=0}^{\infty} d_k^2$ and $U = \sum_{k=0}^{\infty} d_k d_{k+1}$:

$$\begin{aligned}
\sum_{k=0}^{\infty} S_k &= \frac{1}{C}\left((1+\beta-4\beta z^2(1-\beta))\sum_{k=0}^{\infty} d_k^2 + (1+\beta)\sum_{k=0}^{\infty} d_{k+1}^2 - 4\beta^{3/2} z \sum_{k=0}^{\infty} d_k d_{k+1}\right) \\
&= \frac{1}{C}\left((1+\beta-4\beta z^2(1-\beta))D_\infty + (1+\beta)(D_\infty - d_0^2) - 4\beta^{3/2} z U\right) \\
&= \frac{1}{C}\left((2(1+\beta) - 4\beta z^2(1-\beta))D_\infty - 4\beta^{3/2} z U\right).
\end{aligned}$$

Substituting this back into the expression for E_∞, we get

$$\begin{aligned}
E_\infty &= \frac{(2(1+\beta) - 4\beta z^2(1-\beta))D_\infty - 4\beta^{3/2} z U}{C} - D_\infty \\
&= \frac{(2(1+\beta) - 4\beta z^2(1-\beta) - C)D_\infty - 4\beta^{3/2} z U}{C}.
\end{aligned}$$

The coefficient of D_∞ simplifies to

$$2(1+\beta) - 4\beta z^2(1-\beta) - (1-\beta)((1+\beta)^2 - 4\beta z^2) = (1+\beta)(1+\beta^2).$$

Therefore,

$$E_\infty = \frac{(1+\beta)(1+\beta^2)D_\infty - 4\beta^{3/2} z U}{C}.$$

Using the specific results for the sequence d_k where $d_0 = 0, d_1 = 1$:

$$D_\infty = \frac{1+\beta}{C}, \quad U = \frac{2\sqrt{\beta} z}{C},$$

we obtain the final expression for E_∞:

$$\begin{aligned}
E_\infty &= \frac{1}{C}\left((1+\beta)(1+\beta^2)\frac{1+\beta}{C} - 4\beta^{3/2} z \frac{2\sqrt{\beta} z}{C}\right) \\
&= \frac{(1+\beta)^2(1+\beta^2) - 8\beta^2 z^2}{(1-\beta)^2\left((1+\beta)^2 - 4\beta z^2\right)^2}.
\end{aligned}$$

To summarize, the limits are given as follows:

$$Y_\infty = \frac{(1+\beta)(1+\beta^2) - 4\beta\sqrt{\beta}z}{(1-\beta)((1+\beta)^2 - 4\beta z^2)} \tag{11a}$$

$$D_\infty = \frac{1+\beta}{(1-\beta)((1+\beta)^2 - 4\beta z^2)} \tag{11b}$$

$$E_\infty = \frac{(1+\beta)^2(1+\beta^2) - 8\beta^2 z^2}{(1-\beta)^2\left((1+\beta)^2 - 4\beta z^2\right)^2}. \tag{11c}$$

Using the relation $(\eta L)^2 = 4\beta(B-z)^2$, we also have

$$(\eta L)^2 M D_\infty = M\,\frac{(1+\beta)(B-z)}{(1-\beta)(B+z)} \tag{12a}$$

$$(\eta L)^2 M E_\infty = M\,\frac{(1+\beta)^2(1+\beta^2) - 8\beta^2 z^2}{4\beta(1-\beta)^2(B+z)^2}. \tag{12b}$$

Since

$$\begin{aligned}\lim_{T\to\infty}\sum_{k=0}^{T-1}\left(1-\frac{k}{T}\right)d_k^2 &= \lim_{T\to\infty}\sum_{k=0}^{T-1} d_k^2 - \frac{1}{T}\sum_{k=0}^{T-1} k d_k^2 \\ &= D_\infty - \lim_{T\to\infty}\frac{E_\infty}{T} \\ &= D_\infty\end{aligned}$$

and

$$\begin{aligned}D_\infty - \sum_{k=0}^{T-1}\left(1-\frac{k}{T}\right)d_k^2 &= \sum_{k=0}^{\infty} d_k^2 - \sum_{k=0}^{T-1} d_k^2 + \frac{1}{T}\sum_{k=0}^{T-1} k d_k^2 \\ &\le \sum_{k=T}^{\infty} d_k^2 + \frac{1}{T}\sum_{k=0}^{\infty} k d_k^2 \\ &= \sum_{k=T}^{\infty} d_k^2 + \frac{E_\infty}{T},\end{aligned}$$

we have

$$D_T - \frac{E_T}{T} \ge D_\infty - \left(\sum_{k=T}^{\infty} d_k^2 + \frac{E_\infty}{T}\right).$$

Thus, we have

$$\frac{Y_T}{T} + (\eta L)^2 M\left(D_T - \frac{E_T}{T}\right) \tag{13}$$

$$\ge \frac{Y_\infty - \sum_{k=T}^{\infty}\tilde{y}_k^2}{T} + (\eta L)^2 M\left(D_\infty - \sum_{k=T}^{\infty} d_k^2 - \frac{E_\infty}{T}\right)$$

$$= \frac{Y_\infty}{T} + (\eta L)^2 M\left(D_\infty - \frac{E_\infty}{T}\right) - \left(\frac{1}{T}\sum_{k=T}^{\infty}\tilde{y}_k^2 + (\eta L)^2 M\sum_{k=T}^{\infty} d_k^2\right). \tag{14}$$

Therefore, by finding appropriate upper bounds for the tail sums, we can obtain a lower bound for (10).

E.4. Chebyshev Polynomials

In this subsection, we introduce Chebyshev polynomials and show how they can be used to express the sequences $\{\tilde{y}_k\}$ and $\{d_k\}$.

The Chebyshev polynomials of the second kind are defined recursively as

$$U_0(x) = 1, \quad U_1(x) = 2x, \quad U_k(x) = 2xU_{k-1}(x) - U_{k-2}(x) \quad \text{for } k \geq 2.$$

$\{U_n\}$ also satisfies the following identities:

$$\begin{aligned}
U_n(\cos\theta) &= \frac{\sin((n+1)\theta)}{\sin\theta}, \quad \text{if } \sin\theta \neq 0 \\
U_n(\cosh\theta) &= \frac{\sinh((n+1)\theta)}{\sinh\theta}, \quad \text{if } \sinh\theta \neq 0 \\
U_n(1) &= n+1 \\
U_n(-x) &= (-1)^n U_n(x) \\
U_n(x) &= \frac{(x+\sqrt{x^2-1})^{n+1} - (x-\sqrt{x^2-1})^{n+1}}{2\sqrt{x^2-1}}, \quad \text{if } x \in \mathbb{R}.
\end{aligned}$$

Now, we can express $\tilde{y}_k$ and d_k in terms of U_k. Let

$$z = \frac{1 - \eta L + \beta}{2\sqrt{\beta}}.$$

Then,

$$\begin{aligned}
\tilde{y}_k &= \sqrt{\beta}^k \left(U_k(z) - \sqrt{\beta} U_{k-1}(z) \right) \\
d_k &= \sqrt{\beta}^{k-1} U_{k-1}(z).
\end{aligned}$$

Here, we define $U_{-1} \equiv 0$.

Proof For convenience, we define the set of sequences which follow the recurrence relation:

$$\mathcal{R} := \{\{c_k\} : c_{k+1} = (1 - \eta L + \beta)c_k - \beta c_{k-1} \quad \forall k\}.$$

Then, we have the following facts:

- If $\{c_k\} \in \mathcal{R}$ and $\{c'_k\} \in \mathcal{R}$, then $\{\alpha c_k + \alpha' c'_k\} \in \mathcal{R}$ for any $\alpha, \alpha' \in \mathbb{R}$.
- If $\{c_k\} \in \mathcal{R}$ and $\{c'_k\} \in \mathcal{R}$ and $c_0 = c'_0$ and $c_1 = c'_1$, then $\{c_k\} \equiv \{c'_k\}$.

Now, we show that $\{\sqrt{\beta}^k U_k(z)\} \in \mathcal{R}$. Since $U_{k+1} = 2zU_k - U_{k-1}$, by multiplying $\sqrt{\beta}^{k+1}$ both sides, we have

$$\sqrt{\beta}^{k+1} U_{k+1}(z) = 2\sqrt{\beta} z \sqrt{\beta}^k U_k(z) - \beta \sqrt{\beta}^{k-1} U_{k-1}(z).$$

Thus, both $\{\sqrt{\beta}^k (U_k(z) - \sqrt{\beta} U_{k-1}(z))\}$ and $\{\sqrt{\beta}^{k-1} U_{k-1}(z)\}$ are in $\mathcal{R}$. It remains to verify that the initial conditions coincide.

- $\tilde{y}_k$: From the relation $U_{k+1} = 2zU_k - U_{k-1}$, we set $U_{-2} = -1$. Then, we have $\sqrt{\beta}^{-1}(U_{-1}(z) - \sqrt{\beta}U_{-2}(z)) = 1$ and $U_0(z) - \sqrt{\beta}U_{-1}(z) = 1$, which exactly coincide $\tilde{y}_{-1} = \tilde{y}_0 = 1$.
- d_k: Since $U_{-1} \equiv 0$, we have $\sqrt{\beta}^{-1}U_{-1}(z) = 0$ and $U_0(z) = 1$. These coincide $d_0 = 0$ and $d_1 = 1$.

■

Let $\lambda_{1,2}$ be the roots of characteristic polynomial of $\tilde{y}_k$ (and d_k) and $B = \frac{1+\beta}{2\sqrt{\beta}} \geq 1$. Then,

- $|z| < 1$ if and only if λ_1 and λ_2 are imaginary.
- $z = \pm 1$ if and only if $\lambda_1 = \lambda_2 = \pm\sqrt{\beta}$.
- $0 < \eta L < 2(1+\beta)$ if and only if $|z| < B$.

Thus, the convergence condition on the step size can be equivalently expressed as $|z| < B$.

E.5. Region-wise Lower Bound Analysis

We aim to characterize the parameter regimes where SHB becomes slower than SGD. To this end, we derive region-wise lower bounds on the SHB and compare them with the SGD upper bound.

We divide the step size domain into three regions, where $z = \frac{1-\eta L+\beta}{2\sqrt{\beta}}$:

$$\text{Region I: } z \in [1, B), \qquad \text{Region II: } z \in [-1, 1], \qquad \text{Region III: } z \in (-B, -1].$$

In terms of η, those can be written as

$$\text{Region I: } \eta \in \left[0, \frac{(1-\sqrt{\beta})^2}{L}\right),$$
$$\text{Region II: } \eta \in \left[\frac{(1-\sqrt{\beta})^2}{L}, \frac{(1+\sqrt{\beta})^2}{L}\right],$$
$$\text{Region III: } \eta \in \left(\frac{(1+\sqrt{\beta})^2}{L}, 2(1+\beta)\right].$$

Note that this partition is determined by whether the characteristic polynomial of the SHB recursion has complex or real roots.

We now outline the derivation of the region-wise lower bounds.

Proof Outline.

- In Region I, we obtain a different tail sum upper bound and show that when the step size is sufficiently small, SHB again becomes slower than SGD. Combining these results yields a region-wise lower bound on SHB, which we then compare with the SGD upper bound to characterize when SHB is slower in the asymptotic regime.
- In Region II, we derive upper bounds on the tail sums, which lead to lower bounds on the SHB convergence rate.

- In Region III, we show that SHB is always slower than SGD when $T \geq 1 + \frac{1}{\sqrt{M}}$ for $M \geq 1$ (or $M = 0$), and when $\beta \geq \frac{M}{4}$ for $0 < M < 1$.

In this subsection, we assume

$$T \geq \max\left\{25, \frac{1}{1-\beta}\left(5\log\frac{1}{1-\beta} + 8\right), 4M + 13\right\}$$

and

$$\beta \geq \frac{M}{4} \quad \text{if } 0 < M < 1.$$

We obtain a lower bound by deriving upper bounds on the tail sums.

E.5.1. REGION I

In this region, we have

$$\sum_{k=T}^{\infty} \tilde{y}_k^2 \leq \frac{\exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(T(1-\beta)-\beta)(z-B)\right)}{1-\exp\left(\frac{4\sqrt{\beta}}{1-\beta}(z-B)\right)}$$

$$\sum_{k=T}^{\infty} d_k^2 \leq \frac{1}{(1-\beta)^2}\frac{\exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(T(1-\beta)-(1+\beta))(z-B)\right)}{1-\exp\left(\frac{4\sqrt{\beta}}{1-\beta}(z-B)\right)}.$$

Proof The bounds are obtained by deriving exponential upper bounds for the individual terms $\tilde{y}_k^2(z)$ and $d_k^2(z)$ for $k \geq T$, and subsequently summing the resulting geometric series.

This derivation relies on the shared structural property of the components: both $\tilde{y}_k^2(z)$ and $d_k^2(z)$ exhibit *log-concavity* on the interval $[1, B)$. This property, combined with the derivatives at the boundary $z = B$, directly yields the exponential decay. We first focus on bounding $\tilde{y}_k^2(z)$.

Log-concavity of $\tilde{y}_k^2(z)$ on $[1, B)$. We first prove the following lemma:

Lemma 13 *If a polynomial $p(x)$ has only real roots, then the function $\log|p(x)|$ is concave on the intervals where $p(x) \neq 0$.*

Proof A function $f(x)$ is concave if its second derivative is non-positive, *i.e.*, $f''(x) \leq 0$. If $p(x)$ has only real roots, we can write it as $p(x) = c\prod_{i=1}^{n}(x - r_i)$ for real roots r_i. Then, $\log|p(x)| = \log|c| + \sum_{i=1}^{n}\log|x - r_i|$.

Since the sum of concave functions is also concave, we only need to show that each term $\log|x - r_i|$ is concave.

The first derivative is $\frac{d}{dx}\log|x - r_i| = \frac{1}{x-r_i}$, and the second derivative is $\frac{d^2}{dx^2}\log|x - r_i| = -\frac{1}{(x-r_i)^2} \leq 0$. Thus, each term is concave, which implies that $\log|p(x)|$ is also concave. ■

Let $H_k(z) := U_k(z) - \sqrt{\beta}U_{k-1}(z)$. Since $\tilde{y}_k^2 = \beta^k H_k(z)^2$, $\log\tilde{y}_k^2 = k\log\beta + 2\log|H_k(z)|$. Thus, it suffices to show that $\log|H_k(z)|$ is concave. To this end, we prove the following:

1. H_k is a polynomial.
2. H_k has only real roots.

3. $H_k \neq 0$ on $[1, B)$.

Since both U_k and U_{k-1} are polynomials, so is H_k. Also, by **Theorem 3.3.2** in Szegő [33], the roots of U_k lie strictly between any two consecutive roots of U_{k-1}. Let $z_1, \ldots, z_{k-1}$ and $w_1, \ldots, w_k$ be the roots of U_{k-1} and U_k, respectively. Also, let $z_0 = -\infty$ and $z_k = +\infty$. Then, $w_1 < z_1 < \cdots < z_{k-1} < w_k$. Since $H_k(w_i) = -\sqrt{\beta} U_{k-1}(w_i)$, and U_{k-1} does not have multiple roots, the signs of U_{k-1} in (z_i, z_{i+1}) and (z_{i+1}, z_{i+2}) are always different. By the Intermediate Value Theorem, H_k has at least $k-1$ real roots. Additionally, since H_k is of degree k, it has at most k roots. As we have already identified $k-1$ real roots, and the coefficients of H_k are all real, the remaining root must also be real. Therefore, all roots of H_k are real. By Lemma 13, $\log |H_k(z)|$ is concave. Finally, if $z \in [1, B)$, then there exists $t \geq 0$ such that $\cosh(t) = z$. Then,

$$H_k(z) = \frac{\sinh((k+1)t) - \sqrt{\beta}\sinh(kt)}{\sinh t}.$$

Since

$$\frac{\sinh((k+1)t)}{\sinh(kt)} = e^t \cdot \frac{1 - e^{-2(k+1)t}}{1 - e^{-2kt}} > 1$$

and $\sqrt{\beta} < 1$, we have $H_k(z) > 0$ on $[1, B)$.

Derivative of $\tilde{y}_k^2(z)$ at $z = B$. Using the derivative formula

$$U_k'(z) = \frac{(k+1)T_{k+1} - zU_k}{z^2 - 1}, \tag{15}$$

where T_k denotes the Chebyshev polynomial of the first kind, we have

$$\tilde{y}_k'(z) = \frac{1}{z^2-1}\left[\sqrt{\beta}^k\left((k+1)T_{k+1}(z) - k\sqrt{\beta}T_k(z)\right) - z\tilde{y}_k(z)\right].$$

To evaluate this at $z = B$, we use the following identities: $B = \cosh\left(\log\frac{1}{\sqrt{\beta}}\right)$, $T_n(\cosh x) = \cosh(nx)$ and $U_n(\cosh x) = \frac{\sinh((n+1)x)}{\sinh x}$. Then, we have

$$\begin{aligned}
T_n(B) &= \frac{\sqrt{\beta}^n + \sqrt{\beta}^{-n}}{2} \\
U_n(B) &= \frac{\sqrt{\beta}^{n+1} - \sqrt{\beta}^{-(n+1)}}{\sqrt{\beta} - \sqrt{\beta}^{-1}} \\
\tilde{y}_n(B) &= 1 \\
B^2 - 1 &= \frac{(1-\beta)^2}{4\beta}.
\end{aligned}$$

Then,

$$\begin{aligned}
(B^2-1)\tilde{y}_k'(B) &= \beta^{k/2}\left((k+1)T_{k+1}(B) - k\sqrt{\beta}T_k(B)\right) - B\tilde{y}_k(B) \\
&= \frac{\beta^{k/2}}{2}\left((k+1)\left(\beta^{-\frac{k+1}{2}} + \beta^{\frac{k+1}{2}}\right) - k\beta^{\frac{1}{2}}\left(\beta^{-\frac{k}{2}} + \beta^{\frac{k}{2}}\right)\right) - \frac{1+\beta}{2\sqrt{\beta}} \\
&= \frac{1}{2}\left((k+1)\left(\beta^{-\frac{1}{2}} + \beta^{k+\frac{1}{2}}\right) - k\left(\beta^{\frac{1}{2}} + \beta^{k+\frac{1}{2}}\right)\right) - \frac{1+\beta}{2\sqrt{\beta}} \\
&= \frac{1}{2}\left(\frac{k+1}{\sqrt{\beta}} - k\sqrt{\beta} + \beta^{k+\frac{1}{2}}\right) - \frac{1+\beta}{2\sqrt{\beta}} \\
&= \frac{1}{2\sqrt{\beta}}\left((k+1-k\beta+\beta^{k+1}) - (1+\beta)\right) \\
&= \frac{1}{2\sqrt{\beta}}\left(k(1-\beta) - \beta(1-\beta^k)\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
\tilde{y}_k'(B) &= \frac{4\beta}{(1-\beta)^2}\frac{1}{2\sqrt{\beta}}\left(k(1-\beta) - \beta(1-\beta^k)\right) \\
&= \frac{2\sqrt{\beta}}{(1-\beta)^2}\left(k(1-\beta) - \beta(1-\beta^k)\right).
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}z}\log\tilde{y}_k^2(z)\Big|_{z=B} &= \frac{\frac{\mathrm{d}}{\mathrm{d}z}\tilde{y}_k^2(z)\big|_{z=B}}{\tilde{y}_k^2(B)} \\
&= \frac{4\sqrt{\beta}}{(1-\beta)^2}\left(k(1-\beta) - \beta(1-\beta^k)\right) \\
&\geq \frac{4\sqrt{\beta}}{(1-\beta)^2}\left(k(1-\beta) - \beta\right).
\end{aligned}$$

Combining log-concavity of $\tilde{y}_k^2$ on $[1, B)$ and its derivative at $z = B$, we have

$$\tilde{y}_k^2 \leq \exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(k(1-\beta)-\beta)(z-B)\right).$$

The series over $k = T$ to ∞ on the right-hand side converges when $z < B$. Since $z \in [1, B)$, the condition holds, and therefore

$$\sum_{k=T}^{\infty}\tilde{y}_k^2 \leq \frac{\exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(T(1-\beta)-\beta)(z-B)\right)}{1-\exp\left(\frac{4\sqrt{\beta}}{1-\beta}(z-B)\right)}.$$

The bound for $\sum_{k=T}^{\infty} d_k^2$ follows from an analogous argument applied to $d_k^2(z)$. First, $d_k^2(z)$ also shares the property of log-concavity on $[1, B)$. Second, we evaluate the derivative of $\log d_k^2(z)$ at the boundary $z = B$.

Log-concavity of $d_k^2(z)$ on $[1, B)$. Since $d_k^2 = \beta^{k-1} U_{k-1}^2(z)$, we have $\log d_k^2 = (k-1)\log\beta + 2\log|U_{k-1}(z)|$. Thus, it suffices to show that $\log|U_{k-1}(z)|$ is concave. Since $U_{k-1}(z)$ has $k-1$ distinct real roots and $U_{k-1} > 0$ on $[1, B)$, by Lemma 13, $\log|U_{k-1}(z)|$ is concave.

Derivative of $d_k^2(z)$ at $z = B$. By (15), we have

$$d_k'(z) = \frac{1}{z^2-1}\left(k\sqrt{\beta}^{k-1} T_k(z) - z d_k(z)\right)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}z}\log d_k^2(z)\Big|_{z=B} = \frac{2d_k'(B)}{d_k(B)}.$$

Since $d_k(B) = \frac{1-\beta^k}{1-\beta}$,

$$\begin{aligned}
(B^2-1)\frac{d_k'(B)}{d_k(B)} &= k\beta^{\frac{k-1}{2}}\frac{T_k(B)}{d_k(B)} - B \\
&= k\beta^{\frac{k-1}{2}}\frac{\sqrt{\beta}^k + \sqrt{\beta}^{-k}}{2}\frac{1-\beta}{1-\beta^k} - \frac{1+\beta}{2\sqrt{\beta}} \\
&= \frac{k(1+\beta^k)}{2\sqrt{\beta}}\frac{1-\beta}{1-\beta^k} - \frac{1+\beta}{2\sqrt{\beta}} \\
&= \frac{1}{2\sqrt{\beta}}\left(k\frac{(1+\beta^k)(1-\beta)}{1-\beta^k} - (1+\beta)\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}z}\log d_k^2(z)\Big|_{z=B} &= \frac{2}{B^2-1}\frac{1}{2\sqrt{\beta}}\left(k\frac{(1+\beta^k)(1-\beta)}{1-\beta^k} - (1+\beta)\right) \\
&= \frac{4\sqrt{\beta}}{(1-\beta)^2}\left(k\frac{(1+\beta^k)(1-\beta)}{1-\beta^k} - (1+\beta)\right) \\
&\geq \frac{4\sqrt{\beta}}{1-\beta}\left(k - \frac{1+\beta}{1-\beta}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
d_k^2 &\leq \left(\frac{1-\beta^k}{1-\beta}\right)^2 \exp\left(\frac{4\sqrt{\beta}}{1-\beta}\left(k - \frac{1+\beta}{1-\beta}\right)(z-B)\right) \\
&\leq \frac{1}{(1-\beta)^2}\exp\left(\frac{4\sqrt{\beta}}{1-\beta}\left(k - \frac{1+\beta}{1-\beta}\right)(z-B)\right).
\end{aligned}$$

The series over $k = T$ to ∞ on the right-hand side converges whenever $z < B$. Since $z \in [1, B)$, this condition is satisfied, and hence

$$\sum_{k=T}^{\infty} d_k^2 \leq \frac{1}{(1-\beta)^2}\frac{\exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(T(1-\beta) - (1+\beta))(z-B)\right)}{1 - \exp\left(\frac{4\sqrt{\beta}}{1-\beta}(z-B)\right)}.$$

■

Define

$$R_{Y,T}^{(1)} = \frac{\exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(T(1-\beta)-\beta)(z-B)\right)}{1-\exp\left(\frac{4\sqrt{\beta}}{1-\beta}(z-B)\right)}$$
$$R_{D,T}^{(1)} = \frac{1}{(1-\beta)^2}\frac{\exp\left(\frac{4\sqrt{\beta}}{(1-\beta)^2}(T(1-\beta)-(1+\beta))(z-B)\right)}{1-\exp\left(\frac{4\sqrt{\beta}}{1-\beta}(z-B)\right)}.$$

Then, we can express the lower bound as

$$\Lambda_1 = \frac{Y_\infty - R_{Y,T}^{(1)}}{T} + (\eta L)^2 M\left(D_\infty - R_{D,T}^{(1)} - \frac{E_\infty}{T}\right). \tag{16}$$

E.5.2. REGION II

Next, we provide upper bounds for $\sum_{k=T}^{\infty}\tilde{y}_k^2$ and $\sum_{k=T}^{\infty} d_k^2$, on $z \in [-1,1]$. When $z \in [-1,1]$, we have

$$\sum_{k=T}^{\infty}\tilde{y}_k^2 \le \frac{41}{32}\left(1+\sqrt{\beta}\right)^2\frac{(T+1)^2\beta^T}{1-\beta}$$
$$\sum_{k=T}^{\infty} d_k^2 \le \frac{41}{32}\frac{T^2\beta^{T-1}}{1-\beta}.$$

Proof Since $\tilde{y}_k = \sqrt{\beta}^k\left(U_k(z) - \sqrt{\beta}U_{k-1}(z)\right)$, the difference between the limit and the partial sum is

$$\begin{aligned}\sum_{k=T}^{\infty}\tilde{y}_k^2 &= \sum_{k=T}^{\infty}\left(U_k(z) - \sqrt{\beta}U_{k-1}(z)\right)^2\beta^k \\ &\le \sum_{k=T}^{\infty}\left((1+\sqrt{\beta})k+1\right)^2\beta^k.\end{aligned}$$

The second inequality holds because $U_k(z)$ is bounded by $|U_k(z)| \le k+1$ for all $z \in [-1,1]$ with equality at $z=-1$. By applying the triangle inequality to the terms in the sum, we obtain the upper bound, which is maximized when $z=-1$.

Similarly,

$$\begin{aligned}\sum_{k=T}^{\infty} d_k^2 &= \sum_{k=T}^{\infty}\beta^{k-1}U_{k-1}^2(z) \\ &\le \sum_{k=T}^{\infty}\beta^{k-1}k^2 \\ &= \frac{\beta^{T-1}\left((1-\beta)^2T^2 + 2\beta(1-\beta)T + \beta(1+\beta)\right)}{(1-\beta)^3}.\end{aligned}$$

Our first goal is to show that

$$\frac{\beta^{T-1}\left((1-\beta)^2T^2+2\beta(1-\beta)T+\beta(1+\beta)\right)}{(1-\beta)^3}\leq\frac{41}{32}\frac{T^2\beta^{T-1}}{1-\beta}$$

holds for $T\geq\frac{8}{1-\beta}$. By dividing both sides by $\frac{\beta^{T-1}T^2}{1-\beta}$, it becomes

$$1+\frac{2\beta}{1-\beta}\frac{1}{T}+\frac{\beta(1+\beta)}{(1-\beta)^2}\frac{1}{T^2}\leq\frac{41}{32}.$$

Since the left-hand side is decreasing in T, it suffices to show that the inequality holds when $T=\frac{8}{1-\beta}$. For $T=\frac{8}{1-\beta}$, the left-hand side is $1+\frac{2\beta}{8}+\frac{\beta(1+\beta)}{64}$, and we have

$$\sup_{0\leq\beta<1}1+\frac{2\beta}{8}+\frac{\beta(1+\beta)}{64}=\frac{41}{32}.$$

Therefore,

$$1+\frac{2\beta}{1-\beta}\frac{1}{T}+\frac{\beta(1+\beta)}{(1-\beta)^2}\frac{1}{T^2}\leq\frac{41}{32}$$

holds for all $T\geq\frac{8}{1-\beta}$.

From the previous result, we have

$$\begin{aligned}\sum_{k=T}^{\infty}\tilde{y}_k^2&=\sum_{k=T}^{\infty}\left(U_k\left(z\right)-\sqrt{\beta}U_{k-1}\left(z\right)\right)^2\beta^k\\&\leq\sum_{k=T}^{\infty}\left((1+\sqrt{\beta})k+1\right)^2\beta^k\\&\leq\left(1+\sqrt{\beta}\right)^2\sum_{n=T+1}^{\infty}n^2\beta^{n-1}\\&\leq\frac{41}{32}\left(1+\sqrt{\beta}\right)^2\frac{(T+1)^2\beta^T}{1-\beta}\end{aligned}$$

if $T+1\geq\frac{8}{1-\beta}$. ■

Define

$$\begin{aligned}R_{Y,T}^{(2)}&=\frac{41}{32}\left(1+\sqrt{\beta}\right)^2\frac{(T+1)^2\beta^T}{1-\beta}\\R_{D,T}^{(2)}&=\frac{41}{32}\frac{T^2\beta^{T-1}}{1-\beta}.\end{aligned}$$

Then, we can express the lower bound as

$$\Lambda_2=\frac{Y_\infty-R_{Y,T}^{(2)}}{T}+(\eta L)^2M\left(D_\infty-R_{D,T}^{(2)}-\frac{E_\infty}{T}\right).\tag{17}$$

E.5.3. REGION III

In this region, we will show that (10) does not attain a minimum.

Before proceeding, we simplify the expressions of Y_∞, D_∞, and E_∞ by introducing the variable

$$u = \frac{B-z}{B+z} = \frac{\eta L}{2(1+\beta)-\eta L}.$$

In terms of u, the three regions can be equivalently written as

$$\text{Region I: } u \in \left(0, \frac{B-1}{B+1}\right], \text{ Region II: } u \in \left[\frac{B-1}{B+1}, \frac{B+1}{B-1}\right], \text{ Region III: } u \in \left[\frac{B+1}{B-1}, \infty\right).$$

We note that

$$\frac{B-1}{B+1} = \left(\frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}\right)^2.$$

Under this substitution, we have

$$\begin{aligned}
Y_\infty(u) &= \frac{1}{4(1-\beta^2)}\left(\frac{(1-\beta)^2}{u} + (1+\beta)^2\right)(1+u) \\
&= \frac{1}{4}\left(\left(\frac{1-\beta}{1+\beta}\right)\frac{1}{u} + \frac{2(1+\beta^2)}{1-\beta^2} + \left(\frac{1+\beta}{1-\beta}\right)u\right) \\
(\eta L)^2 M D_\infty(u) &= M\,\frac{(1+\beta)}{(1-\beta)}u \\
(\eta L)^2 M E_\infty(u) &= \frac{M}{4}\left(1 + 2\left(\frac{1+\beta}{1-\beta}\right)^2 u + u^2\right),
\end{aligned}$$

which are all expressed as rational functions of u.

We have

$$\frac{Y_\infty}{T} + (\eta L)^2 M\left(D_\infty - \frac{E_\infty}{T}\right) = \frac{c_{-1}}{u} + c_0 + c_1 u + c_2 u^2,$$

where

$$c_{-1} = \frac{1}{4T}\frac{1-\beta}{1+\beta} \tag{18a}$$

$$c_0 = \frac{1}{2T}\frac{1+\beta^2}{1-\beta^2} - \frac{M}{4T} \tag{18b}$$

$$\begin{aligned}
c_1 &= \frac{1}{4T}\frac{1+\beta}{1-\beta} - \frac{M}{2T}\frac{(1+\beta)^2}{(1-\beta)^2} + \frac{M(1+\beta)}{1-\beta} \\
&= \frac{1}{4T}\frac{1+\beta}{1-\beta}\left(1 - 2M\frac{1+\beta}{1-\beta} + 4MT\right)
\end{aligned} \tag{18c}$$

$$c_2 = -\frac{M}{4T}. \tag{18d}$$

Since

$$\begin{aligned}(10) &= \frac{1}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 + (\eta L)^2 M \sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)|d_j|^2 \\ &\le \frac{1}{T}\sum_{k=0}^{\infty}|\tilde{y}_k|^2 + (\eta L)^2 M \sum_{j=1}^{\infty}|d_j|^2 \\ &= \frac{Y_\infty}{T} + (\eta L)^2 M D_\infty,\end{aligned}$$

we have

$$\frac{Y_\infty}{T} + (\eta L)^2 M D_\infty = \frac{1}{4T}\frac{1-\beta}{1+\beta}\frac{1}{u} + \frac{1}{2T}\frac{1+\beta^2}{1-\beta^2} + \frac{1}{4T}\frac{1+\beta}{1-\beta}(1+4MT)$$

and

$$\min_{u>0}\frac{Y_\infty}{T} + (\eta L)^2 M D_\infty = \frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2} + \sqrt{1+4MT}\right). \tag{19}$$

Therefore, it is sufficient to show that

$$\frac{1}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 + (\eta L)^2 M \sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)|d_j|^2 \ge \frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2} + \sqrt{1+4MT}\right)$$

in Region III.

We first show that both $\tilde{y}_k^2$ and d_k^2 are non-increasing in z on this region. We use the substitution $z = -\cosh x$, which maps the interval $z \in (-B, -1]$ to $x \in [\alpha, 0]$, where $\cosh\alpha = B$ and $\alpha < 0$. In this region, we have the identity $U_{k-1}(z) = (-1)^{k-1}\frac{\sinh(kx)}{\sinh x}$. Since $x \mapsto z = -\cosh x$ is increasing on $(-\infty, 0)$, it is enough to check the sign of derivative with respect to x.

Now, we prove the following lemma:

Lemma 14 *The mapping $x \mapsto \frac{\sinh(kx)}{\sinh x}$ is non-decreasing in x on $(0,\infty)$ for all $k \in \mathbb{N}$.*

Proof Let $\psi_k(x) := \frac{\sinh(kx)}{\sinh x}$. Then,

$$\frac{\psi_k'(x)}{\psi_k(x)} = k\coth(kx) - \coth x.$$

Letting $\phi(t) = t\coth t$, we have

$$k\coth(kx) - \coth x = \frac{\phi(kx)-\phi(x)}{x},$$

and

$$\phi'(t) = \frac{\frac{1}{2}\sinh(2t) - t}{\sinh^2 t} > 0.$$

Thus, ϕ is non-decreasing, and $\phi(kx) \ge \phi(x)$ for $k \ge 1$, which implies ψ_k is also non-decreasing. ■

From the lemma, $x \mapsto \frac{\sinh(kx)}{\sinh x}$ is non-increasing in x on $(-\infty, 0)$, because $x \mapsto \frac{\sinh(kx)}{\sinh x}$ is an even function.

For $z \in (-B, -1)$, or equivalently $x \in (\alpha, 0)$, we have

$$\begin{aligned}\operatorname{sign}\left(U_{k-1}(z)\right) &= \operatorname{sign}\left((-1)^{k-1}\frac{\sinh(kx)}{\sinh x}\right) = (-1)^{k-1}\\ \operatorname{sign}\left(U'_{k-1}(z)\right) &= (-1)^{(k-1)}\operatorname{sign}\left(\frac{\mathrm{d}}{\mathrm{d}x}\frac{\sinh(kx)}{\sinh x}\right) = (-1)^{k}.\end{aligned}$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}z}d_k^2(z) = 2\beta^{k-1}U_{k-1}(z)U'_{k-1}(z),$$

$d_k^2(z)$ is non-increasing in z on $(-B, -1)$, if $k \geq 1$.

We also have

$$\frac{\mathrm{d}}{\mathrm{d}z}\tilde{y}_k^2(z) = 2\beta^k\left(U_k(z) - \sqrt{\beta}U_{k-1}(z)\right)\left(U'_k(z) - \sqrt{\beta}U'_{k-1}(z)\right).$$

Since

$$\begin{aligned}U_k(z) - \sqrt{\beta}U_{k-1}(z) &= \frac{(-1)^k}{\sinh x}\left(\sinh((k+1)x) + \sqrt{\beta}\sinh(kx)\right)\\ \operatorname{sign}\left(U'_k(z)\right) &= (-1)^{k+1}\\ \operatorname{sign}\left(U'_{k-1}(z)\right) &= (-1)^{k},\end{aligned}$$

we have

$$\operatorname{sign}\left(U'_k(z) - \sqrt{\beta}U'_{k-1}(z)\right) = (-1)^{k+1},$$

so

$$\operatorname{sign}\left(\frac{\mathrm{d}}{\mathrm{d}z}\tilde{y}_k^2(z)\right) = -1.$$

Consequently, $\tilde{y}_k^2(z)$ is also non-increasing.

If $k = 0$, both $\tilde{y}_k^2$ and d_k^2 are constant, so they are non-increasing.

Note that

$$\begin{aligned}&\frac{1}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 + (\eta L)^2 M\sum_{j=1}^{T-1}\left(1 - \frac{j}{T}\right)|d_j|^2\\ &\geq \frac{1}{T}\sum_{k=0}^{T-1}\beta^k(k(1+\sqrt{\beta})+1)^2 + \left(1+\sqrt{\beta}\right)^4 M\sum_{j=1}^{T-1}\left(1 - \frac{j}{T}\right)\beta^{j-1}j^2.\end{aligned}$$

Case 1. $0 < M < 1$ In this case, we assume that $\beta \geq \frac{M}{4}$. Then,

$$\begin{aligned}
&\frac{1}{T}(\tilde{y}_0^2 + \tilde{y}_1^2) + (\eta L)^2 M \left(1 - \frac{1}{T}\right) d_1^2 \\
&\geq \frac{1}{T}\left(1 + \beta(\sqrt{\beta}+2)^2\right) + M\left(1+\sqrt{\beta}\right)^4 \left(1 - \frac{1}{T}\right) \\
&\geq \frac{1}{T}(1+4\beta) + M\left(1 - \frac{1}{T}\right) \\
&\geq M + \frac{1}{T}. && (\because 4\beta \geq M)
\end{aligned}$$

We also have

$$\begin{aligned}
\frac{1}{T}\sum_{k=2}^{T-1} \beta^k (k(1+\sqrt{\beta})+1)^2 &\geq \frac{1}{T}\sum_{k=2}^{T-1} \beta^k \cdot 9 \qquad \left(\because k\left(1+\sqrt{\beta}\right)+1 \geq 3 \quad \forall k \geq 2\right) \\
&= \frac{9}{T}\frac{\beta^2(1-\beta^{T-2})}{1-\beta}.
\end{aligned} \tag{20}$$

Meanwhile, we have

$$\begin{aligned}
\frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2} + \sqrt{1+4MT}\right) &\leq \frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2} + 1 + 2MT\right) \quad \left(\because \sqrt{1+x} \leq 1 + \frac{x}{2} \quad \forall x \geq 0\right) \\
&\leq M + \frac{1}{T(1-\beta^2)}.
\end{aligned}$$

Thus, it is sufficient to show that

$$M + \frac{1}{T} + \frac{9}{T}\frac{\beta^2(1-\beta^{T-2})}{1-\beta} \geq M + \frac{1}{T(1-\beta^2)}.$$

Since $\beta \in (0,1)$, it is enough to show that

$$1 - \beta^{T-2} \geq \frac{1}{9},$$

which is equivalent to

$$T \geq 2 + \frac{\log(9/8)}{1-\beta}.$$

Since $\frac{8}{1-\beta} \geq 2 + \frac{\log(9/8)}{1-\beta}$, the inequality is true.

Case 2. $M \geq 1$ Since $T \geq 2$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{k=0}^{T-1} \beta^k (k(1+\sqrt{\beta})+1)^2 + \left(1+\sqrt{\beta}\right)^4 M \sum_{j=1}^{T-1} \left(1 - \frac{j}{T}\right)\beta^{j-1} j^2 \\
&\geq \left(1+\sqrt{\beta}\right)^4 M\left(1 - \frac{1}{T}\right) \\
&\geq \frac{M}{2}.
\end{aligned}$$

We also have

$$\begin{aligned}
&\frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2}+\sqrt{1+4MT}\right)\\
&\leq \frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2}+1+2\sqrt{MT}\right) \quad \left(\because \sqrt{1+x}\leq 1+\sqrt{x} \quad \forall x\geq 0\right)\\
&= \frac{1}{T(1-\beta^2)}+1+2\sqrt{MT}\\
&\leq \frac{1}{8(1+\beta)}+\sqrt{\frac{M}{T}} \qquad \left(\because T\geq \frac{8}{1-\beta}\right)\\
&\leq \frac{1}{8(1+\beta)}+\frac{\sqrt{M}}{5} \qquad (\because T\geq 25)\\
&\leq \frac{1}{8}+\frac{\sqrt{M}}{5}.
\end{aligned}$$

For all $M \geq 1$, we have $\frac{1}{8}+\frac{\sqrt{M}}{5} \leq \frac{M}{2}$, as desired.

Case 3. $M = 0$ Since $\tilde{y}_k^2$ is non-increasing, we have $\tilde{y}_k^2 \geq \beta^k(k(1+\sqrt{\beta})+1)^2$ on Region III. Thus, we have

$$\frac{1}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 \geq \frac{1}{T}\sum_{k=0}^{T-1}\beta^k(k(1+\sqrt{\beta})+1)^2,$$

as desired.

E.6. SHB for Small Step Sizes

Now, we show that SHB is slower than SGD for small step sizes. We define

$$\bar{u} := \frac{1}{6}\frac{1}{\sqrt{4M+1}}\frac{1-\beta}{1+\beta}\frac{1}{\sqrt{T}} \tag{21a}$$

$$\bar{s} := B_0\frac{\bar{u}}{1+\bar{u}} \tag{21b}$$

$$\bar{z} := B\frac{1-\bar{u}}{1+\bar{u}}. \tag{21c}$$

Since $\bar{u} \leq \frac{1}{6}$, we have $\bar{u} \leq 1 \leq \frac{B+1}{B-1}$ for all $T \geq 1$. Hence, $\bar{u}$ lies in either Region I or Region II. We will now show that for $0 < u \leq \bar{u}$, SHB is always slower than SGD. If $\bar{u}$ lies in Region I, we compare the minimum of Λ_1 on $(0, \bar{u}]$ with the SGD upper bound. If $\bar{u}$ lies in Region II, we compare the minimum of Λ_1 on Region I and that of Λ_2 on $\left[\frac{B-1}{B+1}, \bar{u}\right]$ with the SGD upper bound.

We first consider the case where $\bar{u}$ falls into Region I. In this case, (16),

$$\Lambda_1 = \frac{1}{T}\left(\frac{1}{B_0 u}-\frac{\exp\left(-A_0\frac{u}{1+u}\right)}{1-\exp\left(-B_0\frac{u}{1+u}\right)}\right)-(\eta L)^2 M R_{D,T}^{(1)}+c_0+c_1 u+c_2 u^2,$$

where $B_0 = \frac{4(1+\beta)}{1-\beta}$, $A_0 = B_0\left(T-\frac{\beta}{1-\beta}\right)$ and c_0, c_1, c_2 are defined as in (18).

We show that

$$\frac{1}{B_0 u} - \frac{\exp\left(-A_0 \frac{u}{1+u}\right)}{1 - \exp\left(-B_0 \frac{u}{1+u}\right)}$$

is decreasing in u. Let $s = \frac{B_0 u}{1+u} \in (0, B_0)$. Then, $(\eta L)^2 = \frac{(1-\beta)^2}{4} s^2$ and

$$\frac{1}{B_0 u} - \frac{\exp\left(-A_0 \frac{u}{1+u}\right)}{1 - \exp\left(-B_0 \frac{u}{1+u}\right)} = \frac{1}{s} - \frac{\exp\left(-\frac{A_0}{B_0} s\right)}{1 - \exp(-s)} - \frac{1}{B_0}.$$

Since $u = \frac{s}{B_0 - s}$ is increasing in s, it suffices to show that the right-hand side is decreasing in s.

Here, we provide the following lemma:

Lemma 15

$$\frac{1}{x} - \frac{e^{-cx}}{1 - e^{-x}}$$

is decreasing for $x > 0$ *if* $c \geq 1$.

Proof Let $g(x) := \frac{1}{x} - \frac{e^{-cx}}{1-e^{-x}}$ and $\phi(x) := (e^x - 1)^{-1}$. Then,

$$g(x) = \frac{1}{x} - e^{-(c-1)x}\phi(x)$$
$$g'(x) = -\frac{1}{x^2} + e^{-(c-1)x}(c\phi(x) + \phi(x)^2).$$

Thus,

$$g'(x) < 0 \iff e^{-(c-1)x}(c\phi(x) + \phi(x)^2) < \frac{1}{x^2}.$$

Let $G(c, x) := e^{-(c-1)x}(c\phi(x) + \phi(x)^2)$. Then,

$$\frac{\partial}{\partial c} G = e^{-(c-1)x}\phi(x)(1 - cx - x\phi(x)).$$

Since

$$1 - cx - x\phi(x) = 1 - cx - \frac{x}{e^x - 1} \leq 1 - x - \frac{x}{e^x - 1} < 0,$$

we have $\frac{\partial}{\partial c} G < 0$. That is, $G(c, x)$ decreases in c. This implies that it suffices to show that $G(1, x) < \frac{1}{x^2}$. Since

$$G(1, x) = \phi(x) + \phi(x)^2 = \frac{e^x}{(e^x - 1)^2},$$

we need to show that

$$\frac{e^x}{(e^x - 1)^2} < \frac{1}{x^2},$$

which is equivalent to show that

$$x e^{\frac{x}{2}} < e^x - 1.$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(e^x - 1 - xe^{\frac{x}{2}}\right) = \frac{1}{2}e^{\frac{x}{2}}\left(2e^{\frac{x}{2}} - 2 - x\right) \geq 0,$$

we have $xe^{\frac{x}{2}} < e^x - 1$ for all $x > 0$, as desired. ■

As $\frac{A_0}{B_0} = T - \frac{\beta}{1-\beta}$,

$$\frac{1}{B_0 u} - \frac{\exp\left(-A_0 \frac{u}{1+u}\right)}{1 - \exp\left(-B_0 \frac{u}{1+u}\right)}$$

is decreasing in u, because $T - \frac{\beta}{1-\beta} \geq 1$.

We first note that $c_1 \geq 0$. Then, $(\eta L)^2 = \frac{(1-\beta)^2}{4}\bar{s}^2$, and if we assume $0 < u \leq \bar{u}$ (equivalently, $\bar{z} \leq z < B$), we have

$$\Lambda_1 \geq \underbrace{\frac{1}{T}\frac{1}{B_0\bar{u}} - \frac{1}{T}\frac{\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right)}{1 - \exp\left(-\bar{s}\right)}}_{\text{(I)}} \underbrace{- \sup_{0<s\leq\bar{s}} \frac{Ms^2}{4}\frac{\exp\left(-\left(T - \frac{1+\beta}{1-\beta}\right)s\right)}{1 - \exp(-s)}}_{\text{(II)}} + c_0 - \underbrace{|c_2||\bar{u}|^2}_{\text{(III)}}.$$

Term I. We use the fact that $1 - e^{-x} \geq x(1 - x/2)$ for $x \geq 0$. We also note that

$$\bar{u} \leq \frac{1}{6\sqrt{T}}$$
$$\bar{s} \leq B_0\bar{u} = \frac{2}{3}\frac{1}{\sqrt{4M+1}}\frac{1}{\sqrt{T}} \leq \frac{2}{3\sqrt{T}}.$$

Thus,

$$\begin{aligned}
\frac{1}{TB_0\bar{u}} - \frac{1}{T}\frac{\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right)}{1 - \exp\left(-\bar{s}\right)} &\geq \frac{1}{TB_0\bar{u}} - \frac{1}{TB_0\bar{u}}\frac{1+\bar{u}}{1-\bar{s}/2}\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right) \\
&\geq \frac{1}{TB_0\bar{u}}\left(1 - \frac{1+\bar{u}}{1-\bar{s}/2}\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right)\right) \\
&= \frac{3}{2}\frac{\sqrt{4M+1}}{\sqrt{T}}\left(1 - \frac{1+\bar{u}}{1-\bar{s}/2}\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right)\right).
\end{aligned}$$

Since $T \geq 25$, we have $\bar{u} \leq \frac{1}{30}$ and $\bar{s} \leq \frac{2}{15}$. Consequently,

$$\frac{1+\bar{u}}{1-\bar{s}/2} \leq \frac{31}{28}.$$

Also, since $T \geq \frac{8}{1-\beta}$, we have $T - \frac{\beta}{1-\beta} \geq \frac{7}{8}T$. Therefore,

$$\begin{aligned}
1 - \frac{1+\bar{u}}{1-\bar{s}/2} \exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right) &\geq 1 - \frac{31}{28}\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)\bar{s}\right) \\
&\geq 1 - \frac{31}{28}\exp\left(-\frac{7T}{8}\frac{B_0\bar{u}}{1+\bar{u}}\right) \\
&\geq 1 - \frac{31}{28}\exp\left(-\frac{7T}{8}\frac{B_0\bar{u}}{31/30}\right) \qquad \left(\because \bar{u} \leq \frac{1}{30}\right) \\
&= 1 - \frac{31}{28}\exp\left(-\frac{105}{124}\frac{\sqrt{T}}{\sqrt{4M+1}}\right).
\end{aligned}$$

Therefore, we obtain

$$\text{(I)} \geq \frac{3}{2}\frac{\sqrt{4M+1}}{\sqrt{T}}\left(1 - \frac{31}{28}\exp\left(-\frac{105}{124}\frac{\sqrt{T}}{\sqrt{4M+1}}\right)\right). \tag{22}$$

Term II. We obtain

$$\begin{aligned}
\frac{M}{4}\frac{s^2\exp\left(-\left(T - \frac{1+\beta}{1-\beta}\right)s\right)}{1-\exp(-s)} &\leq \frac{M}{4}\cdot 2s\exp\left(-\left(T - \frac{1+\beta}{1-\beta}\right)s\right) \\
&\leq \frac{M}{4}\frac{2}{e\left(T - \frac{1+\beta}{1-\beta}\right)} \\
&\leq \frac{2}{3e}\frac{M}{T}.
\end{aligned} \tag{23}$$

The first line holds because $s \leq \bar{s} \leq \frac{2}{3\sqrt{T}} < 1$ and $\frac{x}{1-e^{-x}} \leq 2$ for all $x \in (0,1)$, and the second line follows from the fact that $x\exp(-Ax) \leq \frac{1}{eA}$ for all $x \geq 0$ and $A > 0$. Finally, the last inequality uses the assumption on T, which ensures $T \geq \frac{8}{1-\beta}$.

Term III. We have

$$|c_2||\bar{u}|^2 = \frac{1}{144T^2}\left(\frac{1-\beta}{1+\beta}\right)^2\frac{M}{4M+1} \leq \frac{1}{144T}, \tag{24}$$

because $T \geq 1$.

Combining (22)–(24), we have

$$\Lambda_1 \geq \frac{3}{2}\frac{\sqrt{4M+1}}{\sqrt{T}}\left(1 - \frac{31}{28}\exp\left(-\frac{105}{124}\frac{\sqrt{T}}{\sqrt{4M+1}}\right)\right) - \left(\frac{2}{3e} + \frac{1}{4}\right)\frac{M}{T} - \frac{1}{144T} + c_0. \tag{25}$$

Let Λ_1' be the right-hand side of (25). Now, we will show that $\Lambda_1' \geq \frac{1}{2T}\left(\frac{1+\beta^2}{1-\beta^2} + \sqrt{1+4MT}\right)$. Using $\sqrt{1+x} \leq 1 + \sqrt{x}$, which holds for all $x \geq 0$, it suffices to show that

$$\Lambda_1' \geq \frac{1}{2T}\frac{1+\beta^2}{1-\beta^2} + \frac{1+2\sqrt{MT}}{2T}. \tag{26}$$

Since

$$\begin{aligned}
&\Lambda_1' - \frac{1}{2T}\frac{1+\beta^2}{1-\beta^2} \\
&= \frac{3}{2}\frac{\sqrt{4M+1}}{\sqrt{T}}\left(1 - \frac{31}{28}\exp\left(-\frac{105}{124}\frac{\sqrt{T}}{\sqrt{4M+1}}\right)\right) - \frac{1}{T}\left(\left(\frac{2}{3e}+\frac{1}{2}\right)M + \frac{1}{144}\right) \\
&\geq \frac{3}{2}\frac{\sqrt{4M+1}}{\sqrt{T}}\left(1 - \frac{31}{28}\exp\left(-\frac{105}{124}\frac{\sqrt{T}}{\sqrt{4M+1}}\right)\right) - \frac{1}{T}\left(\frac{3}{4}M + \frac{1}{144}\right),
\end{aligned}$$

it suffices to show that

$$\frac{3}{2}\frac{\sqrt{4M+1}}{\sqrt{T}}(1-X) - \frac{1}{T}\left(\frac{3}{4}M + \frac{1}{144}\right) \geq \frac{1}{2T} + \sqrt{\frac{M}{T}},$$

where $X := \frac{31}{28}\exp\left(-\frac{105}{124}\frac{\sqrt{T}}{\sqrt{4M+1}}\right)$. Multiplying T and rearranging yields

$$\left(\frac{3}{2}\sqrt{4M+1} - \sqrt{M}\right)\sqrt{T} - \frac{3}{2}X\sqrt{4M+1} \geq \frac{3}{4}M + \frac{73}{144}.$$

Since $\sqrt{M} \leq \frac{1}{2}\sqrt{4M+1}$ and $1 - \frac{3}{2}X \geq 1 - \frac{93}{56}\exp\left(-\frac{105}{124}\right) \geq 0.2878$, we need

$$T \geq \frac{\left(\frac{3}{4}M + \frac{73}{144}\right)^2}{0.2878^2(4M+1)}.$$

Under the assumption, $T \geq 2M+4$, and $0.2878^2(4M+1)(2M+4) \geq \left(\frac{3}{4}M + \frac{73}{144}\right)^2$ for all $M \geq 0$. This concludes the proof of (26).

Next, consider the case where $\bar{u}$ lies in Region II. Since the result for the previous case did not rely on the condition $\bar{u} \leq \frac{B-1}{B+1}$, $\Lambda_1 \geq \frac{1+\sqrt{1+4MT}}{2T}$ for all $u \in \left(0, \frac{B-1}{B+1}\right]$. Then, the only remaining part is Region II, $u \in \left[\frac{B-1}{B+1}, \bar{u}\right]$. In this range, our lower bound is

$$\begin{aligned}
\Lambda_2 = &\frac{1}{T}\frac{1}{B_0 u} + c_0 + c_1 u + c_2 u^2 \\
&- \left(\frac{41}{32}\frac{(1+\sqrt{\beta})^2}{T}\frac{(T+1)^2\beta^T}{1-\beta} + 4M(1+\beta)^2\left(\frac{u}{1+u}\right)^2\frac{41}{32}\frac{T^2\beta^{T-1}}{1-\beta}\right),
\end{aligned}$$

because $2(1+\beta)\frac{u}{1+u} = \eta L$.

We further obtain

$$\begin{aligned}
&\frac{41}{32}\frac{(1+\sqrt{\beta})^2}{T}\frac{(T+1)^2\beta^T}{1-\beta} + 4M(1+\beta)^2\left(\frac{u}{1+u}\right)^2\frac{41}{32}\frac{T^2\beta^{T-1}}{1-\beta} \\
&\leq \frac{41}{32}\frac{T\beta^{T-1}}{1-\beta}\left(\beta(1+\sqrt{\beta})^2\left(1+\frac{1}{T}\right)^2 + 4M(1+\beta)^2 T\bar{u}^2\right) \\
&\leq \frac{41}{32}\frac{T\beta^{T-1}}{1-\beta}\left(\beta(1+\sqrt{\beta})^2\left(\frac{26}{25}\right)^2 + \frac{1}{36}\frac{4M}{4M+1}(1-\beta)^2\right) \qquad (\because T \geq 25) \\
&\leq \frac{41}{32}\cdot 4\cdot\left(\frac{26}{25}\right)^2\frac{T\beta^{T-1}}{1-\beta}.
\end{aligned}$$

Let $C_0 = \frac{41}{32} \cdot 4 \cdot \left(\frac{26}{25}\right)^2 = 5.5432$. Now, we will show that if $T \geq \frac{1}{1-\beta}\left(5\log\frac{1}{1-\beta} + 8\right)$, then $C_0\frac{T\beta^{T-1}}{1-\beta} \leq \frac{1}{2\sqrt{T}}$. To show this, we first prove the following lemma:

Lemma 16 *Let $C > 0$, $\gamma > 0$, and $\delta > 0$. Assume that A_1 and A_0 are positive constants satisfying:*

1. $A_1 > \delta$,
2. $A_0 \geq \frac{\gamma A_1}{A_1 - \delta}$,
3. $A_0 - \gamma\log(A_0) \geq \log(C)$.

For any $\varepsilon \in (0, 1]$, if $y \geq A_1\log(1/\varepsilon) + A_0$, then

$$Cy^\gamma e^{-y}\varepsilon^{-\delta} \leq 1.$$

Proof Let $t = \log(1/\varepsilon)$, so that $t \in [0, \infty)$. Then, $y \geq A_1 t + A_0$ and we need to show the inequality is $Cy^\gamma e^{-y}e^{\delta t} \leq 1$. Since $f(y) = y^\gamma e^{-y}$ is decreasing for $y \geq A_0 > \gamma$, the left-hand side is maximized at the minimum value $y = A_1 t + A_0$.

It suffices to show that $g(t) \leq 1$ for all $t \geq 0$, where

$$g(t) = C(A_1 t + A_0)^\gamma \exp\left(-(A_1 - \delta)t - A_0\right).$$

We differentiate $L(t) := \log(g(t)/C)$:

$$L'(t) = \frac{\gamma A_1}{A_1 t + A_0} - (A_1 - \delta).$$

From Condition 2, $A_0 \geq \frac{\gamma A_1}{A_1 - \delta}$, which means $\frac{\gamma A_1}{A_0} \leq A_1 - \delta$. Thus,

$$L'(t) \leq \frac{\gamma A_1}{A_0} - (A_1 - \delta) \leq 0.$$

Since $L'(t) \leq 0$, $g(t)$ is non-increasing for $t \geq 0$. The maximum occurs at $t = 0$.

We verify $g(0) \leq 1$ using Condition 3:

$$A_0 - \gamma\log(A_0) \geq \log(C) \implies CA_0^\gamma e^{-A_0} \leq 1.$$

Since $g(t) \leq g(0) = CA_0^\gamma e^{-A_0} \leq 1$ for all $t \geq 0$, the inequality holds. ■

Now, let $\varepsilon = 1 - \beta$ and $y = \varepsilon T$. Then, our goal becomes to show that

$$2C_0 y^{3/2}(1-\varepsilon)^{T-1}\varepsilon^{-5/2} \leq 1$$

and using the fact that $(1-\varepsilon)^{T-1} \leq e^{-\varepsilon(T-1)} \leq e^{-y+1}$, we only need to show

$$2C_0 e y^{3/2} e^{-y}\varepsilon^{-5/2} \leq 1.$$

We can check all the three conditions in Lemma 16 hold with $A_1 = 5$, $A_0 = 8$, $C = 2C_0 e \approx 30.136$, $\gamma = 1.5$ and $\delta = 2.5$, as desired.

Now, we have

$$\begin{aligned}\Lambda_2 &\geq \frac{1}{T}\frac{1}{B_0\bar{u}} - \frac{M}{4T} - \frac{1}{144T^2} - \frac{1}{2\sqrt{T}} \\ &= \frac{3\sqrt{4M+1}}{2\sqrt{T}} - \frac{M}{4T} - \frac{1}{144T^2} - \frac{1}{2\sqrt{T}}\end{aligned}$$

on Region II. To show that

$$\frac{3\sqrt{4M+1}}{2\sqrt{T}} - \frac{M}{4T} - \frac{1}{144T^2} - \frac{1}{2\sqrt{T}} \geq \frac{1+\sqrt{MT}}{T},$$

it suffices to show that

$$\left(\sqrt{4M+1} - \frac{1}{2}\right)\sqrt{T} \geq \frac{M}{4} + \frac{145}{144},$$

and this holds for all $T \geq 4M + 13$.

E.7. Bounding the Ratio of the Tail Sums to the Infinite Sums

Next, we bound the ratio of the tail sum to the infinite sum.

Lemma 17 *If* $T \geq \max\left\{25, \frac{1}{1-\beta}\left(5\log\frac{1}{1-\beta} + 9\right), 39M + 10\right\}$, *then*

$$\left|\frac{R_{Y,T}^{(i)}/T + (\eta L)^2 M R_{D,T}^{(i)}}{Y_\infty/T + (\eta L)^2 M\left(D_\infty - E_\infty/T\right)}\right| \leq \frac{5}{T}$$

with $i \in \{1, 2\}$.

Proof We first consider Region II. For the term Y_∞/T, we have $\inf_{|z|\leq 1} Y_\infty = \frac{1}{1-\beta^2}$. Hence,

$$\begin{aligned}\left|\frac{R_{Y,T}^{(2)}/T}{Y_\infty/T}\right| &\leq \frac{R_{Y,T}^{(2)}}{\inf_{|z|\leq 1} Y_\infty} \\ &= \frac{41}{32}(T+1)^2(1+\beta)\left(1+\sqrt{\beta}\right)^2\beta^T \\ &= \frac{41}{4}\left(\frac{26}{25}\right)^2 T^2\beta^T. \qquad (\because T \geq 25)\end{aligned}$$

For the term $D_\infty - \frac{E_\infty}{T}$, we first show that if $T \geq \frac{9}{1-\beta}$, then its minimum is attained at $z = 0$. Recall that

$$D_\infty - \frac{E_\infty}{T} = \frac{1+\beta}{1-\beta}\frac{1}{C} - \frac{1}{T}\frac{(1-\beta^2)^2 + 2\beta C}{(1-\beta)^2 C^2},$$

where $C = (1+\beta)^2 - 4\beta z^2$. Thus, $D_\infty - \frac{E_\infty}{T}$ is an even function in z and it suffices to show that it is increasing in z^2. Since C is decreasing in z^2, we will show that it decreases in C. Note that $(1-\beta)^2 \leq C \leq (1+\beta)^2$. Now, we have

$$\frac{\partial}{\partial C}\left(D_\infty - \frac{E_\infty}{T}\right) = \frac{1}{C^3}\left(-\frac{1+\beta}{1-\beta}C + \frac{1}{T}\frac{2C\beta + 2(1-\beta^2)^2}{(1-\beta)^2}\right).$$

Since $C > 0$, we only need to show that

$$-\frac{1+\beta}{1-\beta}C + \frac{1}{T}\frac{2C\beta + 2(1-\beta^2)^2}{(1-\beta)^2} \leq 0,$$

or equivalently

$$\frac{2C\beta + 2(1-\beta^2)^2}{1-\beta^2} \leq T.$$

Under the assumption $T \geq \frac{9}{1-\beta}$, it holds for all $C \in [(1-\beta)^2, (1+\beta)^2]$, as desired.

Thus,

$$\begin{aligned}\left|\frac{(\eta L)^2 M R_{D,T}^{(2)}}{(\eta L)^2 M(D_\infty - E_\infty/T)}\right| &\leq \frac{R_{D,T}^{(2)}}{\inf_{|z|\leq 1}(D_\infty - E_\infty/T)} \\ &= \frac{41}{32}\frac{T^2\beta^{T-1}}{1-\beta}\left(\frac{1}{1-\beta^2} - \frac{1}{T}\frac{1+\beta^2}{(1-\beta^2)^2}\right)^{-1}.\end{aligned}$$

Under the assumption $T \geq \frac{8}{1-\beta}$, we have

$$1 - \beta^2 - \frac{1}{T}(1+\beta^2) \geq \frac{7}{8}(1-\beta^2),$$

which implies

$$\left(\frac{1}{1-\beta^2} - \frac{1}{T}\frac{1+\beta^2}{(1-\beta^2)^2}\right)^{-1} \leq \frac{8}{7}(1-\beta^2).$$

We now aim to show that

$$\left|\frac{R_{Y,T}^{(2)}/T}{Y_\infty/T}\right| \leq \frac{5}{T}, \quad \left|\frac{(\eta L)^2 M R_{D,T}^{(2)}}{(\eta L)^2 M(D_\infty - E_\infty/T)}\right| \leq \frac{5}{T},$$

which implies

$$\left|\frac{R_{Y,T}^{(2)}/T + (\eta L)^2 M R_{D,T}^{(2)}}{Y_\infty/T + (\eta L)^2 M\left(D_\infty - E_\infty/T\right)}\right| \leq \frac{5}{T}.$$

From the above, it suffices to show that

$$\frac{41}{4}\left(\frac{26}{25}\right)^2 T^2\beta^T \leq \frac{5}{T}, \quad \frac{41}{32}\cdot\frac{8}{7}(1+\beta)T^2\beta^{T-1} \leq \frac{5}{T}.$$

The above inequalities can be proved by Lemma 16. Using the substitution $\varepsilon = 1-\beta$ and $y = \varepsilon T$, those inequalities can be written as

$$\begin{aligned}C_1 y^3 e^{-y}\varepsilon^{-3} &\leq 1 \\ C_2 y^3 e^{-y}\varepsilon^{-3} &\leq 1,\end{aligned}$$

where $C_1 = \frac{41}{20}\left(\frac{26}{25}\right)^2 = 2.21728$ and $C_2 = \frac{41}{70}e \approx 1.5921$. Setting $A_1 = 5$ and $A_0 = 9$, by Lemma 16, both inequalities can be verified.

Next, we consider Region I. In this case, we bound

$$\left| \frac{R_{Y,T}^{(1)}/T + (\eta L)^2 M R_{D,T}^{(1)}}{Y_\infty/T + (\eta L)^2 M \left(D_\infty - E_\infty/T\right)} \right|.$$

We first find a lower bound of $\frac{Y_\infty}{T} + (\eta L)^2 M \left(D_\infty - \frac{E_\infty}{T}\right)$. We have

$$\frac{Y_\infty}{T} + (\eta L)^2 M \left(D_\infty - \frac{E_\infty}{T}\right) = \frac{c_{-1}}{u} + c_0 + c_1 u + c_2 u^2,$$

where the coefficients are defined in (18).

Now, we prove the following lemma on the minima of the rational functions:

Lemma 18 *Let $F(u) = \frac{c_{-1}}{u} + c_0 + c_1 u + c_2 u^2$ and $G(u) = \frac{c_{-1}}{u} + c_0 + c_1 u$. Then, F has a unique minimizer, denoted by u_f, and $u_f \in \left[\bar{u}, \frac{B+1}{B-1}\right]$, where $\bar{u} = \frac{1}{6}\frac{1}{\sqrt{4M+1}}\frac{1-\beta}{1+\beta}\frac{1}{\sqrt{T}}$. Furthermore,*

$$\left| F(u_f) - \min_{u>0} G(u) \right| \leq \frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2,$$

where $\Psi = 1 - 2M\frac{1+\beta}{1-\beta} + 4MT$.

Proof We first show that the rational part has a unique minimizer in $\left(0, \frac{B-1}{B+1}\right]$ (in terms of u), or equivalently $z \in [1, B)$. To show this, we use the fact that there exists $a > 0$ such that the rational part is strictly convex on $(0, a)$ and strictly concave on (a, ∞). Then, if the derivative at $a' > 0$ is non-negative, the function has a unique minimizer on $(0, a']$. We now evaluate the derivative of the rational part at $u_0 = \frac{B+1}{B-1} = \left(\frac{1+\sqrt{\beta}}{1-\sqrt{\beta}}\right)^2$. The derivative is

$$-\frac{c_{-1}}{u^2} + c_1 + 2c_2 u,$$

and we have

$$-\frac{c_{-1}}{u_0^2} + c_1 + 2c_2 u_0 = \frac{B_1(\beta)}{4T} + M\left(B_2(\beta) - \frac{B_3(\beta)}{2T}\right),$$

where

$$B_1(\beta) := \frac{1+\beta}{1-\beta} - \frac{1-\beta}{1+\beta}\frac{1}{u_0^2}$$
$$B_2(\beta) := \frac{1+\beta}{1-\beta}$$
$$B_3(\beta) := \left(\frac{1+\beta}{1-\beta}\right)^2 + u_0.$$

To show that $-\frac{c_{-1}}{u_0^2} + c_1 + 2c_2 u_0 \geq 0$, it is enough to show that

$$\frac{1}{T}\left(\frac{B_1}{4} - \frac{MB_3}{2}\right) + MB_2 \geq 0.$$

If $M = 0$, it holds for all $T \geq 0$, because $B_1 \geq 0$. Otherwise, we need $T \geq \frac{B_3}{2B_2} - \frac{B_1}{4MB_2}$, and it suffices that $T \geq \frac{B_3}{2B_2}$ since $\frac{B_1}{4MB_2} \geq 0$. We have

$$\sup_{0 \leq \beta < 1} \frac{B_3}{2B_2}(1-\beta) = \sup_{0 \leq \beta < 1} \frac{(1+\beta)^2 + (1+\sqrt{\beta})^4}{2(1+\beta)} = 5.$$

Since the standing assumption on T implies $T \geq \frac{5}{1-\beta}$, this is sufficient to establish the desired inequality.

Meanwhile, G has a unique minimizer

$$u_g = \sqrt{\frac{c_{-1}}{c_1}} = \frac{1-\beta}{1+\beta} \frac{1}{\sqrt{1 - 2M\frac{1+\beta}{1-\beta} + 4MT}} = \frac{1-\beta}{1+\beta} \frac{1}{\sqrt{\Psi}}$$

on $(0, \infty)$. Note that $\bar{u} \leq u_g$.

Define $I = \left[\frac{u_g}{2}, \frac{3u_g}{2}\right]$. Since $F''(u) = 2c_{-1}u^{-3} + 2c_2$, we have

$$F''(u) \geq 2c_{-1}\left(\frac{3u_g}{2}\right)^{-3} + 2c_2 = \frac{16c_1^{3/2}}{27c_{-1}^{1/2}} + 2c_2$$

when $u \in I$.

Now, we show that

$$|c_2| \leq \frac{4}{27}\frac{c_1^{3/2}}{c_{-1}^{1/2}}.$$

This is equivalent to

$$\frac{M}{4T} \leq \frac{4}{27}\frac{1}{4T}\left(\frac{1+\beta}{1-\beta}\right)^2 \Psi^{3/2}.$$

Since $T \geq \frac{8}{1-\beta}$, we have $\Psi = 1 - 2M\frac{1+\beta}{1-\beta} + 4MT \geq 1 + \frac{7}{2}MT$. Thus, it is sufficient to show that

$$\frac{M}{4T} \leq \frac{4}{27}\frac{1}{4T}\left(\frac{1+\beta}{1-\beta}\right)^2 \left(1 + \frac{7}{2}MT\right)^{3/2},$$

or

$$M^{2/3} \leq \frac{2^{4/3}}{9}\left(\frac{1+\beta}{1-\beta}\right)^{4/3}\left(1 + \frac{7}{2}MT\right).$$

This holds for all $M \geq 0$ and $T \geq 1$. Then, $F''(u) \geq \frac{8}{27}\frac{c_1^{3/2}}{c_{-1}^{1/2}}$ on I. Let $\mu = \frac{8}{27}\frac{c_1^{3/2}}{c_{-1}^{1/2}}$. Now, we evaluate F' at the boundary points of I. Then,

$$F'(u_g/2) = c_2 u_g - 3c_1, \quad F'(3u_g/2) = 3c_2 u_g + \frac{5}{9}c_1.$$

Since $c_2 \leq 0$, $u_g > 0$ and $c_1 > 0$, $F'(u_g/2) < 0$. On the other hand,

$$\begin{aligned} F'(3u_g/2) &= 3c_2 u_g + \frac{5}{9}c_1 \\ &\geq -\frac{4}{9}c_1 + \frac{5}{9}c_1 \\ &= \frac{c_1}{9} > 0. \end{aligned}$$

Here, the second line follows from $|c_2| = -c_2 \leq \frac{4}{27}\frac{c_1^{3/2}}{c_{-1}^{1/2}}$. Thus, $u_f \in I$. By the Mean Value Theorem, there exists ξ between u_f and u_g such that $F'(u_f) - F'(u_g) = F''(\xi)(u_f - u_g)$. Since $F'(u_f) = 0$ and $F'(u_g) = G'(u_g) + 2c_2 u_g = 2c_2 u_g$, we have

$$\begin{aligned} |u_f - u_g| &= \frac{2|c_2|u_g}{F''(\xi)} \\ &\leq \frac{2|c_2|u_g}{\mu}. \end{aligned}$$

We now bound the gap between the the minimal values:

$$\begin{aligned} |G(u_g) - F(u_f)| &\leq |G(u_g) - (G(u_f) + c_2 u_f^2)| \\ &\leq |G(u_g) - G(u_f)| + |c_2||u_f|^2 \\ &= |G(u_g) - G(u_f)| + |c_2|\left(|u_g|^2 + 2|u_g||u_g - u_f| + |u_g - u_f|^2\right) \end{aligned}$$

Since $F(u_f) \leq F(u_g)$ and $F(u) = G(u) + c_2 u^2$, we obtain

$$\begin{aligned} G(u_f) + c_2 u_f^2 &\leq G(u_g) + c_2 u_g^2 \\ G(u_f) - G(u_g) &\leq c_2(u_g^2 - u_f^2) \\ |G(u_f) - G(u_g)| &\leq |c_2|\,|u_g - u_f|\,(u_g + u_f) && (\because G(u_f) \geq G(u_g)) \\ &\leq |c_2|\,|u_g - u_f|\,(2u_g + |u_g - u_f|). \end{aligned}$$

Thus,

$$\begin{aligned} |G(u_g) - F(u_f)| &\leq |c_2|(u_g^2 + 4u_g|u_g - u_f| + 2|u_g - u_f|^2) \\ &\leq \frac{M}{4T}\left(u_g + 2|u_g - u_f|\right)^2. \end{aligned}$$

Using $|u_f - u_g| \leq \frac{2|c_2|u_g}{\mu}$, we obtain

$$\begin{aligned} \frac{|u_g - u_f|}{u_g} &\leq \frac{2|c_2|}{\mu} \\ &\leq \frac{27}{16}M\left(\frac{1-\beta}{1+\beta}\right)^2\Psi^{-3/2} \\ &\leq \frac{27}{16}M\Psi^{-3/2} \\ &\leq \frac{27}{16}M\left(1 + \frac{7}{2}MT\right)^{-3/2} \\ &\leq \frac{27}{16}M\left(1 + \frac{7}{2}\cdot 25M\right)^{-3/2} && (\because T \geq 25) \\ &\leq \frac{27}{4}\frac{1}{3\sqrt{3}}\frac{1}{7\cdot 25} && \left(\because \sup_{x\geq 0}\frac{x}{(1+ax)^{3/2}} = \frac{2}{3\sqrt{3}a} \quad \forall a > 0\right) \\ &\leq 0.0075. \end{aligned}$$

Therefore,

$$|G(u_g) - F(u_f)| \leq \frac{M}{4T} \cdot (1 + 2 \cdot 0.0075)^2 \left(\frac{1-\beta}{1+\beta}\right)^2 \frac{1}{\Psi}$$
$$\leq \frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2. \tag{27}$$

■

Thus, we can find a lower bound of the denominator:

$$F(u_f) \geq G(u_g) - \frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2$$
$$\geq \frac{1}{2}\left(\sqrt{1+\frac{7}{2}MT} + 1 - \frac{M}{2}\right) - \frac{1}{3}\frac{M}{T}\left(1+\frac{7}{2}MT\right)^{-1}.$$

Now, we will show that

$$\frac{1}{2}\left(\sqrt{1+\frac{7}{2}MT} + 1 - \frac{M}{2}\right) - \frac{1}{3}\frac{M}{T}\left(1+\frac{7}{2}MT\right)^{-1} - \frac{7}{15}\sqrt{1+\frac{7}{2}MT}$$
$$= \frac{1}{30}\sqrt{1+\frac{7}{2}MT} + \frac{1}{2} - \frac{M}{4} - \frac{1}{3}\frac{M}{T}\left(1+\frac{7}{2}MT\right)^{-1}$$
$$\geq 0$$

if $T \geq 18M$. If $M = 0$, the inequality trivially holds. If $M \neq 0$, let $X := 1 + \frac{7}{2}MT$. Since X is increasing in T, it is enough to show that

$$\frac{1}{30}\sqrt{1+\frac{7}{2}MT} - \frac{1}{3}\frac{M}{T}\left(1+\frac{7}{2}MT\right)^{-1} = \frac{1}{30}\sqrt{X} - \frac{7}{6}\frac{M^2}{X^2 - X}$$

is increasing in X. Differentiating gives

$$\frac{1}{60\sqrt{X}} + \frac{7M^2}{6} \cdot \frac{2X-1}{(X^2-X)^2} \geq 0,$$

as desired. Thus, the left-hand side is non-decreasing in T for all $M \geq 0$, we only need to check the value at $T = 18M$. If $T = 18M$, $X = 63M^2 + 1$. Then,

$$\frac{1}{30}\sqrt{X} - \frac{7}{6}\frac{M^2}{X(X-1)} + \frac{1}{2} - \frac{M}{4} \geq \frac{1}{30}\sqrt{X} - \frac{7}{6 \cdot 63} + \frac{1}{2} - \frac{\sqrt{X}}{12\sqrt{7}}$$
$$\geq 0.$$

Here, we used the fact that $\frac{M^2}{X(X-1)} \leq \frac{1}{63X} \leq \frac{1}{63}$. Thus, the denominator is $\geq \frac{7}{15}\sqrt{1+\frac{7}{2}MT}$.

Next, consider the numerator $R^{(1)}_{Y,T}/T + (\eta L)^2 M R^{(1)}_{D,T}$. We represent this part using the variable u. We will show that $R^{(1)}_{Y,T}$ is decreasing in u. For $R^{(1)}_{Y,T}$, we again use the substitution $s = B_0\frac{u}{1+u} = 4\frac{1+\beta}{1-\beta}\frac{u}{1+u}$. Then,

$$R^{(1)}_{Y,T} = \frac{\exp\left(-\left(T - \frac{\beta}{1-\beta}\right)s\right)}{1 - \exp\left(-s\right)},$$

and it is decreasing in s. Since s is increasing in u, it implies $R_{Y,T}^{(1)}$ is decreasing in u. On the other hand, for $(\eta L)^2 M R_{D,T}^{(1)}$, we first use the fact that $T \geq \frac{8}{1-\beta}$ to obtain

$$\begin{aligned}(\eta L)^2 M R_{D,T}^{(1)} &= \frac{Ms^2}{4} \frac{\exp\left(-\left(T - \frac{1+\beta}{1-\beta}\right)s\right)}{1-\exp(-s)} \\ &\leq \frac{Ms^2}{4} \frac{\exp\left(-\frac{3}{4}Ts\right)}{1-\exp(-s)}.\end{aligned}$$

Now, we will show that $\frac{M}{4}s^2\frac{\exp\left(-\frac{3}{4}Ts\right)}{1-\exp(-s)}$ is non-increasing on $(\bar{s}, \infty)$. If $M = 0$, it is trivially non-increasing because it is 0, so we will assume $M \neq 0$. To show this, we take the logarithm and differentiate:

$$\frac{\partial}{\partial s} \log\left(\frac{M}{4}s^2 \frac{\exp\left(-\frac{3}{4}Ts\right)}{1-\exp(-s)}\right) = \frac{2}{s} - \frac{3}{4}T - \frac{1}{e^s - 1}.$$

Since $\frac{M}{4}s^2\frac{\exp\left(-\frac{3}{4}Ts\right)}{1-\exp(-s)} > 0$, the sign of its derivative is the same as that of log-derivative. By differentiating the log-derivative again, we have

$$\frac{\partial}{\partial s}\left(\frac{2}{s} - \frac{3}{4}T - \frac{1}{e^s - 1}\right) = -\frac{2}{s^2} + \frac{e^s}{(e^s-1)^2} < 0$$

for all $s > 0$. Thus, if

$$\frac{2}{\bar{s}} - \frac{3}{4}T - \frac{1}{e^{\bar{s}} - 1} \leq 0,$$

$\frac{M}{4}s^2\frac{\exp\left(-\frac{3}{4}Ts\right)}{1-\exp(-s)}$ is non-increasing on $(\bar{s}, \infty)$. Since

$$\begin{aligned}\bar{s} &= B_0 \frac{\bar{u}}{1+\bar{u}} \\ &= \frac{2}{3}\frac{1}{\sqrt{4M+1}}\frac{1}{\sqrt{T}}\frac{1}{1+\bar{u}} \\ &\geq \frac{20}{31}\frac{1}{\sqrt{4M+1}}\frac{1}{\sqrt{T}} \qquad (\because T \geq 25),\end{aligned}$$

and $\frac{2}{s} - \frac{1}{e^s-1}$ is decreasing in s and $\leq \frac{3}{2s}$ for $s \leq 1$,

$$\begin{aligned}\frac{2}{\bar{s}} - \frac{3}{4}T - \frac{1}{e^{\bar{s}}-1} &\leq \frac{3}{2\bar{s}} - \frac{3}{4}T \\ &\leq \frac{93}{40}\sqrt{4M+1}\sqrt{T} - \frac{3}{4}T \\ &\leq 0,\end{aligned}$$

whenever $T \geq 39M + 10$. Therefore, $R_{Y,T}^{(1)}/T + (\eta L)^2 M R_{D,T}^{(1)}$ is decreasing on $(\bar{u}, \infty)$ and we can upper bound it by the value at $\bar{u}$. Since $\bar{s} \geq \frac{20}{31}\frac{1}{\sqrt{4M+1}}\frac{1}{\sqrt{T}}$ and $\bar{s} \leq \frac{2}{15}$, we have

$$\begin{aligned}\frac{1}{1-\exp(-\bar{s})} &\leq \frac{93}{56}\sqrt{4M+1}\sqrt{T} \\ \exp(-CT\bar{s}) &\leq \exp\left(-\frac{20}{31}C\sqrt{\frac{T}{4M+1}}\right)\end{aligned}$$

for any $C > 0$. Then, we have

$$\begin{aligned}
\frac{R^{(1)}_{Y,T}}{T} + (\eta L)^2 M R^{(1)}_{D,T} &\leq \frac{1}{1-\exp(-\bar{s})}\left(\frac{1}{T}\exp\left(-\frac{7}{8}T\bar{s}\right) + M\bar{s}^2\exp\left(-\frac{3}{4}T\bar{s}\right)\right) \\
&\leq \frac{\exp\left(-\frac{3}{4}T\bar{s}\right)}{1-\exp(-\bar{s})}\left(\frac{1}{T} + M\bar{s}^2\right) \\
&\leq \frac{93}{56}\sqrt{4M+1}\sqrt{T}\left(\frac{1}{T} + M\bar{s}^2\right)\exp\left(-\frac{15}{31}\sqrt{\frac{T}{4M+1}}\right) \\
&\leq \frac{93}{56}\sqrt{4M+1}\left(\frac{1}{\sqrt{T}} + \frac{1}{9}\frac{4M}{4M+1}\frac{1}{\sqrt{T}}\right)\exp\left(-\frac{15}{31}\sqrt{\frac{T}{4M+1}}\right) \\
&\leq \frac{155}{84}\sqrt{\frac{4M+1}{T}}\exp\left(-\frac{15}{31}\sqrt{\frac{T}{4M+1}}\right).
\end{aligned}$$

Here, the fourth inequality holds because $\bar{s} \leq \frac{2}{3}\frac{1}{\sqrt{4M+1}}\frac{1}{\sqrt{T}}$.
Then, we obtain

$$\left|\frac{R^{(1)}_{Y,T}/T + (\eta L)^2 M R^{(1)}_{D,T}}{Y_\infty/T + (\eta L)^2 M\left(D_\infty - E_\infty/T\right)}\right| \leq \frac{\frac{155}{84}\sqrt{\frac{4M+1}{T}}\exp\left(-\frac{15}{31}\sqrt{\frac{T}{4M+1}}\right)}{\frac{7}{15}\sqrt{1+\frac{7}{2}MT}}.$$

We will show that the ratio is $\leq \frac{5}{T}$, when $T \geq 25$. It is enough to show that

$$\frac{155}{196}\frac{\sqrt{4M+1}\sqrt{T}\exp\left(-\frac{15}{31}\sqrt{\frac{T}{4M+1}}\right)}{\sqrt{1+\frac{7}{2}MT}} \leq 1. \tag{28}$$

We first prove that the left-hand side is decreasing in T for all $M \geq 0$ and $T \geq 5$. To show this, we define

$$g(T) := \frac{\sqrt{4M+1}\sqrt{T}\exp\left(-\frac{15}{31}\sqrt{\frac{T}{4M+1}}\right)}{\sqrt{1+\frac{7}{2}MT}}.$$

We take the logarithm, because it is enough to show that $\log g(T)$ is decreasing in T. To this end, we compute

$$\frac{\partial}{\partial T}\log g(T) = \frac{1}{2T} - \frac{15}{62\sqrt{4M+1}}\frac{1}{\sqrt{T}} - \frac{7}{4}\frac{1}{1+\frac{7}{2}MT}.$$

To show that the log derivative is non-positive, it is enough to show that

$$4M+1 \leq \frac{225}{961}T\left(1+\frac{7}{2}MT\right)^2.$$

Then,

$$\frac{225}{961}T\left(1+\frac{7}{2}MT\right)^2 - (4M+1) = \frac{11025T^3}{3844}M^2 + \left(\frac{1575T^2}{961} - 4\right)M + \left(\frac{225T}{961} - 1\right),$$

where all the coefficients are positive when $T \geq 5$, implying that the left-hand side is increasing in M. Now, it remains to check the value at $M = 0$. Since

$$\frac{225}{961}T\left(1+\frac{7}{2}MT\right)^2 - (4M+1)\Big|_{M=0} = \frac{225}{961}T - 1 > 0,$$

we have the desired result.

Having shown that $g(T)$ is decreasing in T for all $M \geq 0$ and $T \geq 5$, we will evaluate the value at $T = 25$, to conclude that $\frac{155}{196}g(T) \leq 1$ for all $T \geq 25$. To proceed, we define

$$h(M) := \frac{155}{196}g(25) = \frac{775}{196}\frac{\sqrt{4M+1}}{\sqrt{\frac{175M}{2}+1}}\exp\left(-\frac{75}{31\sqrt{4M+1}}\right).$$

Taking logarithm and derivative with respect to M yields

$$\frac{52500M - 5177\sqrt{4M+1} + 600}{62\,(4M+1)^{\frac{3}{2}}\,(175M+2)}.$$

Notice that the numerator is positive on $(0, M_\star)$, where $M_\star \approx 0.1063$ is the solution of $52500M - 5177\sqrt{4M+1} + 600 = 0$, and is negative on $(M_\star, \infty)$. Therefore, we only need to check whether

$$h(0) \leq 1, \quad \lim_{M\to\infty} h(M) \leq 1.$$

Since

$$h(0) = \frac{775}{196}e^{-75/31} \leq 0.352$$
$$\lim_{M\to\infty} h(M) = \frac{155\sqrt{14}}{686} \leq 0.846,$$

the inequality (28) has been proved. ■

We have

$$\Lambda_i = \frac{Y_\infty}{T} + (\eta L)^2 M\left(D_\infty - \frac{E_\infty}{T}\right) - \left(\frac{R^{(i)}_{Y,T}}{T} + (\eta L)^2 M R^{(i)}_{D,T}\right), \quad i \in \{1,2\}.$$

Applying the ratio bound in Lemma 17, we obtain

$$\Lambda_i \geq \left(1-\frac{5}{T}\right)\left(\frac{Y_\infty}{T} + (\eta L)^2 M\left(D_\infty - \frac{E_\infty}{T}\right)\right), \quad i \in \{1,2\}.$$

Then, we have

$$\begin{aligned}
\min_\eta \frac{Y_\infty}{T} + (\eta L)^2 M\left(D_\infty - \frac{E_\infty}{T}\right) &= F(u_f) \\
&\geq G(u_g) - |F(u_f) - G(u_g)| \\
&\geq \frac{1}{2T}\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right) - \frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2.
\end{aligned}$$

Here, the minimum is taken over step sizes η in Regions I and II.

We now show that

$$\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} > 0.$$

We use the lower bound for Ψ:

$$\Psi \geq 1 + \frac{7}{2}MT \geq 1 + \frac{7}{2}M(39M+10).$$

Then,

$$\begin{aligned}\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} &\geq \sqrt{\Psi} + 1 - \frac{M}{2} \\ &\geq \sqrt{1 + \frac{7}{2}M(39M+10)} + 1 - \frac{M}{2} \\ &> 0\end{aligned}$$

for all $M \geq 0$.

Finally, we find an upper bound of the ratio

$$\frac{\frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2}{\frac{1}{2T}\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right)}.$$

We have

$$\begin{aligned}T \cdot \frac{\frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2}{\frac{1}{2T}\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right)} &\leq \frac{2M}{3\Psi}\frac{T}{\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}} \\ &\leq \sup_{M\geq 0}\sup_{T\geq 39M+10}\frac{2M}{3\Psi}\frac{T}{\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}} \\ &\leq \sup_{M\geq 0}\sup_{T\geq 39M+10}\frac{2M}{3\left(1+\frac{7}{2}MT\right)}\frac{T}{\sqrt{1+\frac{7}{2}MT} + 1 - \frac{M}{2}}.\end{aligned}$$

Let

$$\begin{aligned}\psi &:= 1 + \frac{7}{2}MT \\ S &:= \frac{2M}{3\left(1+\frac{7}{2}MT\right)}\frac{T}{\sqrt{1+\frac{7}{2}MT} + 1 - \frac{M}{2}}.\end{aligned}$$

Then,

$$S = \frac{4}{21}\frac{\psi - 1}{\psi\left(\sqrt{\psi} + 1 - \frac{M}{2}\right)}.$$

The constraint $T \geq 39M + 10$ translates into the requirement $\psi \geq \frac{273}{2}M^2 + 35M + 1$.

Now, we will show that $\sup_{M,\psi} S \leq \frac{1}{20}$. It suffices to show that

$$M < 2\sqrt{\psi} - \frac{118}{21} + \frac{160}{21\psi}$$

for all (M, ψ) that satisfy the constraints.

First, if $\psi \geq 9$, we use $\frac{273}{2}M^2 \leq \psi$. Then,

$$M \leq \sqrt{\frac{2}{273}}\sqrt{\psi} \leq 0.0856\sqrt{\psi} \leq 2\sqrt{\psi} - \frac{118}{21} \leq 2\sqrt{\psi} - \frac{118}{21} + \frac{160}{21\psi}$$

holds because $\sqrt{\psi} \geq 3$.

Next, if $1 \leq \psi \leq 9$, we use $35M \leq \psi - 1$. Then, we need to show that

$$\frac{\psi - 1}{35} \leq 2\sqrt{\psi} - \frac{118}{21} + \frac{160}{21\psi}.$$

Since

$$2\sqrt{\psi} - \frac{118}{21} + \frac{160}{21\psi} - \frac{\psi - 1}{35} \geq 0$$

whenever $\psi \in [1, 9]$, the inequality also holds for this case.

Thus, the ratio is bounded by

$$\frac{\frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2}{\frac{1}{2T}\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right)} \leq \frac{1}{20T},$$

and finally

$$\begin{aligned}
&\left(1 - \frac{5}{T}\right)\left(\frac{1}{2T}\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right) - \frac{1}{3}\frac{M}{\Psi T}\left(\frac{1-\beta}{1+\beta}\right)^2\right) \\
&\geq \frac{1}{2T}\left(1 - \frac{5}{T}\right)\left(1 - \frac{1}{20T}\right)\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right) \\
&\geq \frac{1}{2T}\left(1 - \frac{5.05}{T}\right)\left(\sqrt{\Psi} + \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2}\right).
\end{aligned}$$

Appendix F. Proof of Theorem 2

F.1. Proof Sketch

The proof starts from a finite heavy-ball trajectory that is interpolable by a smooth function. For such a tuple $\{(\boldsymbol{x}_t, \boldsymbol{g}_t, f_t)\}_{t=0}^{T-1}$ with $f_0 = 0$, we set

$$G_T^2 := \min_{0\le t<T} \|\boldsymbol{g}_t\|_2^2, \quad \Delta_T := \max_{0\le t<T} \left(-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2\right).$$

A scaling argument then gives the lower bound $\frac{G_T^2}{2\Delta_T}$ for the normalized class $\mathcal{F}_1(1/2)$. Thus the proof reduces to constructing, for every step size, an interpolable heavy-ball trajectory for which this ratio is large enough after the εT scaling.

The interpolation conditions are checked using a pairing form. The values f_t are chosen recursively so that the smooth interpolation conditions reduce to two inequalities, $S_{k,\ell}^- \ge 0$ and $S_{k,\ell}^+ \ge 0$. This form is convenient for the trajectories used below, because their increments have explicit pairwise inner products. In particular, for helical trajectories

$$\boldsymbol{x}_t = (e^{i\phi_t}, \lambda\phi_t),$$

the sums $S_{k,\ell}^{\pm}$ split into an xy-part and a z-part. For the helix with constant angle increment $\phi_t = t\theta$, the gradient norm, the function-value increments, and the ordered sums have closed forms.

We split the step size into four regimes. In Regime I, where $0 < \eta \le a\varepsilon^2$, a constant-gradient construction gives the bound $\frac{1}{2a}$ after multiplying by εT. In Regime IV, where $\eta \ge 1 + \sqrt{3}$, we use orthogonal gradients and constant function values. The condition $\eta \ge 1+\sqrt{3}$ makes the interpolation check direct, and the εT-scaled lower bound diverges.

The two middle regimes use helical trajectories. In Regime III, where $\kappa\varepsilon \le \eta \le 1 + \sqrt{3}$, we first use a finite prefix to match the zero initial momentum and then attach a helix with constant angle increment. The vertical scale is chosen so that the z-part controls the possible negative contribution from the xy-part in $S_{k,\ell}^{\pm}$. This gives interpolation, while the gradient norm stays bounded below and Δ_T stays bounded as $T \to \infty$. Hence the εT-scaled lower bound diverges in this regime.

Regime II, where $a\varepsilon^2 \le \eta \le b\varepsilon$, gives the finite constant in the theorem. The main coordinate is a helix with constant angle increment satisfying $\theta^2 = K\eta$ and $K = \frac{2}{1+\beta}$.

The closed-form estimates show that the tail interpolation inequalities hold, that the tail function-value increment M is negative of order ε, and that $\frac{G^2}{-2M/\varepsilon}$ is at least the target constant. The interpolation inequalities involving the initial index are handled by adding one scalar auxiliary coordinate with a prefix and an exponentially decaying tail. Its scale is chosen to compensate for the finitely many initial-index terms without changing the asymptotic coefficient. Therefore,

$$G_T^2 \ge G^2, \quad \Delta_T \le C_{\Delta,\varepsilon} + (-M)T,$$

and consequently

$$\liminf_{T\to\infty} \varepsilon T \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{G^2}{-2M/\varepsilon} \ge c_0$$

uniformly over $a\varepsilon^2 \le \eta \le b\varepsilon$.

Finally, we choose

$$\varepsilon_{\max} = \frac{1}{8}, \quad a = 4, \quad b = \kappa = 5, \quad \tau = \frac{5}{3}, \quad c_0 = \frac{1}{8}.$$

The remaining estimates verify the required conditions in Regimes II and III. Regimes I and II give the constant $1/8$, while Regimes III and IV give divergent scaled lower bounds. Since these regimes cover all $\eta > 0$, we obtain

$$\liminf_{T\to\infty} \varepsilon T \inf_{\eta>0} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{1}{8}.$$

This proves the claimed lower bound for every $\beta \ge 7/8$.

F.2. Interpolation Conditions

We begin with the interpolation conditions used to certify the finite trajectories constructed below. The first one characterizes when prescribed points, gradients, and function values can be realized by an L-smooth function.

Proposition 19 *[27] $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, f_i)\}_{i\in\mathcal{I}}$ is $\mathcal{F}_L$-interpolable, if and only if for every pair of indices (i, j), with $i, j \in \mathcal{I}$, it holds that*

$$\frac{1}{2L}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|_2^2 - \frac{L}{4}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j - \frac{1}{L}(\boldsymbol{g}_i - \boldsymbol{g}_j)\right\|_2^2 \le f_i - f_j - \langle \boldsymbol{g}_j, \boldsymbol{x}_i - \boldsymbol{x}_j\rangle \tag{29}$$

The following proposition will also be useful for controlling the minimum value of an interpolating function.

Proposition 20 *[7] Set $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, f_i)\}_{i\in\mathcal{I}}$ is $\mathcal{F}_L$-interpolable. Then, there exists a function $f \in \mathcal{F}_L$ such that $f_t = f(\boldsymbol{x}_t)$, $\nabla f(\boldsymbol{x}_t) = \boldsymbol{g}_t$ and*

$$f^\star := \min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = f\left(\boldsymbol{x}_j - \frac{1}{L}\boldsymbol{g}_j\right) = f_j - \frac{1}{2L}\|\boldsymbol{g}_j\|_2^2,$$

where $j \in \arg\min_t f_t - \frac{1}{2L}\|\boldsymbol{g}_t\|_2^2$.

We consider the convergence metric $\mathbb{M}(f, A) = \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2$ for a function f and an algorithm A. Here, $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ is the output of the algorithm A on f.

We define the following function class:

Definition 21 *For some $\Delta > 0$ and $L > 0$, we define*

$$\mathcal{F}_L(\Delta) := \{f : \mathbb{R}^d \to \mathbb{R} : f(\boldsymbol{0}) - f^\star \le \Delta,\ f \text{ is } L\text{-smooth}\}.$$

We next prepare two elementary tools used in the finite construction. The first turns an interpolable heavy-ball trajectory into a lower bound for the normalized class. The second rewrites the interpolation conditions as ordered sums after a particular choice of the values f_t.

Throughout the proof, when we say that $\{\boldsymbol{x}_t\}$ is generated from $\{\boldsymbol{g}_t\}$ by the zero-initialized heavy-ball recursion, we mean

$$\boldsymbol{x}_{-1} = \boldsymbol{x}_0, \quad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \boldsymbol{g}_t + \beta(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$$

for all relevant indices t.

When $\{\boldsymbol{x}_t\}$ is specified first, we define $\{\boldsymbol{g}_t\}$ by the heavy-ball relation as

$$\boldsymbol{g}_t = \frac{(1+\beta)\boldsymbol{x}_t - \boldsymbol{x}_{t+1} - \beta \boldsymbol{x}_{t-1}}{\eta}. \tag{30}$$

Lemma 22 *Fix $T \geq 1$. Suppose that there exists a finite tuple*

$$\mathcal{T}_T := \{(\boldsymbol{x}_t, \boldsymbol{g}_t, f_t)\}_{t=0}^{T-1}$$

with $f_0 = 0$, which is $\mathcal{F}_1$-interpolable and that $\{\boldsymbol{x}_t\}$ is generated from $\{\boldsymbol{g}_t\}$ by the heavy-ball recursion. Define

$$G_T^2 := \min_{0 \leq t < T} \|\boldsymbol{g}_t\|_2^2$$

and

$$\Delta_T := f_0 - \min_{0 \leq t < T} \left(f_t - \frac{1}{2}\|\boldsymbol{g}_t\|_2^2 \right) = \max_{0 \leq t < T} \left(-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2 \right).$$

If $\Delta_T > 0$, then

$$\sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{G_T^2}{2\Delta_T}.$$

Proof By Proposition 20, there exists a 1-smooth function $\tilde{f}$ interpolating the tuple, namely

$$\tilde{f}(\boldsymbol{x}_t) = f_t, \quad \nabla \tilde{f}(\boldsymbol{x}_t) = \boldsymbol{g}_t \quad 0 \leq t < T,$$

and satisfying

$$\tilde{f}^\star = \min_{0 \leq t < T} \left(f_t - \frac{1}{2}\|\boldsymbol{g}_t\|_2^2 \right) = -\Delta_T.$$

After translating the input space, we may assume that the initial point is the origin. Since $f_0 = 0$, the initial gap of $\tilde{f}$ is Δ_T.

Define the rescaled function

$$f(\boldsymbol{x}) := \frac{1}{2\Delta_T} \tilde{f}\left(\sqrt{2\Delta_T}\, \boldsymbol{x} \right).$$

Then $f \in \mathcal{F}_L(\Delta)$. Moreover, if $\boldsymbol{y}_t := \boldsymbol{x}_t / \sqrt{2\Delta_T}$, then

$$\nabla f(\boldsymbol{y}_t) = \frac{1}{\sqrt{2\Delta_T}} \boldsymbol{g}_t.$$

The heavy-ball recursion is invariant under this rescaling, so $\{\boldsymbol{y}_t\}_{t=0}^{T-1}$ is the heavy-ball trajectory on f. Therefore

$$\min_{0 \leq t < T} \|\nabla f(\boldsymbol{y}_t)\|_2^2 = \frac{1}{2\Delta_T} \min_{0 \leq t < T} \|\boldsymbol{g}_t\|_2^2 = \frac{G_T^2}{2\Delta_T}.$$

Taking the supremum over $f \in \mathcal{F}_L(\Delta)$ gives the claim. ■

Lemma 23 (Pairing form of the interpolation conditions) *Let $\{(\boldsymbol{x}_t, \boldsymbol{g}_t)\}_{t=0}^N$ be an arbitrary finite tuple. Define*

$$\Delta \boldsymbol{x}_t := \boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \quad \Delta \boldsymbol{g}_t := \boldsymbol{g}_{t+1} - \boldsymbol{g}_t \quad (0 \le t < N).$$

For $0 \le i, j \le N$, define

$$Q_{i,j} := \frac{1}{4}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|_2^2 - \frac{1}{4}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 + \frac{1}{2}\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{g}_i - \boldsymbol{g}_j\rangle.$$

Set $f_0 = 0$ and define f_t recursively by

$$f_{t+1} - f_t = \langle \boldsymbol{g}_{t+1}, \Delta \boldsymbol{x}_t\rangle - Q_{t,t+1} \tag{31}$$

for $0 \le t < N$. Then the tuple $\{(\boldsymbol{x}_t, \boldsymbol{g}_t, f_t)\}_{t=0}^N$ is $\mathcal{F}_1$-interpolable if and only if, for every $0 \le k < \ell \le N$,

$$S_{k,\ell}^- := \frac{1}{2} \sum_{k \le t < s \le \ell - 1} \langle \Delta \boldsymbol{x}_t - \Delta \boldsymbol{g}_t, \Delta \boldsymbol{x}_s + \Delta \boldsymbol{g}_s\rangle \ge 0 \tag{32}$$

$$S_{k,\ell}^+ := \frac{1}{2} \sum_{k \le t \le s \le \ell - 1} \langle \Delta \boldsymbol{x}_t + \Delta \boldsymbol{g}_t, \Delta \boldsymbol{x}_s - \Delta \boldsymbol{g}_s\rangle \ge 0. \tag{33}$$

Proof For $0 \le i, j \le N$, the 1-smooth interpolation condition from Proposition 19 is

$$Q_{i,j} \le f_i - f_j - \langle \boldsymbol{g}_j, \boldsymbol{x}_i - \boldsymbol{x}_j\rangle.$$

Fix $0 \le k < \ell \le N$. Define

$$\widetilde{S}_{k,\ell}^- := f_k - f_\ell - \langle \boldsymbol{g}_\ell, \boldsymbol{x}_k - \boldsymbol{x}_\ell\rangle - Q_{k,\ell}.$$

We first compute $\widetilde{S}_{k,\ell}^-$. Since

$$\boldsymbol{x}_\ell - \boldsymbol{x}_k = \sum_{t=k}^{\ell-1} \Delta \boldsymbol{x}_t, \quad \boldsymbol{g}_\ell - \boldsymbol{g}_k = \sum_{t=k}^{\ell-1} \Delta \boldsymbol{g}_t,$$

and

$$f_k - f_\ell = -\sum_{t=k}^{\ell-1} \langle \boldsymbol{g}_{t+1}, \Delta \boldsymbol{x}_t\rangle + \sum_{t=k}^{\ell-1} Q_{t,t+1},$$

we have

$$\widetilde{S}_{k,\ell}^- = \sum_{t=k}^{\ell-1} \langle \boldsymbol{g}_\ell - \boldsymbol{g}_{t+1}, \Delta \boldsymbol{x}_t\rangle + \sum_{t=k}^{\ell-1} Q_{t,t+1} - Q_{k,\ell}.$$

Because

$$\boldsymbol{g}_\ell - \boldsymbol{g}_{t+1} = \sum_{s=t+1}^{\ell-1} \Delta \boldsymbol{g}_s,$$

the first term becomes

$$\sum_{t=k}^{\ell-1}\langle \boldsymbol{g}_\ell - \boldsymbol{g}_{t+1}, \Delta\boldsymbol{x}_t\rangle = \sum_{k\le t<s\le \ell-1}\langle \Delta\boldsymbol{x}_t, \Delta\boldsymbol{g}_s\rangle.$$

On the other hand, expanding $Q_{i,j}$ gives

$$\sum_{t=k}^{\ell-1} Q_{t,t+1} - Q_{k,\ell} = \frac{1}{2}\sum_{k\le t<s\le \ell-1}\left(\langle \Delta\boldsymbol{x}_t, \Delta\boldsymbol{x}_s\rangle - \langle \Delta\boldsymbol{g}_t, \Delta\boldsymbol{g}_s\rangle - \langle \Delta\boldsymbol{x}_t, \Delta\boldsymbol{g}_s\rangle - \langle \Delta\boldsymbol{g}_t, \Delta\boldsymbol{x}_s\rangle\right).$$

Combining the above yields

$$\widetilde{S}^-_{k,\ell} = \frac{1}{2}\sum_{k\le t<s\le \ell-1}\langle \Delta\boldsymbol{x}_t - \Delta\boldsymbol{g}_t, \Delta\boldsymbol{x}_s + \Delta\boldsymbol{g}_s\rangle.$$

Thus $\widetilde{S}^-_{k,\ell} = S^-_{k,\ell}$.

Now define

$$\widetilde{S}^+_{k,\ell} := f_\ell - f_k - \langle \boldsymbol{g}_k, \boldsymbol{x}_\ell - \boldsymbol{x}_k\rangle - Q_{\ell,k}.$$

Since $Q_{\ell,k} = Q_{k,\ell}$, we have

$$\widetilde{S}^-_{k,\ell} + \widetilde{S}^+_{k,\ell} = \langle \boldsymbol{g}_\ell - \boldsymbol{g}_k, \boldsymbol{x}_\ell - \boldsymbol{x}_k\rangle - 2Q_{k,\ell}.$$

Writing

$$X := \boldsymbol{x}_\ell - \boldsymbol{x}_k, \quad G := \boldsymbol{g}_\ell - \boldsymbol{g}_k,$$

the right-hand side is

$$\langle G, X\rangle - \frac{1}{2}\|G\|_2^2 + \frac{1}{2}\|X\|_2^2 - \langle X, G\rangle = \frac{1}{2}\langle X - G, X + G\rangle.$$

Since

$$X \pm G = \sum_{t=k}^{\ell-1}(\Delta\boldsymbol{x}_t \pm \Delta\boldsymbol{g}_t),$$

we obtain

$$\widetilde{S}^-_{k,\ell} + \widetilde{S}^+_{k,\ell} = \frac{1}{2}\sum_{k\le t,s\le \ell-1}\langle \Delta\boldsymbol{x}_t - \Delta\boldsymbol{g}_t, \Delta\boldsymbol{x}_s + \Delta\boldsymbol{g}_s\rangle.$$

Subtracting the expression for $S^-_{k,\ell}$ and using symmetry of the inner product gives

$$\widetilde{S}^+_{k,\ell} = \frac{1}{2}\sum_{k\le t\le s\le \ell-1}\langle \Delta\boldsymbol{x}_t + \Delta\boldsymbol{g}_t, \Delta\boldsymbol{x}_s - \Delta\boldsymbol{g}_s\rangle.$$

Thus, $\widetilde{S}^+_{k,\ell} = S^+_{k,\ell}$.

The interpolation condition must hold for every ordered pair of indices. For $k < \ell$, the ordered pair (k, ℓ) is equivalent to $S^-_{k,\ell} \ge 0$, while the ordered pair (ℓ, k) is equivalent to $S^+_{k,\ell} \ge 0$. Hence the tuple is $\mathcal{F}_1$-interpolable if and only if $S^\pm_{k,\ell} \ge 0$ hold for all $k < \ell$. ■

F.3. Helical Trajectory

We collect the elementary identities for helical trajectories. Throughout this subsection, we identify $\mathbb{R}^2$ with $\mathbb{C}$, and use the real inner product

$$\langle u, v\rangle := \mathrm{Re}(\overline{u}v) \quad (u, v \in \mathbb{C}).$$

Lemma 24 *Let $\{\phi_t\}_{t=-1}^{N+1}$ be a real sequence and define*

$$e_t := e^{i\phi_t}, \quad \delta_t := \phi_t - \phi_{t-1}.$$

For $\lambda > 0$, define the helical trajectory

$$\boldsymbol{x}_t := (e_t, \lambda\phi_t) \in \mathbb{C} \times \mathbb{R}.$$

Define the gradients $\{\boldsymbol{g}_t\}$ by the heavy-ball relation (30) Then

$$\boldsymbol{g}_t = (h_t e_t, g_t^z),$$

where

$$h_t := \frac{(1+\beta) - e^{i\delta_{t+1}} - \beta e^{-i\delta_t}}{\eta}$$

and

$$g_t^z = \frac{\lambda(\beta\delta_t - \delta_{t+1})}{\eta}.$$

Moreover, if

$$\rho_t := \frac{(1+\beta)\delta_{t+1} - \delta_{t+2} - \beta\delta_t}{\eta},$$

then

$$\Delta\boldsymbol{x}_t^z = \lambda\delta_{t+1}, \quad \Delta\boldsymbol{g}_t^z = \lambda\rho_t.$$

Consequently, the z-parts of the interpolation conditions in pairing form are

$$S_{k,\ell}^{-,z} = \frac{\lambda^2}{2} \sum_{k\le t<s\le \ell-1} (\delta_{t+1} - \rho_t)(\delta_{s+1} + \rho_s)$$
$$S_{k,\ell}^{+,z} = \frac{\lambda^2}{2} \sum_{k\le t\le s\le \ell-1} (\delta_{t+1} + \rho_t)(\delta_{s+1} - \rho_s).$$

For the xy-part, define

$$U_t^- := (e^{i\delta_{t+1}} - 1) - (h_{t+1}e^{i\delta_{t+1}} - h_t)$$
$$U_t^+ := (e^{i\delta_{t+1}} - 1) + (h_{t+1}e^{i\delta_{t+1}} - h_t).$$

Then

$$\Delta\boldsymbol{x}_t^{xy} - \Delta\boldsymbol{g}_t^{xy} = e_t U_t^-, \quad \Delta\boldsymbol{x}_t^{xy} + \Delta\boldsymbol{g}_t^{xy} = e_t U_t^+.$$

Hence the xy-parts of the interpolation conditions in pairing form are

$$S_{k,\ell}^{-,xy} = \frac{1}{2} \sum_{k \le t < s \le \ell - 1} \mathrm{Re}\left(\overline{U_t^-} U_s^+ e^{i(\phi_s - \phi_t)} \right) \tag{34}$$

$$S_{k,\ell}^{+,xy} = \frac{1}{2} \sum_{k \le t \le s \le \ell - 1} \mathrm{Re}\left(\overline{U_t^+} U_s^- e^{i(\phi_s - \phi_t)} \right). \tag{35}$$

In particular,

$$S_{k,\ell}^{-} = S_{k,\ell}^{-,xy} + S_{k,\ell}^{-,z}, \quad S_{k,\ell}^{+} = S_{k,\ell}^{+,xy} + S_{k,\ell}^{+,z}.$$

Proof The formula for $\boldsymbol{g}_t$ follows from

$$e_{t+1} = e_t e^{i\delta_{t+1}}, \quad e_{t-1} = e_t e^{-i\delta_t},$$

and

$$\phi_{t+1} - \phi_t = \delta_{t+1}, \quad \phi_t - \phi_{t-1} = \delta_t.$$

Meanwhile, we have

$$\Delta \boldsymbol{x}_t^{xy} = e_t (e^{i\delta_{t+1}} - 1)$$

and

$$\Delta \boldsymbol{g}_t^{xy} = h_{t+1} e_{t+1} - h_t e_t = e_t (h_{t+1} e^{i\delta_{t+1}} - h_t).$$

Substituting these identities into Lemma 23 gives the claimed decomposition. ■

Lemma 25 *Let $0 < \theta < 2\pi$, $\lambda > 0$, and*

$$\boldsymbol{x}_t := (e^{it\theta}, \lambda t\theta) \in \mathbb{C} \times \mathbb{R}.$$

Set

$$\varepsilon := 1 - \beta$$

and define the heavy-ball gradients

$$\boldsymbol{g}_t := \frac{(1+\beta)\boldsymbol{x}_t - \boldsymbol{x}_{t+1} - \beta \boldsymbol{x}_{t-1}}{\eta}.$$

Then

$$\boldsymbol{g}_t = \left(C e^{it\theta}, -\frac{\lambda \varepsilon \theta}{\eta} \right),$$

where

$$C := \frac{(1+\beta)(1-\cos\theta) - i\varepsilon \sin\theta}{\eta}.$$

Writing $C = C_r + iC_i$, we have

$$C_i = -\frac{\varepsilon \sin\theta}{\eta}.$$

The gradient norm is independent of t, and equals

$$G^2 := \|\boldsymbol{g}_t\|_2^2 = |C|^2 + \frac{\lambda^2 \varepsilon^2 \theta^2}{\eta^2}.$$

For an interval of length $n = \ell - k$, the ordered sums depend only on n. They are

$$S_n^- = \frac{1}{2}\sum_{m=1}^{n-1}(n-m)\left[2\alpha\left((1-|C|^2)\cos(m\theta) - 2C_i\sin(m\theta)\right) + \lambda^2\theta^2\right],$$

and

$$S_n^+ = \frac{1}{2}\sum_{m=0}^{n-1}(n-m)\left[2\alpha\left((1-|C|^2)\cos(m\theta) + 2C_i\sin(m\theta)\right) + \lambda^2\theta^2\right],$$

where

$$\alpha := 1 - \cos\theta.$$

Equivalently, define

$$\begin{aligned}
A_n^- &:= \sum_{m=1}^{n-1}(n-m)\cos(m\theta) = \frac{1}{2}\left[\left(\frac{\sin(n\theta/2)}{\sin(\theta/2)}\right)^2 - n\right] \\
A_n^+ &:= \sum_{m=0}^{n-1}(n-m)\cos(m\theta) = \frac{1}{2}\left[\left(\frac{\sin(n\theta/2)}{\sin(\theta/2)}\right)^2 + n\right] \\
D_n &:= n\sin\theta - \sin(n\theta).
\end{aligned}$$

Then

$$2S_n^\pm = \lambda^2\theta^2\frac{n(n\pm 1)}{2} + 2\alpha(1-|C|^2)A_n^\pm \pm 2C_iD_n.$$

Proof In this case, $\delta_t = \theta$ for all t. Hence $h_t = C$, and

$$U_t^- = (1-C)(e^{i\theta}-1), \quad U_t^+ = (1+C)(e^{i\theta}-1).$$

Thus U_t^- and U_t^+ are independent of t. We write them as U^- and U^+.

Since

$$\overline{U^-}U^+ = |e^{i\theta}-1|^2(1-\overline{C})(1+C),$$

we get

$$\overline{U^-}U^+ = 2(1-\cos\theta)(1-|C|^2+C-\overline{C}).$$

Therefore

$$\overline{U^-}U^+ = 2\alpha(1-|C|^2+2iC_i).$$

Similarly,

$$\overline{U^+}U^- = 2\alpha(1-|C|^2-2iC_i).$$

Taking real parts after multiplication by $e^{im\theta}$ gives

$$\begin{aligned}
\mathrm{Re}\left(\overline{U^-}U^+e^{im\theta}\right) &= 2\alpha\left((1-|C|^2)\cos(m\theta) - 2C_i\sin(m\theta)\right) \\
\mathrm{Re}\left(\overline{U^+}U^-e^{im\theta}\right) &= 2\alpha\left((1-|C|^2)\cos(m\theta) + 2C_i\sin(m\theta)\right).
\end{aligned}$$

The z-increments satisfy

$$\Delta\boldsymbol{x}_t^z = \lambda\theta, \quad \Delta\boldsymbol{g}_t^z = 0.$$

Substitution into Lemma 24 gives the formulas for S_n^- and S_n^+.

It remains to compute the trigonometric sums. We have

$$\sum_{m=-(n-1)}^{n-1}(n-|m|)e^{im\theta} = \left|\sum_{r=0}^{n-1} e^{ir\theta}\right|^2 = \left(\frac{\sin(n\theta/2)}{\sin(\theta/2)}\right)^2.$$

Taking real parts yields the stated formulas for A_n^- and A_n^+. Also,

$$\sum_{m=1}^{n-1}(n-m)\sin(m\theta) = \frac{n\sin\theta - \sin(n\theta)}{2(1-\cos\theta)} = \frac{D_n}{2\alpha}.$$

Substituting these closed forms gives the final expressions. ■

Lemma 26 *Consider the helix trajectory from Lemma 25, and define f_t by*

$$f_{t+1} - f_t = \langle \boldsymbol{g}_{t+1}, \Delta\boldsymbol{x}_t\rangle - Q_{t,t+1}.$$

Then the function value increment

$$M := f_{t+1} - f_t$$

is independent of t, and equals

$$M = \frac{1-\cos\theta}{2}(1-|C|^2) - \frac{\varepsilon\sin^2\theta}{\eta} + \lambda^2\theta^2\left(\frac{1}{4} - \frac{\varepsilon}{\eta}\right).$$

Assume in addition that

$$K := \frac{2}{1+\beta}, \quad \theta^2 = K\eta.$$

Define

$$H := \frac{2(1-\cos\theta)}{\theta^2}, \quad R := \frac{\sin\theta}{\theta}, \quad s := \frac{\varepsilon\theta}{\eta}.$$

Then

$$C_r = H, \quad C_i = -sR, \quad |C|^2 = H^2 + s^2R^2, \quad G^2 = H^2 + s^2(R^2+\lambda^2).$$

Moreover,

$$M = A_{xy}(\eta,\theta) + K\left(\frac{\eta}{4} - \varepsilon\right)\lambda^2,$$

where

$$A_{xy}(\eta,\theta) := \frac{1-\cos\theta}{2}(1-H^2-s^2R^2) - \frac{\varepsilon\sin^2\theta}{\eta}.$$

Equivalently,

$$M = K\varepsilon\left[-R^2 - \lambda^2 + \frac{\eta}{4\varepsilon}\left(H(1-H^2-s^2R^2) + \lambda^2\right)\right].$$

Proof By the shift-invariance, it suffices to compute M at a single index. For the xy-part,

$$\Delta \boldsymbol{x}_t^{xy} = e^{it\theta}(e^{i\theta} - 1), \quad \Delta \boldsymbol{g}_t^{xy} = Ce^{it\theta}(e^{i\theta} - 1).$$

Thus

$$\left\langle \boldsymbol{g}_{t+1}^{xy}, \Delta \boldsymbol{x}_t^{xy} \right\rangle = \mathrm{Re}\left(\overline{C}e^{-i\theta}(e^{i\theta} - 1)\right) = C_r(1 - \cos\theta) + C_i \sin\theta.$$

Also,

$$Q_{t,t+1}^{xy} = \frac{1 - \cos\theta}{2}(|C|^2 - 1) + (1 - \cos\theta)C_r.$$

Therefore

$$\left\langle \boldsymbol{g}_{t+1}^{xy}, \Delta \boldsymbol{x}_t^{xy} \right\rangle - Q_{t,t+1}^{xy} = \frac{1 - \cos\theta}{2}(1 - |C|^2) + C_i \sin\theta.$$

Since

$$C_i = -\frac{\varepsilon \sin\theta}{\eta},$$

the xy-contribution is

$$\frac{1 - \cos\theta}{2}(1 - |C|^2) - \frac{\varepsilon \sin^2\theta}{\eta}.$$

For the z-part,

$$\Delta \boldsymbol{x}_t^{z} = \lambda\theta, \quad \Delta \boldsymbol{g}_t^{z} = 0, \quad \boldsymbol{g}_{t+1}^{z} = -\frac{\lambda\varepsilon\theta}{\eta}.$$

Hence

$$\left\langle \boldsymbol{g}_{t+1}^{z}, \Delta \boldsymbol{x}_t^{z} \right\rangle = -\frac{\lambda^2\varepsilon\theta^2}{\eta},$$

and

$$Q_{t,t+1}^{z} = -\frac{\lambda^2\theta^2}{4}.$$

The z-contribution is therefore

$$\lambda^2\theta^2\left(\frac{1}{4} - \frac{\varepsilon}{\eta}\right).$$

Adding the two gives the first formula for M.

Now assume $\theta^2 = K\eta$, where $K = 2/(1 + \beta)$. Since

$$1 - \cos\theta = \frac{\theta^2}{2}H, \quad \sin\theta = \theta R,$$

we obtain

$$C_r = \frac{(1 + \beta)(1 - \cos\theta)}{\eta} = H \tag{36}$$

$$C_i = -\frac{\varepsilon \sin\theta}{\eta} = -sR. \tag{37}$$

The formulas for $|C|^2$, G^2, and M follow immediately.

■

F.4. Regime Decomposition

We prove the following heavy-ball lower bound:

$$\liminf_{T\to\infty} \varepsilon T \inf_{\eta>0} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge c,$$

for a universal constant $c > 0$.

We prove the lower bound by splitting the step size into four regimes. Throughout the proof, we write $\varepsilon := 1 - \beta$. The constants $a > 0$, $b > 0$, and $\varepsilon_{\max} > 0$ will be fixed in the final admissibility check. For now, assume $0 < \varepsilon \le \varepsilon_{\max}$. We decompose the step size range as follows:

$$0 < \eta \le a\varepsilon^2, \quad a\varepsilon^2 \le \eta \le b\varepsilon, \quad b\varepsilon \le \eta \le 1+\sqrt{3}, \quad \eta \ge 1+\sqrt{3}.$$

The proof in each regime uses a different finite interpolable trajectory.

Regime I. For $0 < \eta \le a\varepsilon^2$, we use a constant-gradient construction.

Regime II. For $a\varepsilon^2 \le \eta \le b\varepsilon$, we use an exact helical tail together with an auxiliary coordinate. This is the only regime that determines the final finite constant.

Regime III. For $b\varepsilon \le \eta \le 1+\sqrt{3}$, we choose the initial phase increments so that the trajectory has zero initial momentum, and then set the phase increment equal to θ for all later indices. After choosing the scale λ large enough, the values f_t increase from that point onward, so $\max_{0\le t<T}\left\{-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2\right\}$ is independently of T. As a result, the lower bound diverges.

Regime IV. For $\eta \ge 1+\sqrt{3}$, we use an orthogonal-gradient construction. The threshold $1+\sqrt{3}$ appears from a interpolation inequality, and this regime also gives a divergent lower bound.

Combining the four regimes gives a uniform lower bound over all step sizes. The elementary Regimes I and IV are handled first, followed by the helical Regimes III and II.

F.5. Regimes I and IV

In this subsection, we handle the two elementary step size regimes.

Proposition 27 (Regime I) *Fix $a > 0$, $0 < \varepsilon < 1$, and $\beta = 1 - \varepsilon$. Then*

$$\liminf_{T\to\infty} \varepsilon T \inf_{0<\eta\le a\varepsilon^2} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{1}{2a}.$$

Proof Fix $T \ge 1$ and $0 < \eta \le a\varepsilon^2$. Let $\boldsymbol{g}$ be a unit vector, and set

$$\boldsymbol{g}_t \equiv \boldsymbol{g}$$

for $0 \le t < T$. Define the trajectory by the heavy-ball recursion

$$\boldsymbol{x}_{-1} = \boldsymbol{x}_0 = \boldsymbol{0}, \quad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\boldsymbol{g}_t + \beta(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}),$$

for $0 \le t < T$. Finally, define

$$f_t := \langle \boldsymbol{g}, \boldsymbol{x}_t\rangle.$$

Then $f_0 = 0$.

We first check interpolation. For every i, j,

$$\boldsymbol{g}_i - \boldsymbol{g}_j = \mathbf{0},$$

and hence

$$Q_{i,j} = -\frac{1}{4}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \le 0.$$

On the other hand,

$$f_i - f_j - \langle \boldsymbol{g}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle = \langle \boldsymbol{g}, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle - \langle \boldsymbol{g}, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle = 0.$$

Thus the tuple is $\mathcal{F}_1$-interpolable.

It remains to bound G_T^2 and Δ_T. Clearly,

$$G_T^2 = \min_{0 \le t < T} \|\boldsymbol{g}_t\|_2^2 = 1.$$

Let

$$\boldsymbol{d}_t := \boldsymbol{x}_t - \boldsymbol{x}_{t-1}.$$

Since $\boldsymbol{d}_0 = \mathbf{0}$, the recursion gives

$$\boldsymbol{d}_{t+1} = -\eta \boldsymbol{g} + \beta \boldsymbol{d}_t = -\eta \sum_{j=0}^{t} \beta^j \boldsymbol{g}.$$

Therefore, for $0 \le t < T$,

$$-f_t = -\langle \boldsymbol{g}, \boldsymbol{x}_t \rangle = \eta \sum_{r=0}^{t-1} \sum_{j=0}^{r} \beta^j \le \frac{\eta t}{\varepsilon} \le \frac{\eta T}{\varepsilon}.$$

Since $\|\boldsymbol{g}_t\|_2 = 1$, we obtain

$$\Delta_T = \max_{0 \le t < T} \left(-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2 \right) \le \frac{1}{2} + \frac{\eta T}{\varepsilon} \le \frac{1}{2} + a\varepsilon T.$$

By Lemma 22,

$$\sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \le t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{1}{1 + 2a\varepsilon T}.$$

Thus

$$\varepsilon T \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \le t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{\varepsilon T}{1 + 2a\varepsilon T}.$$

Taking the infimum over $0 < \eta \le a\varepsilon^2$ and then taking $\liminf_{T \to \infty}$ gives

$$\liminf_{T \to \infty} \varepsilon T \inf_{0 < \eta \le a\varepsilon^2} \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \le t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{1}{2a}.$$

■

Proposition 28 (Regime IV) *Fix $0 < \varepsilon < 1$, and let $\beta = 1 - \varepsilon$. Then*

$$\liminf_{T\to\infty} \varepsilon T \inf_{\eta \geq 1+\sqrt{3}} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\leq t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 = +\infty.$$

Proof Fix $T \geq 1$ and $\eta \geq 1 + \sqrt{3}$. We construct T-dimensional instance. Let $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{T-1}$ be the standard orthonormal basis. Define

$$\boldsymbol{g}_t := \boldsymbol{e}_t, \quad f_t := 0,$$

for $0 \leq t < T$. Let $\{\boldsymbol{x}_t\}$ be generated from $\{\boldsymbol{g}_t\}$ by the zero-initialized heavy-ball recursion.

We check the 1-smooth interpolation condition. Fix $0 \leq i < j < T$, and set

$$z_{i,j} := \langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{e}_i \rangle.$$

Let $\boldsymbol{d}_t := \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$. Since

$$\boldsymbol{d}_{t+1} = -\eta \boldsymbol{e}_t + \beta \boldsymbol{d}_t,$$

the coefficient of $\boldsymbol{e}_i$ in $\boldsymbol{x}_j - \boldsymbol{x}_i$ is

$$-\eta \sum_{r=0}^{j-i-1} \beta^r.$$

Therefore

$$z_{i,j} = \eta \sum_{r=0}^{j-i-1} \beta^r \geq \eta \geq 1 + \sqrt{3}.$$

Moreover,

$$\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{e}_j \rangle = 0, \quad \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \geq z_{i,j}.$$

For the ordered pair (i, j), we have

$$\boldsymbol{g}_i - \boldsymbol{g}_j = \boldsymbol{e}_i - \boldsymbol{e}_j, \quad \|\boldsymbol{g}_i - \boldsymbol{g}_j\|_2^2 = 2.$$

Hence

$$Q_{i,j} = \frac{1}{2} - \frac{1}{4}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 + \frac{1}{2}\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{e}_i - \boldsymbol{e}_j \rangle,$$

and

$$Q_{i,j} \leq \frac{1}{2} - \frac{1}{4} z_{i,j}^2 + \frac{1}{2} z_{i,j}.$$

Since $z_{i,j} \geq 1 + \sqrt{3}$, the scalar inequality

$$\frac{1}{2} - \frac{1}{4}z^2 + \frac{1}{2}z \leq 0, \quad \forall z \geq 1 + \sqrt{3}$$

implies $Q_{i,j} \leq 0$. On the other hand,

$$f_i - f_j - \langle \boldsymbol{g}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle = -\langle \boldsymbol{e}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle = 0.$$

Thus the interpolation condition holds for the ordered pair (i, j).

For the reverse ordered pair (j, i), the left-hand side is the same because $Q_{j,i} = Q_{i,j}$. The right-hand side is

$$f_j - f_i - \langle \boldsymbol{g}_i, \boldsymbol{x}_j - \boldsymbol{x}_i \rangle = \langle \boldsymbol{e}_i, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle = z_{i,j} \geq 0.$$

Since $Q_{j,i} = Q_{i,j} \leq 0$, the reverse interpolation condition also holds. Therefore the tuple is $\mathcal{F}_1$-interpolable.

Finally,

$$G_T^2 = \min_{0 \leq t < T} \|\boldsymbol{g}_t\|_2^2 = 1,$$

and

$$\Delta_T = \max_{0 \leq t < T} \left(-f_t + \frac{1}{2} \|\boldsymbol{g}_t\|_2^2 \right) = \frac{1}{2}.$$

By Lemma 22,

$$\sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq 1.$$

Consequently,

$$\varepsilon T \inf_{\eta \geq 1+\sqrt{3}} \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \varepsilon T.$$

Letting $T \to \infty$ proves the claim. ■

F.6. Regime III

In this subsection, we handle Regime III. We consider the interval

$$[\kappa\varepsilon, 1 + \sqrt{3}],$$

for some constant $\kappa > 4$. Throughout this subsection, we assume that this interval is nonempty:

$$\kappa\varepsilon \leq 1 + \sqrt{3}. \tag{38}$$

We also choose $\tau > 0$ such that, with

$$N := \left\lceil \frac{\tau}{\varepsilon} \right\rceil,$$

the following two admissibility conditions hold:

$$N < \frac{\sqrt{\beta}}{1 - \sqrt{\beta}}, \tag{39}$$

and

$$p_{\max} := \frac{\varepsilon(1 + \pi^2/\tau^2)}{\kappa} < 1. \tag{40}$$

Proposition 29 (Regime III) *Under (38)–(40),*

$$\liminf_{T \to \infty} \varepsilon T \inf_{\eta \in [\kappa\varepsilon, 1+\sqrt{3}]} \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 = +\infty.$$

We prove the proposition through a zero-initialized sine prefix followed by an exact helical tail.

Construction. Fix $\eta \in [\kappa\varepsilon, 1+\sqrt{3}]$, and set

$$K := \frac{2}{1+\beta} = \frac{2}{2-\varepsilon}, \quad \theta^2 := K\eta.$$

We also assume that (38)–(40) hold. Then, there exists

$$\omega \in \left(0, \frac{\pi}{N+1}\right)$$

such that

$$\sqrt{\beta}\sin((N+1)\omega) = \sin(N\omega). \tag{41}$$

Indeed, the function

$$F(\omega) := \sqrt{\beta}\sin((N+1)\omega) - \sin(N\omega)$$

satisfies

$$\lim_{\omega\downarrow 0}\frac{F(\omega)}{\omega} = \sqrt{\beta}(N+1) - N > 0$$

by (39), while

$$F\left(\frac{\pi}{N+1}\right) = -\sin\left(\frac{N\pi}{N+1}\right) < 0.$$

Hence the root exists by continuity.

Define $\{\delta_t\}$ by

$$\delta_t := \begin{cases} \theta\beta^{(t-N)/2}\frac{\sin(t\omega)}{\sin(N\omega)}, & 0 \le t \le N+1, \\ \theta, & t \ge N+1. \end{cases}$$

Then, $\delta_t > 0$ for $t \ge 1$.

Set

$$\phi_t := \begin{cases} 0, & t \in \{-1, 0\}, \\ \sum_{j=1}^{t}\delta_j, & t \ge 1. \end{cases}$$

For a parameter $\lambda > 0$, define

$$\boldsymbol{x}_t := (e^{i\phi_t}, \lambda\phi_t) \in \mathbb{C}\times\mathbb{R}.$$

Then $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$. Define $\boldsymbol{g}_t$ as (30) so that the heavy-ball recursion holds.

Interpolation conditions.

Lemma 30 *Define*

$$m_t := \frac{\delta_{t+1} - \beta\delta_t}{\eta}.$$

Then

$$m_t \ge s := \frac{\varepsilon\theta}{\eta},$$

for all $t \ge 0$. Moreover, there exists a sequence $\{p_t\}$ with $0 \le p_t \le p_{\max} < 1$ for all $t \ge 0$, such that

$$\Delta\boldsymbol{g}_t^z = \lambda p_t \delta_{t+1}.$$

Consequently,

$$\Delta\boldsymbol{x}_t^z \pm \Delta\boldsymbol{g}_t^z = \lambda(1 \pm p_t)\delta_{t+1}.$$

Proof For $0 \leq t < N$, the sequence $\{\delta_t\}$ satisfies

$$\delta_{t+2} - 2\sqrt{\beta}\cos\omega\,\delta_{t+1} + \beta\delta_t = 0.$$

Therefore

$$m_{t+1} - m_t = \frac{\delta_{t+2} - (1+\beta)\delta_{t+1} + \beta\delta_t}{\eta} = -p\delta_{t+1},$$

where

$$p := \frac{(1-\sqrt{\beta})^2 + 2\sqrt{\beta}(1-\cos\omega)}{\eta}.$$

Thus

$$m_{t+1} = m_t - p\delta_{t+1},$$

for $0 \leq t < N$. Since $p \geq 0$ and $\delta_{t+1} > 0$, the sequence m_t is non-increasing on $0 \leq t \leq N$.

For $t \geq N$, we have $\delta_t = \theta$, so

$$m_t = \frac{\theta - \beta\theta}{\eta} = \frac{\varepsilon\theta}{\eta} = s.$$

In particular, $m_N = s$, and hence $m_t \geq s$ for every $t \geq 0$.

Thus, it remains to bound p. Since

$$1 - \sqrt{\beta} = \frac{\varepsilon}{1+\sqrt{\beta}} \leq \varepsilon$$

and

$$1 - \cos\omega \leq \frac{\omega^2}{2},$$

we get

$$p\eta \leq \varepsilon^2 + \omega^2.$$

Using

$$\omega < \frac{\pi}{N+1} \leq \frac{\pi\varepsilon}{\tau},$$

which holds because $N + 1 \geq \frac{\tau}{\varepsilon}$, we obtain

$$p\eta \leq \varepsilon^2\left(1 + \frac{\pi^2}{\tau^2}\right).$$

Since $\eta \geq \kappa\varepsilon$,

$$p \leq \frac{\varepsilon(1+\pi^2/\tau^2)}{\kappa} = p_{\max}.$$

Define

$$p_t := \begin{cases} p, & t < N, \\ 0, & t \geq N. \end{cases}$$

Then, $m_{t+1} = m_t - p_t\delta_{t+1}$ holds.

Finally, the z-component of $\boldsymbol{g}_t$ is

$$\boldsymbol{g}_t^z = \frac{\lambda(\beta\delta_t - \delta_{t+1})}{\eta} = -\lambda m_t.$$

Hence

$$\Delta\boldsymbol{g}_t^z = \boldsymbol{g}_{t+1}^z - \boldsymbol{g}_t^z = \lambda(m_t - m_{t+1}) = \lambda p_t\delta_{t+1}.$$

Since $\Delta\boldsymbol{x}_t^z = \lambda\delta_{t+1}$, the identities $\Delta\boldsymbol{x}_t^z \pm \Delta\boldsymbol{g}_t^z = \lambda(1 \pm p_t)\delta_{t+1}$ hold. ■

Function values.

Lemma 31 *There exists $\lambda_{\mathrm{IC}} = \lambda_{\mathrm{IC}}(\varepsilon, \kappa, \tau) > 0$ such that, for every $\lambda \geq \lambda_{\mathrm{IC}}$, the tuple $\{(\boldsymbol{x}_i, \boldsymbol{g}_i, f_i)\}$ is $\mathcal{F}_1$-interpolable, if $f_0 = 0$ and $\{f_t\}$ is defined by (31).*

Proof We first obtain a positive lower bound for the z-factors. Define

$$a_{\min} := \min\left\{1, \min_{1\leq j\leq N} \beta^{(j-N)/2} \frac{\sin(j\omega)}{\sin(N\omega)}\right\} > 0.$$

Also,

$$\theta^2 = K\eta, \quad K \geq 1, \quad \eta \geq \kappa\varepsilon.$$

Therefore, for all $t \geq 0$,

$$\delta_{t+1} \geq a_{\min}\theta \geq a_{\min}\sqrt{\kappa\varepsilon} > 0.$$

Set $\delta_{\min} := a_{\min}\sqrt{\kappa\varepsilon} > 0$ and

$$Z_\varepsilon := (1 - p_{\max})\delta_{\min}^2 > 0.$$

Since $0 \leq p_t \leq p_{\max} < 1$ for all $t \geq 0$ by Lemma 30,

$$(1 - p_t)(1 + p_s)\delta_{t+1}\delta_{s+1} \geq Z_\varepsilon$$
$$(1 + p_t)(1 - p_s)\delta_{t+1}\delta_{s+1} \geq Z_\varepsilon$$

for all $t, s \geq 0$.

Now we bound the xy-part. We have

$$\boldsymbol{g}_t^{xy} = h_t e^{i\phi_t}, \quad h_t = \frac{(1+\beta) - e^{i\delta_{t+1}} - \beta e^{-i\delta_t}}{\eta}.$$

Since $\eta \geq \kappa\varepsilon$, we have

$$|h_t| \leq \frac{1 + \beta + 1 + \beta}{\eta} \leq \frac{4}{\kappa\varepsilon}.$$

Thus

$$|\Delta\boldsymbol{g}_t^{xy}| \leq |h_{t+1}| + |h_t| \leq \frac{8}{\kappa\varepsilon},$$

while

$$|\Delta\boldsymbol{x}_t^{xy}| \leq 2.$$

Now, define

$$U_\varepsilon := 2 + \frac{8}{\kappa\varepsilon}.$$

Then

$$|\Delta\boldsymbol{x}_t^{xy} \pm \Delta\boldsymbol{g}_t^{xy}| \leq U_\varepsilon.$$

Fix $0 \leq k < \ell$, and set $m := \ell - k$.

Write $S_{k,\ell}^{\pm} = S_{k,\ell}^{\pm,xy} + S_{k,\ell}^{\pm,z}$. This decomposition holds because the inner product of u and v is simply the sum of the contributions from their xy and z parts.

Then, the xy-part of $S_{k,\ell}^{-}$ satisfies

$$|S_{k,\ell}^{-,xy}| \leq \frac{U_\varepsilon^2}{2}\frac{m(m-1)}{2} = \frac{U_\varepsilon^2}{4}m(m-1).$$

The z-part of $S^-_{k,\ell}$ satisfies

$$S^{-,z}_{k,\ell} = \frac{\lambda^2}{2} \sum_{k \le t < s \le \ell - 1} (1 - p_t)(1 + p_s)\delta_{t+1}\delta_{s+1} \ge \frac{\lambda^2 Z_\varepsilon}{4} m(m-1).$$

Similarly,

$$|S^{+,xy}_{k,\ell}| \le \frac{U_\varepsilon^2}{4} m(m+1),$$
$$S^{+,z}_{k,\ell} \ge \frac{\lambda^2 Z_\varepsilon}{4} m(m+1).$$

Choose

$$\lambda_{\mathrm{IC}}^2 := \frac{U_\varepsilon^2}{Z_\varepsilon}.$$

Then, for every $\lambda \ge \lambda_{\mathrm{IC}}$,

$$S^-_{k,\ell} = S^{-,xy}_{k,\ell} + S^{-,z}_{k,\ell} \ge 0,$$
$$S^+_{k,\ell} = S^{+,xy}_{k,\ell} + S^{+,z}_{k,\ell} \ge 0$$

for every $k < \ell$. Thus, the interpolation condition holds by Lemma 23. ■

Lemma 32 *There exists $\lambda_f(\varepsilon, \kappa, \tau) > 0$ such that, for every $\lambda \ge \lambda_f$, $\{f_t\}$ defined as (31) satisfies*

$$f_{t+1} - f_t \ge 1,$$

for all $t \ge N$.

Proof For $t \ge N$, we have

$$\delta_t = \delta_{t+1} = \theta.$$

By Lemma 26, we have

$$M_0(\eta, \lambda) := f_{t+1} - f_t = A_{xy}(\eta) + K\left(\frac{\eta}{4} - \varepsilon\right)\lambda^2,$$

for all $t \ge N$, where

$$A_{xy}(\eta) = \frac{1 - \cos\theta}{2}(1 - H^2 - s^2R^2) - \frac{\varepsilon \sin^2\theta}{\eta}.$$

Since

$$\eta \ge \kappa\varepsilon, \quad \kappa > 4,$$

we have

$$\frac{\eta}{4} - \varepsilon \ge \left(\frac{\kappa}{4} - 1\right)\varepsilon.$$

Define

$$c_\kappa := \frac{\kappa}{4} - 1 > 0.$$

As $K = \frac{2}{1+\beta} \geq 1$, we have $K\left(\frac{\eta}{4} - \varepsilon\right) \geq c_\kappa \varepsilon$.

The function $A_{xy}(\eta)$ is continuous on the compact interval

$$\eta \in [\kappa\varepsilon, 1 + \sqrt{3}].$$

Hence

$$B_\varepsilon := \max_{\eta \in [\kappa\varepsilon, 1+\sqrt{3}]} |A_{xy}(\eta)| < \infty.$$

Choose

$$\lambda_f^2 := \frac{B_\varepsilon + 1}{c_\kappa \varepsilon}.$$

Then for every $\lambda \geq \lambda_f$,

$$M_0(\eta, \lambda) \geq -B_\varepsilon + c_\kappa \varepsilon \lambda^2 \geq 1.$$

This proves the claim. ■

Gradient norm lower bound.

Lemma 33 *Choose*

$$\lambda \geq \max\{\lambda_{\mathrm{IC}}, \lambda_f\}.$$

Then there exist constants

$$G_\varepsilon > 0, \quad C_\varepsilon < \infty,$$

depending on $\varepsilon, \kappa, \tau, \lambda$, but not on T or η, such that

$$G_T^2 := \min_{0 \leq t < T} \|\boldsymbol{g}_t\|_2^2 \geq G_\varepsilon^2,$$

and

$$\Delta_T := \max_{0 \leq t < T} \left(-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2\right) \leq C_\varepsilon.$$

Proof We first prove the gradient norm lower bound. Let

$$G_\varepsilon^2 := \frac{\lambda^2 \varepsilon^2}{1 + \sqrt{3}} > 0.$$

By Lemma 30,

$$\boldsymbol{g}_t^z = -\lambda m_t, \quad m_t \geq s = \frac{\varepsilon\theta}{\eta}.$$

Therefore

$$\|\boldsymbol{g}_t\|_2^2 \geq \lambda^2 s^2.$$

Since

$$s^2 = \frac{\varepsilon^2\theta^2}{\eta^2} = \frac{K\varepsilon^2}{\eta}, \quad K = \frac{2}{1+\beta} \geq 1, \quad \eta \leq 1 + \sqrt{3},$$

we obtain

$$s^2 \geq \frac{\varepsilon^2}{1 + \sqrt{3}}.$$

Thus

$$\|\boldsymbol{g}_t\|_2^2 \geq G_\varepsilon^2 > 0.$$

Now we prove the initial gap bound. For $t \geq N$, Lemmas 26 and 32 gives

$$f_t = f_N + (t-N)M_0(\eta,\lambda), \quad M_0(\eta,\lambda) \geq 1.$$

Moreover, for $t \geq N$, the gradient norm is constant. Denote this constant by

$$G_{\text{tail}}^2(\eta,\lambda).$$

Then, for $t \geq N$,

$$-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2 = -f_N - (t-N)M_0(\eta,\lambda) + \frac{1}{2}G_{\text{tail}}^2(\eta,\lambda) \leq -f_N + \frac{1}{2}G_{\text{tail}}^2(\eta,\lambda).$$

Therefore the maximum of

$$-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2$$

for $t \geq N$ is attained at $t = N$.

It remains to control the finite prefix $0 \leq t \leq N$. For fixed $\varepsilon, \kappa, \tau$, the integers N, the root ω, and the coefficients δ_t/θ are independent of η. Since $\theta^2 = K\eta$ and $\eta \geq \kappa\varepsilon > 0$, the vectors $\boldsymbol{x}_t$ and $\boldsymbol{g}_t$ are continuous functions of η on $[\kappa\varepsilon, 1+\sqrt{3}]$ for every $0 \leq t \leq N$. Moreover, f_t is continuous by the finite recursion (31). Hence

$$C_\varepsilon := \max_{\eta\in[\kappa\varepsilon,1+\sqrt{3}]} \max_{0\leq t\leq N} \left(-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2\right) < \infty.$$

Therefore,

$$\Delta_T \leq C_\varepsilon$$

for every T and every $\eta \in [\kappa\varepsilon, 1+\sqrt{3}]$. ■

Proof for Regime II. **Proof** [Proof of Proposition 29] Fix $T \geq 1$ and $\eta \in [\kappa\varepsilon, 1+\sqrt{3}]$. Construct the trajectory above, choose $\lambda \geq \max\{\lambda_{\text{IC}}, \lambda_f\}$, define the gradients by heavy-ball, and define f_t as (31).

By construction, the tuple is heavy-ball consistent and has zero initial momentum. By Lemma 31, it is $\mathcal{F}_1$-interpolable. By Lemma 33,

$$G_T^2 \geq G_\varepsilon^2, \quad \Delta_T \leq C_\varepsilon,$$

where $G_\varepsilon > 0$ and $C_\varepsilon < \infty$ do not depend on T or η. Therefore, by Lemma 22,

$$\sup_{f\in\mathcal{F}_1(1/2)} \min_{0\leq t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{G_\varepsilon^2}{2C_\varepsilon}.$$

Taking the infimum over $\eta \in [\kappa\varepsilon, 1+\sqrt{3}]$ preserves the same lower bound:

$$\inf_{\eta\in[\kappa\varepsilon,1+\sqrt{3}]} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\leq t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{G_\varepsilon^2}{2C_\varepsilon}.$$

Multiplying by εT, we get

$$\varepsilon T \inf_{\eta\in[\kappa\varepsilon,1+\sqrt{3}]} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \varepsilon T \frac{G_\varepsilon^2}{2C_\varepsilon}.$$

Since

$$\varepsilon > 0, \quad G_\varepsilon > 0, \quad C_\varepsilon < \infty,$$

the right-hand side diverges as $T \to \infty$. Hence

$$\liminf_{T\to\infty} \varepsilon T \inf_{\eta\in[\kappa\varepsilon,1+\sqrt{3}]} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 = +\infty.$$

■

F.7. Regime II

We now prove the lower bound in Regime II.

Admissible parameters. Fix constants

$$0 < \varepsilon_{\max} < 1, \quad a > 0, \quad b > 0, \quad \tau > 0, \quad c_0 > 0.$$

Define

$$K_{\max} := \frac{2}{2-\varepsilon_{\max}}.$$

For $0 < \varepsilon \le \varepsilon_{\max}$, set

$$\beta := 1-\varepsilon, \quad K := \frac{2}{1+\beta} = \frac{2}{2-\varepsilon}.$$

The interval we consider is

$$a\varepsilon^2 \le \eta \le b\varepsilon.$$

For each η in this interval, define

$$\theta^2 := K\eta, \quad s := \frac{\varepsilon\theta}{\eta}.$$

Then

$$\theta^2 \le K_{\max} b\varepsilon_{\max}, \quad s^2 \le \frac{K_{\max}}{a}.$$

Define

$$\Theta^2 := K_{\max} b\varepsilon_{\max}, \quad S^2 := \frac{K_{\max}}{a}.$$

We assume that there exist constants $h, r, d, \rho > 0$ such that, for every $0 < u \le \Theta$ and every integer $n \ge 1$,

$$H(u) := \frac{2(1-\cos u)}{u^2} \ge h, \tag{42a}$$

$$R(u) := \frac{\sin u}{u} \ge r, \tag{42b}$$

$$0 \le n\sin u - \sin(nu) \le du^2 n(n+1), \tag{42c}$$

$$1 - H(u)^2 \le \rho u^2. \tag{42d}$$

We note that $H \le 1$ and $R \le 1$ always hold.

We also assume the following conditions:

$$\Theta \le 1, \tag{43a}$$
$$S < 1, \tag{43b}$$
$$\frac{1-S^2}{2} - 2Sd > 0, \tag{43c}$$
$$S^2 + \rho\Theta^2 < 1, \tag{43d}$$
$$\frac{b}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) < 1 + r^2, \tag{43e}$$
$$\frac{1}{a}\left(\frac{1}{(1+\sqrt{1-\varepsilon_{\max}})^2} + \frac{\pi^2}{\tau^2}\right) < 1, \tag{43f}$$
$$\frac{h^2}{4K_{\max}} \ge c_0, \tag{43g}$$
$$\left\lceil \frac{\tau}{\varepsilon_{\max}} \right\rceil \ge 2. \tag{43h}$$

Proposition 34 (Regime II) *Assume the admissible parameters (42) and (43). Then, for every fixed $0 < \varepsilon \le \varepsilon_{\max}$,*

$$\liminf_{T\to\infty} \varepsilon T \inf_{a\varepsilon^2 \le \eta \le b\varepsilon} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge c_0.$$

The constants used in the construction may depend on ε, but they do not depend on T or η.

Construction. For $t \ge 0$, define the primary helical coordinate

$$\boldsymbol{x}_t^h := (e^{it\theta}, t\theta) \in \mathbb{C}\times\mathbb{R}.$$

Set $\boldsymbol{x}_{-1}^h := \boldsymbol{x}_0^h$.

The gradient is

$$\boldsymbol{g}_t^h := \frac{(1+\beta)\boldsymbol{x}_t^h - \boldsymbol{x}_{t+1}^h - \beta\boldsymbol{x}_{t-1}^h}{\eta}.$$

For every $t \ge 1$,

$$\boldsymbol{g}_t^{h,xy} = Ce^{it\theta}, \quad \boldsymbol{g}_t^{h,z} = -s,$$

where

$$C = H - isR, \quad H = \frac{2(1-\cos\theta)}{\theta^2}, \quad R = \frac{\sin\theta}{\theta}.$$

The gradient norm is

$$G^2 := \|\boldsymbol{g}_t^h\|_2^2 = H^2 + s^2(R^2+1).$$

We define $\{f_t^h\}$ as (31), with $\{(\boldsymbol{x}_t^h, \boldsymbol{g}_t^h)\}$.

Let $P_{k,\ell}^{\pm}$ denote the expressions in (32) and (33) computed using only $\{(\boldsymbol{x}_t^h, \boldsymbol{g}_t^h)\}$.

Lemma 35 *Under (43a), (43c) and (43d), for every $1 \le k < \ell$, we have*

$$P_{k,\ell}^- \ge 0, \quad P_{k,\ell}^+ \ge 0.$$

Proof Fix $k \geq 1$ and set $\ell - k = n \geq 1$. By the definition of $\{\boldsymbol{x}_t^h\}$, we can apply Lemma 25, with $\lambda = 1$. Thus, we denote $P_{k,\ell}^{\pm} = S_n^{\pm}$. Then,

$$2S_n^- = \theta^2 \frac{n(n-1)}{2} + 2(1-\cos\theta)(1-|C|^2)A_n^- - 2C_i D_n, \tag{44}$$

$$2S_n^+ = \theta^2 \frac{n(n+1)}{2} + 2(1-\cos\theta)(1-|C|^2)A_n^+ + 2C_i D_n, \tag{45}$$

where

$$C_i = -sR,$$

$$A_n^- := \sum_{m=1}^{n-1} (n-m)\cos(m\theta), \quad A_n^+ := \sum_{m=0}^{n-1} (n-m)\cos(m\theta),$$

and

$$D_n := n\sin\theta - \sin(n\theta).$$

We have

$$0 \leq A_n^+ \leq \frac{n(n+1)}{2}, \quad |A_n^-| \leq \frac{n(n-1)}{2}.$$

Since $0 < \theta \leq 1 < \pi$ by (43a), we have $D_n \geq 0$. Also, by (42c),

$$D_n \leq d\theta^2 n(n+1).$$

We first prove $S_n^+ \geq 0$. Since

$$1 - |C|^2 = 1 - H^2 - s^2R^2 \geq -S^2,$$

and

$$2(1-\cos\theta) \leq \theta^2,$$

we obtain

$$2S_n^+ \geq \theta^2 n(n+1)\left(\frac{1-S^2}{2} - 2Sd\right), \tag{46}$$

from (45). By (43c), the right-hand side is non-negative.

Now we prove $S_n^- \geq 0$. Since $C_i = -sR \leq 0$ and $D_n \geq 0$,

$$-2C_i D_n \geq 0.$$

Moreover,

$$|1 - |C|^2| = |1 - H^2 - s^2R^2| \leq 1 - H^2 + s^2R^2 \leq \rho\Theta^2 + S^2.$$

Using (44), we get

$$2S_n^- \geq \theta^2 \frac{n(n-1)}{2}\left(1 - S^2 - \rho\Theta^2\right). \tag{47}$$

By (43d), this is nonnegative for $n \geq 2$. For $n = 1$, the sum S_n^- is 0. ■

Lemma 36 *Under* (43e) *and* (43g)*, for* $t \geq 1$*,*

$$M := f_{t+1}^h - f_t^h$$

satisfies

$$-\frac{M}{K\varepsilon} \geq \tilde{\delta},$$

where

$$\tilde{\delta} := 1 + r^2 - \frac{b}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) > 0.$$

Moreover,

$$\frac{G^2}{-2M/\varepsilon} \geq c_0.$$

Proof By Lemma 26 with $\lambda = 1$, we have

$$M = K\varepsilon\left[-R^2 - 1 + \frac{\eta}{4\varepsilon}\left(1 + H(1 - H^2 - s^2R^2)\right)\right].$$

Let

$$B := 1 + H(1 - H^2 - s^2R^2).$$

Since $0 < H = \frac{2(1-\cos u)}{u^2} \leq 1$,

$$0 \leq H(1 - H^2) \leq \frac{2}{3\sqrt{3}}.$$

Also,

$$-Hs^2R^2 \leq 0.$$

Thus,

$$B \leq 1 + \frac{2}{3\sqrt{3}}.$$

Since $\eta \leq b\varepsilon$ and $R \geq r$ by (42b),

$$\frac{M}{K\varepsilon} \leq -1 - r^2 + \frac{b}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) = -\tilde{\delta}.$$

Hence $M \leq -K\varepsilon\tilde{\delta} < 0$.

For the coefficient, first note that

$$G^2 = H^2 + s^2(R^2 + 1) \geq H^2 \geq h^2$$

by (42a). Also, by (43b) and the fact that $H^2, R^2 \leq 1$ and $s^2 \leq S^2$, we have

$$\begin{aligned} B &= 1 + H(1 - H^2 - s^2R^2) \\ &= 1 + H(1 - H^2) - Hs^2R^2 \\ &\geq 1 - Hs^2R^2 \\ &\geq 1 - S^2 > 0. \end{aligned}$$

Therefore,

$$R^2 + 1 - \frac{\eta}{4\varepsilon} B \le R^2 + 1 \le 2.$$

Using the formula for M,

$$-\frac{2M}{\varepsilon} = 2K \left(R^2 + 1 - \frac{\eta}{4\varepsilon} B \right) \le 4K \le 4K_{\max}.$$

Consequently,

$$\frac{G^2}{-2M/\varepsilon} \ge \frac{h^2}{4K_{\max}} \ge c_0$$

by (43g). ■

We keep the primary helical coordinate and add one auxiliary scalar coordinate.

Let

$$N := \left\lceil \frac{\tau}{\varepsilon} \right\rceil.$$

By (43h), for every $0 < \varepsilon \le \varepsilon_{\max}$, $N \ge 2$.

Choose

$$\omega \in \left(0, \frac{\pi}{N+1}\right)$$

such that

$$\sin((N+1)\omega) = \sqrt{\beta} \sin(N\omega). \tag{48}$$

Such a root exists because the function

$$F(\omega) := \sin((N+1)\omega) - \sqrt{\beta} \sin(N\omega)$$

satisfies

$$\lim_{\omega \downarrow 0} \frac{F(\omega)}{\omega} = N + 1 - \sqrt{\beta} N > 0,$$

while

$$F\left(\frac{\pi}{N+1}\right) = -\sqrt{\beta} \sin\left(\frac{N\pi}{N+1}\right) < 0.$$

Define $\{\gamma_t\}$ by

$$\gamma_t := \begin{cases} \beta^{(t-N)/2} \frac{\sin(t\omega)}{\sin(N\omega)}, & 0 \le t \le N, \\ \beta^{t-N}, & t \ge N \end{cases}$$

Note that the definition of γ_t for $0 \le t \le N$ agrees with $t = N+1$, due to (48).

Define

$$x_t^c := \begin{cases} 0, & t \in \{-1, 0\}, \\ \sum_{j=1}^{t} \gamma_j, & t \ge 1. \end{cases}$$

For $\Gamma > 0$, define the full trajectory

$$\boldsymbol{x}_t := (\boldsymbol{x}_t^h, \Gamma x_t^c).$$

This satisfies $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$. Define gradients $\boldsymbol{g}_t$ according to (30). We write

$$\boldsymbol{g}_t = (\boldsymbol{g}_t^h, \Gamma g_t^c).$$

Equivalently,

$$g_t^c = \frac{\beta\gamma_t - \gamma_{t+1}}{\eta}.$$

For the unscaled auxiliary coordinate, define

$$m_t^c := \frac{\gamma_{t+1} - \beta\gamma_t}{\eta}.$$

For $t < N$, the sequence γ_t satisfies

$$\gamma_{t+2} - 2\sqrt{\beta}\cos\omega\,\gamma_{t+1} + \beta\gamma_t = 0.$$

Hence

$$m_{t+1}^c - m_t^c = -p\gamma_{t+1},$$

where

$$p := \frac{(1-\sqrt{\beta})^2 + 2\sqrt{\beta}(1-\cos\omega)}{\eta}.$$

For $t \geq N$, it holds that

$$\gamma_{t+1} = \beta\gamma_t,$$

and hence $m_t^c = 0$.

Define

$$p_t := \begin{cases} p, & t < N, \\ 0, & t \geq N. \end{cases}$$

Since

$$1 - \sqrt{\beta} = \frac{\varepsilon}{1+\sqrt{\beta}}$$

and $2(1-\cos\omega) \leq \omega^2$, we have

$$p\eta \leq \frac{\varepsilon^2}{(1+\sqrt{\beta})^2} + \sqrt{\beta}\omega^2.$$

Moreover,

$$\omega < \frac{\pi}{N+1} \leq \frac{\pi\varepsilon}{\tau}.$$

Since $0 < \varepsilon \leq \varepsilon_{\max}$, we get

$$p \leq \frac{1}{a}\left(\frac{1}{(1+\sqrt{1-\varepsilon_{\max}})^2} + \frac{\pi^2}{\tau^2}\right).$$

Let

$$p_{\max} := \frac{1}{a}\left(\frac{1}{(1+\sqrt{1-\varepsilon_{\max}})^2} + \frac{\pi^2}{\tau^2}\right).$$

Then, by (43f),

$$0 \leq p_t \leq p_{\max} < 1.$$

For the auxiliary scalar coordinate with Γ,

$$\Delta x_t^c = \gamma_{t+1}, \quad g_t^c = -m_t^c, \quad \Delta g_t^c = p_t\gamma_{t+1}.$$

Thus,

$$\Delta x_t^c \pm \Delta g_t^c = (1 \pm p_t)\gamma_{t+1}.$$

Because $\gamma_{t+1} > 0$ and $0 \le p_t < 1$, $(1 \pm p_t)\gamma_{t+1} \ge 0$ for all $t \ge 0$.

Since $N \ge 2$ and $\omega \in (0, \pi/(N+1))$,

$$\sin\omega > 0, \quad \sin(2\omega) > 0, \quad \sin(N\omega) > 0.$$

Therefore

$$\gamma_1 > 0, \quad \gamma_2 > 0.$$

Define

$$c_+ := \frac{1}{2}(1 - p_{\max})\gamma_1^2 > 0, \quad c_- := \frac{1}{2}(1 - p_{\max})\gamma_1\gamma_2 > 0.$$

Let $C_{k,\ell}^{\pm}$ denote the contribution of the auxiliary coordinate in $S_{k,\ell}^{\pm}$ with $\Gamma = 1$. Since $\Gamma(1 \pm p_t)\gamma_{t+1} \ge 0$ for all $t \ge 0$, we have

$$C_{k,\ell}^{\pm} \ge 0.$$

For $k = 0$, we also note that

$$C_{0,\ell}^{+} \ge c_+, \quad \ell \ge 1$$
$$C_{0,\ell}^{-} \ge c_-, \quad \ell \ge 2.$$

For $\ell = 1$, $C_{0,\ell}^{-} = 0$.

Interpolation conditions. Let $P_{k,\ell}^{\pm}$ be the contribution of the helical coordinates in $S_{k,\ell}^{\pm}$. The full sums are

$$S_{k,\ell}^{\pm} = P_{k,\ell}^{\pm} + \Gamma^2 C_{k,\ell}^{\pm}.$$

If $k \ge 1$, by Lemma 35,

$$P_{k,\ell}^{\pm} \ge 0.$$

Since $C_{k,\ell}^{\pm} \ge 0$, all pairs of (k, ℓ) with $k \ge 1$ satisfy the inequalities.

It remains to check the case $k = 0$. We prove a uniform bound on $|P_{k,\ell}^{\pm}|$.

Define

$$A_t^h := \Delta \boldsymbol{x}_t^h - \Delta \boldsymbol{g}_t^h, \quad B_t^h := \Delta \boldsymbol{x}_t^h + \Delta \boldsymbol{g}_t^h.$$

For $t \ge 1$, we have

$$A_t^h = (e^{it\theta}U^-, \theta), \quad B_t^h = (e^{it\theta}U^+, \theta),$$

where

$$U^- = (1 - C)(e^{i\theta} - 1), \quad U^+ = (1 + C)(e^{i\theta} - 1).$$

Because

$$|C| \le H + sR \le 1 + S,$$

we have

$$|U^-| \le (2 + S)\theta, \quad |U^+| \le (2 + S)\theta.$$

Also,

$$\theta^2 = K\eta \geq \eta \geq a\varepsilon^2,$$

so

$$\theta \geq \sqrt{a}\varepsilon.$$

Since $0 < \theta \leq \Theta \leq 1$ by (43a),

$$|1 - e^{i\theta}| = 2\sin(\theta/2) \geq \frac{\theta}{2}.$$

Therefore, for every $n \geq 1$,

$$\left|\sum_{s=1}^{n} e^{is\theta}\right| \leq \frac{2}{|1 - e^{i\theta}|} \leq \frac{4}{\sqrt{a}\varepsilon}.$$

Using the z-component bound $n\theta \leq n\Theta$, there exists $L_\varepsilon < \infty$ such that, for every $n \geq 1$ and every $\eta \in [a\varepsilon^2, b\varepsilon]$,

$$\left\|\sum_{s=1}^{n} A_s^h\right\|_2 \leq L_\varepsilon(1+n)$$
$$\left\|\sum_{s=1}^{n} B_s^h\right\|_2 \leq L_\varepsilon(1+n).$$

We also need to bound the vectors A_0^h and B_0^h. Since $\boldsymbol{x}_{-1}^h = \boldsymbol{x}_0^h$,

$$\boldsymbol{g}_0^h = \frac{\boldsymbol{x}_0^h - \boldsymbol{x}_1^h}{\eta}.$$

Moreover,

$$\|\boldsymbol{x}_1^h - \boldsymbol{x}_0^h\|_2 \leq |e^{i\theta} - 1| + \theta \leq 2\theta,$$

and hence

$$\|\boldsymbol{g}_0^h\|_2 \leq \frac{2\theta}{\eta} = \frac{2K}{\theta} \leq \frac{2K_{\max}}{\sqrt{a}\varepsilon}.$$

The gradient satisfies

$$\|\boldsymbol{g}_1^h\|_2^2 = G^2 = H^2 + s^2(R^2+1) \leq 1 + 2S^2.$$

Thus, there exists $D_\varepsilon < \infty$ such that

$$\|A_0^h\|_2 \leq D_\varepsilon, \quad \|B_0^h\|_2 \leq D_\varepsilon,$$

for all $\eta \in [a\varepsilon^2, b\varepsilon]$, because both A_0^h and B_0^h are continuous in η and the interval $[a\varepsilon^2, b\varepsilon]$ is compact.

Now fix $\ell \geq 2$ and set $n := \ell - 1$. We obtain

$$P_{0,\ell}^- = P_{1,\ell}^- + \frac{1}{2}\left\langle A_0^h, \sum_{s=1}^{n} B_s^h \right\rangle \tag{49}$$

$$P_{0,\ell}^+ = P_{1,\ell}^+ + \frac{1}{2}\langle B_0^h, A_0^h\rangle + \frac{1}{2}\left\langle B_0^h, \sum_{s=1}^{n} A_s^h \right\rangle. \tag{50}$$

From the right-hand sides of (46) and (47), define the margins

$$m_+ := \frac{1-S^2}{2} - 2Sd, \quad m_- := 1 - S^2 - \rho\Theta^2.$$

Then, $m_+, m_- > 0$ holds due to (43d) and (43c).

We now have

$$P^+_{1,\ell} \geq \frac{1}{2}\theta^2 n(n+1)m_+,$$
$$P^-_{1,\ell} \geq \frac{1}{4}\theta^2 n(n-1)m_-.$$

Using $\theta^2 = K\eta \geq \eta \geq a\varepsilon^2$, set

$$q_+ := \frac{1}{2}a\varepsilon^2 m_+, \quad q_- := \frac{1}{4}a\varepsilon^2 m_-.$$

Then $q_+, q_- > 0$, and

$$P^+_{1,\ell} \geq q_+ n(n+1),$$
$$P^-_{1,\ell} \geq q_- n(n-1).$$

We have

$$\begin{aligned} P^-_{0,\ell} &= P^-_{1,\ell} + \frac{1}{2}\left\langle A^h_0, \sum_{s=1}^n B^h_s \right\rangle \\ &\geq q_- n(n-1) - \frac{1}{2}\|A^h_0\|_2 \left\| \sum_{s=1}^n B^h_s \right\|_2 \\ &\geq q_- n(n-1) - \frac{1}{2}D_\varepsilon L_\varepsilon (n+1) \end{aligned}$$

and

$$\begin{aligned} P^+_{0,\ell} &= P^+_{1,\ell} + \frac{1}{2}\langle B^h_0, A^h_0\rangle + \frac{1}{2}\left\langle B^h_0, \sum_{s=1}^n A^h_s \right\rangle \\ &\geq q_+ n(n+1) - \frac{1}{2}\|B^h_0\|_2\|A^h_0\|_2 - \frac{1}{2}\|B^h_0\|_2 \left\| \sum_{s=1}^n A^h_s \right\|_2 \\ &\geq q_+ n(n+1) - \frac{1}{2}D^2_\varepsilon - \frac{1}{2}D_\varepsilon L_\varepsilon (n+1). \end{aligned}$$

Define

$$C_\varepsilon := \frac{1}{2}D^2_\varepsilon + \frac{1}{2}D_\varepsilon L_\varepsilon.$$

Since

$$-\frac{1}{2}D_\varepsilon L_\varepsilon (n+1) \geq -C_\varepsilon (n+1)$$
$$-\frac{1}{2}D^2_\varepsilon - \frac{1}{2}D_\varepsilon L_\varepsilon (n+1) \geq -C_\varepsilon (n+1),$$

we have

$$P_{0,\ell}^{\pm} \geq q_{\pm} n(n \pm 1) - C_\varepsilon (n+1).$$

Denote $[x]_+ := \max\{x, 0\}$. Then, we obtain

$$\begin{aligned}
[-P_{0,\ell}^{-}]_+ &\leq [C_\varepsilon (n+1) - q_- n(n-1)]_+ \\
&= [C_\varepsilon + (C_\varepsilon + q_-)n - q_- n^2]_+ \\
&\leq C_\varepsilon + \frac{(C_\varepsilon + q_-)^2}{4q_-}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
[-P_{0,\ell}^{+}]_+ &\leq [C_\varepsilon (n+1) - q_+ n(n+1)]_+ \\
&\leq [C_\varepsilon (n+1) - q_+ n^2]_+ \\
&= [C_\varepsilon + C_\varepsilon n - q_+ n^2]_+ \\
&\leq C_\varepsilon + \frac{C_\varepsilon^2}{4q_+}.
\end{aligned}$$

Define

$$B_\varepsilon := D_\varepsilon^2 + C_\varepsilon + \max\left\{ \frac{(C_\varepsilon + q_-)^2}{4q_-}, \frac{C_\varepsilon^2}{4q_+} \right\} < \infty.$$

Then, for every $\eta \in [a\varepsilon^2, b\varepsilon]$ and $\ell \geq 1$, we have

$$[-P_{0,\ell}^{\pm}]_+ \leq B_\varepsilon.$$

Choose Γ so that

$$\Gamma^2 \geq \frac{B_\varepsilon}{\min\{c_+, c_-\}}.$$

Then, for every $\ell \geq 1$,

$$S_{0,\ell}^{+} = P_{0,\ell}^{+} + \Gamma^2 C_{0,\ell}^{+} \geq 0,$$

because $C_{0,\ell}^{+} \geq c_+$ and $[-P_{0,\ell}^{+}]_+ \leq B_\varepsilon$. Similarly, for every $\ell \geq 2$,

$$S_{0,\ell}^{-} = P_{0,\ell}^{-} + \Gamma^2 C_{0,\ell}^{-} \geq 0,$$

because $C_{0,\ell}^{-} \geq c_-$ and $[-P_{0,\ell}^{-}]_+ \leq B_\varepsilon$. For $\ell = 1$, we have $S_{0,1}^{-} = 0$. Hence every interpolation inequality holds.

Gradient norm lower bound. For every $t \geq 1$, we have $\boldsymbol{g}_t = (\boldsymbol{g}_t^h, \Gamma g_t^c)$, so

$$\|\boldsymbol{g}_t\|_2^2 \geq \|\boldsymbol{g}_t^h\|_2^2 = G^2.$$

At $t = 0$, the auxiliary scalar coordinate gives

$$g_0^c = -\frac{\gamma_1}{\eta}.$$

Thus

$$\|\boldsymbol{g}_0\|_2^2 \geq \Gamma^2 \frac{\gamma_1^2}{\eta^2}.$$

For fixed ε, G^2 is bounded above on the compact interval $[a\varepsilon^2, b\varepsilon]$. Define

$$G_{\max,\varepsilon}^2 := \sup_{\eta \in [a\varepsilon^2, b\varepsilon]} G^2 < \infty.$$

Choose Γ so that

$$\Gamma^2 \geq \max\left\{ \frac{B_\varepsilon}{\min\{c_+, c_-\}}, \frac{G_{\max,\varepsilon}^2 b^2 \varepsilon^2}{\gamma_1^2} \right\}.$$

Then

$$\|\boldsymbol{g}_0\|_2^2 \geq G^2.$$

Therefore, for every T,

$$G_T^2 := \min_{0 \leq t < T} \|\boldsymbol{g}_t\|_2^2 \geq G^2.$$

Function values. We define f_t^h and f_t^c as (31) using $\{(\boldsymbol{x}_t^h, \boldsymbol{g}_t^h)\}$ and $\{(x_t^c, g_t^c)\}$, respectively. Then define

$$f_t := f_t^h + \Gamma^2 f_t^c.$$

This agrees with (31) for the full tuple because

$$\boldsymbol{x}_t = (\boldsymbol{x}_t^h, \Gamma x_t^c), \quad \boldsymbol{g}_t = (\boldsymbol{g}_t^h, \Gamma g_t^c).$$

For $t \geq 1$, we have

$$f_{t+1}^h - f_t^h = M, \quad \|\boldsymbol{g}_t^h\|_2^2 = G^2.$$

Therefore, for $t \geq 1$,

$$-f_t^h + \frac{1}{2}\|\boldsymbol{g}_t^h\|_2^2 = \left(-f_1^h + M + \frac{1}{2}G^2\right) + (-M)t.$$

Define

$$C_{h,\varepsilon} := \sup_{\eta \in [a\varepsilon^2, b\varepsilon]} \max\left\{ -f_0^h + \frac{1}{2}\|\boldsymbol{g}_0^h\|_2^2, -f_1^h + M + \frac{1}{2}G^2 \right\}.$$

This is finite because all terms are continuous functions of η, and the interval $[a\varepsilon^2, b\varepsilon]$ is compact. Hence, for every $t \geq 0$,

$$-f_t^h + \frac{1}{2}\|\boldsymbol{g}_t^h\|_2^2 \leq C_{h,\varepsilon} + (-M)t.$$

Now consider the unscaled auxiliary scalar coordinate. For $t \geq N$,

$$\gamma_{t+1} = \beta\gamma_t, \quad g_t^c = 0.$$

Also $\Delta g_t^c = 0$, and hence

$$Q_{t,t+1}^c = -\frac{1}{4}\gamma_{t+1}^2.$$

Thus, for $t \geq N$,

$$f_{t+1}^c - f_t^c = \frac{1}{4}\gamma_{t+1}^2 \geq 0.$$

Consequently, for $t \geq N$,

$$-f_t^c \leq -f_N^c.$$

For $0 \leq t \leq N$, all quantities are finite and uniformly bounded over $\eta \in [a\varepsilon^2, b\varepsilon]$. Therefore there exists

$$C_{c,\varepsilon}^{(0)} < \infty$$

such that

$$-f_t^c + \frac{1}{2}|g_t^c|^2 \leq C_{c,\varepsilon}^{(0)}$$

for all $t \geq 0$. Define

$$C_{c,\varepsilon} := \Gamma^2 C_{c,\varepsilon}^{(0)}.$$

Then,

$$-\Gamma^2 f_t^c + \frac{1}{2}\Gamma^2|g_t^c|^2 \leq C_{c,\varepsilon}$$

for all $t \geq 0$.

Thus, we get

$$-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2 \leq C_{h,\varepsilon} + C_{c,\varepsilon} + (-M)t.$$

Consequently,

$$\Delta_T := \max_{0 \leq t < T}\left(-f_t + \frac{1}{2}\|\boldsymbol{g}_t\|_2^2\right) \leq C_{\Delta,\varepsilon} + (-M)T,$$

where

$$C_{\Delta,\varepsilon} := C_{h,\varepsilon} + C_{c,\varepsilon} < \infty.$$

Proof for Regime II. **Proof** [Proof of Proposition 34] Fix $0 < \varepsilon \leq \varepsilon_{\max}$, $T \geq 1$, and $\eta \in [a\varepsilon^2, b\varepsilon]$. Construct the tuple above, choosing Γ so that

$$\Gamma^2 \geq \max\left\{\frac{B_\varepsilon}{\min\{c_+, c_-\}}, \frac{G_{\max,\varepsilon}^2 b^2 \varepsilon^2}{\gamma_1^2}\right\}.$$

The tuple satisfies the interpolation condition.

Then,

$$\sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{G^2}{2(C_{\Delta,\varepsilon} + (-M)T)}.$$

Multiplying by εT, we obtain

$$\varepsilon T \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{G^2}{-2M/\varepsilon} \cdot \frac{1}{1 + C_{\Delta,\varepsilon}/((-M)T)}.$$

By Lemma 36,

$$-M \geq K\varepsilon\tilde{\delta} \geq \varepsilon\tilde{\delta}.$$

Since $C_{\Delta,\varepsilon} < \infty$ for fixed ε,

$$\frac{C_{\Delta,\varepsilon}}{(-M)T} \le \frac{C_{\Delta,\varepsilon}}{\varepsilon\tilde{\delta}T} \to 0$$

as $T \to \infty$, uniformly over $\eta \in [a\varepsilon^2, b\varepsilon]$. Again by Lemma 36,

$$\frac{G^2}{-2M/\varepsilon} \ge c_0.$$

Taking the infimum over $a\varepsilon^2 \le \eta \le b\varepsilon$ and then taking $\liminf_{T\to\infty}$ gives

$$\liminf_{T\to\infty} \varepsilon T \inf_{a\varepsilon^2\le\eta\le b\varepsilon} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge c_0.$$

■

F.8. Final Admissibility Check

We choose

$$\varepsilon_{\max} = \frac{1}{8}, \quad a = 4, \quad b = \kappa = 5, \quad \tau = \frac{5}{3}, \quad c_0 = \frac{1}{8}.$$

With this choice,

$$K_{\max} = \frac{16}{15}, \quad \Theta^2 = K_{\max} b\varepsilon_{\max} = \frac{2}{3}, \quad S^2 = \frac{K_{\max}}{a} = \frac{4}{15}.$$

We take

$$h = \frac{17}{18}, \quad r = \frac{8}{9}, \quad d = \frac{1}{3}, \quad \rho = \frac{1}{6}.$$

We first verify the envelope conditions in (42). For $0 < u \le \Theta \le 1$, the inequalities

$$\frac{\sin u}{u} \ge 1 - \frac{u^2}{6}, \quad \frac{2(1-\cos u)}{u^2} \ge 1 - \frac{u^2}{12}$$

give

$$H(u) \ge 1 - \frac{\Theta^2}{12} = \frac{17}{18} = h, \quad R(u) \ge 1 - \frac{\Theta^2}{6} = \frac{8}{9} = r.$$

Moreover, since $0 \le H(u) \le 1$,

$$1 - H(u)^2 = (1 - H(u))(1 + H(u)) \le 2(1 - H(u)) \le \frac{u^2}{6} = \rho u^2.$$

Finally, for every integer $n \ge 1$,

$$0 \le n\sin u - \sin(nu) \le nu - \sin(nu) \le \frac{n^2u^2}{3} \le \frac{u^2 n(n+1)}{3} = du^2 n(n+1).$$

Thus (42) holds.

Next, we verify (43). The conditions (43a) and (43b) follow from

$$\Theta^2 = \frac{2}{3} < 1, \quad S^2 = \frac{4}{15} < 1.$$

For (43c), we have

$$\frac{1-S^2}{2} - 2Sd = \frac{11}{30} - \frac{4}{3\sqrt{15}} > 0.$$

For (43d),

$$S^2 + \rho\Theta^2 = \frac{4}{15} + \frac{1}{6}\cdot\frac{2}{3} = \frac{17}{45} < 1.$$

For (43e), we get

$$\frac{b}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) = \frac{5}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) < 1 + r^2 = \frac{145}{81}.$$

For (43f),

$$\frac{1}{a}\left(\frac{1}{(1+\sqrt{1-\varepsilon_{\max}})^2} + \frac{\pi^2}{\tau^2}\right) = \frac{1}{4}\left(\frac{1}{(1+\sqrt{7/8})^2} + \frac{\pi^2}{(5/3)^2}\right) < 1.$$

For (43g),

$$\frac{h^2}{4K_{\max}} = \frac{(17/18)^2}{4\cdot 16/15} > \frac{1}{8} = c_0.$$

Finally,

$$\left\lceil \frac{\tau}{\varepsilon_{\max}} \right\rceil = \left\lceil \frac{40}{3} \right\rceil \geq 2.$$

Therefore, all assumptions of Proposition 34 are satisfied.

We now verify the assumptions of Regime III with $\kappa = b = 5$. The interval condition (38) holds because

$$\kappa\varepsilon \leq 5\varepsilon_{\max} = \frac{5}{8} < 1 + \sqrt{3}.$$

For (39), set $N = \lceil \tau/\varepsilon \rceil$. Since

$$N \leq \frac{\tau}{\varepsilon} + 1,$$

it is enough to show

$$\frac{\tau}{\varepsilon} + 1 < \frac{\sqrt{1-\varepsilon}}{1-\sqrt{1-\varepsilon}}.$$

Using

$$\frac{\sqrt{1-\varepsilon}}{1-\sqrt{1-\varepsilon}} = \frac{\sqrt{1-\varepsilon}+1-\varepsilon}{\varepsilon},$$

the desired inequality follows from

$$\tau + 2\varepsilon - 1 < \sqrt{1-\varepsilon}.$$

For $0 < \varepsilon \leq \frac{1}{8}$, the left-hand side is at most

$$\frac{5}{3} + \frac{1}{4} - 1 = \frac{11}{12},$$

while the right-hand side is at least $\sqrt{\frac{7}{8}}$. Since $\frac{11}{12} < \sqrt{\frac{7}{8}}$, we obtain (39). Finally, (40) holds because

$$\frac{\varepsilon(1+\pi^2/\tau^2)}{\kappa} \leq \frac{1}{8}\cdot\frac{1+9\pi^2/25}{5} < 1.$$

Therefore, all assumptions of Proposition 29 are satisfied.

The four regimes cover all step sizes $\eta > 0$:

$$0 < \eta \le a\varepsilon^2, \quad a\varepsilon^2 \le \eta \le b\varepsilon, \quad b\varepsilon \le \eta \le 1 + \sqrt{3}, \quad \eta \ge 1 + \sqrt{3}.$$

The first two intervals are ordered because $a\varepsilon \le a\varepsilon_{\max} = 1/2 < 5 = b$. By Proposition 27, Regime I gives the constant

$$\frac{1}{2a} = \frac{1}{8}.$$

By Proposition 34, Regime II gives the constant $c_0 = 1/8$. By Propositions 28 and 29, Regimes III and IV give divergent εT-scaled lower bounds. Combining the four regimes yields, for every $0 < \varepsilon \le 1/8$,

$$\liminf_{T\to\infty} \varepsilon T \inf_{\eta>0} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{1}{8}.$$

Equivalently, the bound holds for every $\beta \ge 7/8$.

Appendix G. Proof of Theorem 3

G.1. Proof Sketch

The proof uses two lower bounds on translated quadratic functions. By translation and coordinate separability, it is enough to analyze the one-dimensional Signum iteration on $f(x) = \frac{L}{2}x^2$ and then repeat the same construction over d coordinates.

The first bound compares Signum with SignGD. After perturbing the initialization if necessary, the momentum variable does not hit zero during the first T iterations. Hence the Signum iterates move on a lattice with step size η. Among all such lattice paths, the SignGD path minimizes the cumulative absolute gradient on a shifted quadratic. This gives

$$\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \Lambda_1(\eta),$$

where $\Lambda_1(\eta)$ is the SignGD lower bound.

The second bound uses a deterministic cycle of the Signum iteration. For an odd integer n, we initialize the scalar quadratic at $x_0 = \eta n/2$. If n is at most of order $(1-\beta)^{-1/2}$ and also satisfies the initial gap constraint, the signs of the momentum variable make the iterates move from x_0 to $-x_0$ and then back to x_0. The condition ensuring this cycle is

$$\frac{1-\beta^n}{(1-\beta)(1+\beta^n)} - \frac{n}{2} > 0.$$

A lower bound on the average absolute value along this cycle gives

$$\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \Lambda_2(\eta),$$

where $\Lambda_2(\eta)$ is the cycle-based lower bound.

Combining the two estimates gives, for every step size,

$$\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \max\{\Lambda_1(\eta), \Lambda_2(\eta)\} \geq \frac{3}{4}\Lambda_1(\eta) + \frac{1}{4}\Lambda_2(\eta).$$

The remaining step is an explicit minimization over η. The weighted bound is decreasing for very small η and increasing after

$$\eta' = \frac{6}{7}\sqrt{\frac{2\Delta}{Ld}}\sqrt{1-\beta}.$$

Thus the minimum is attained in the middle interval. Evaluating the scalar minimum gives

$$\inf_{\eta>0}\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \sqrt{\frac{3\Delta Ld}{2T}\left(\frac{9}{16} + \frac{35}{128\sqrt{1-\beta}} - \frac{3}{4T}\right)}.$$

Since $T \geq 20$, this implies

$$\inf_{\eta>0}\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \sqrt{\frac{3\Delta Ld}{2T}\left(\frac{21}{40} + \frac{35}{128\sqrt{1-\beta}}\right)}.$$

This proves Theorem 3.

G.2. Useful Lemmas

Before proceeding, we provide some useful lemmas.

Lemma 37 *Let K be a positive integer such that $K \equiv 2 \pmod 4$ and $x_0 = \eta\frac{K}{4}$. Define*

$$n = \frac{K}{2}, \quad \delta = \frac{1-\beta^n}{(1-\beta)(1+\beta^n)} - \frac{n}{2}.$$

Let $\{x_k\}_{k=0}^{T-1}$ be generated by Signum, with $f(x) = \frac{L}{2}x^2$, $\sigma = 0$ and $T > K$. If $\delta > 0$, for any $k \geq 0$,

$$x_k = \begin{cases} x_0 - \eta l & \text{if } 0 \leq k \leq n, \\ x_0 - \eta(2n-l) & \text{if } n \leq k \leq 2n, \end{cases}$$

where $l \in \{0, \dots, K-1\}$ and $k \equiv l \pmod K$. Therefore, $\{x_k\}$ is K-periodic.

Proof We first define the following operator $F : \mathbb{R}^2 \to \mathbb{R}^2$:

$$(x, m) \mapsto (x - \eta\,\mathrm{sign}(\beta m + Lx), \beta m + Lx).$$

This operator represents one step of the Signum iteration on the quadratic function $\frac{L}{2}x^2$.

Let $\{\tilde{x}_k\}$ be the sequence such that if $k \equiv l \pmod K$ and $l \in \{0, \dots, n-1\}$, then

$$\tilde{x}_k = \begin{cases} x_0 - \eta l & \text{if } 0 \leq l \leq n, \\ x_0 - \eta(2n-l) & \text{if } n \leq l \leq 2n. \end{cases} \tag{51}$$

We will show that if $(x_k, m_k) = F^k(x_0, m_0)$ with $m_0 \in I := \left(-\eta L\delta\frac{\beta^{-n}+1}{1-\beta},\, \eta L\delta\frac{\beta^{-2n}-1}{1-\beta}\right)$, then $x_k = \tilde{x}_k$ for all $k \geq 0$ and $m_K \in I$.

We first show that $m_1, \dots, m_n > 0$ if $m_0 \in I$ by induction. For $k \in \{0, \dots, n-1\}$, suppose that $m_1, \dots, m_k > 0$. If $k = 0$, this is vacuously true. Assuming the inductive hypothesis, we need to show that $m_{k+1} > 0$.

Unrolling the recursion, we have

$$\begin{aligned} m_{k+1} &= \beta^{k+1}m_0 + \frac{\eta L}{1-\beta}\left(\frac{n}{2}(1-\beta^{k+1}) - (k+1) + \frac{1-\beta^{k+1}}{1-\beta}\right) \\ &= \beta^{k+1}m_0 + \frac{\eta L}{1-\beta}\delta(1+\beta^{k+1}), \end{aligned}$$

because $x_0 = \eta\frac{n}{2}$. To show that $m_{k+1} > 0$, we have to prove that $m_0 > L_{k+1}$, where

$$L_k := -\eta L\frac{\beta^{-k}}{1-\beta}\left(\frac{n}{2}(1-\beta^k) - k + \frac{1-\beta^k}{1-\beta}\right).$$

Note that $L_n = -\eta(\beta^{-n}+1)\delta = \inf I$, so $m_0 > L_n$. We will show that $\max_{1\leq k\leq n} L_k = L_n$. We have

$$L_{k+1} - L_k = -\eta L\beta^{-(k+1)}\left(k - \frac{n}{2}\right),$$

so $L_{k+1} - L_k \leq 0$ for $k < n/2$ and $L_{k+1} - L_k \geq 0$ for $k > n/2$ (because n is odd). This implies that the maximum of L_k occurs only at $k = 1$ or $k = n$. We have

$$\begin{aligned}
(1-\beta)\frac{L_n - L_1}{\eta L} &= \beta^{-1}\frac{n}{2} - \beta^{-n}\left(\frac{1-\beta^n}{1-\beta} - \frac{n}{2}\right) \\
&= \beta^{-n}\left(\beta^{n-1}\frac{n}{2} - \left(\sum_{i=0}^{n-1}\beta^i - \frac{n}{2}\right)\right) \\
&= \beta^{-n}\left((1+\beta^{n-1})\frac{n}{2} - \frac{1}{2}\sum_{i=0}^{n-1}(\beta^i + \beta^{n-1-i})\right) \\
&= \frac{\beta^{-n}}{2}\left((1+\beta^{n-1})n - \sum_{i=0}^{n-1}(\beta^i + \beta^{n-1-i})\right) \\
&\geq 0,
\end{aligned}$$

where the last inequality holds because $1 + \beta^{n-1} - (\beta^i + \beta^{n-1-i}) = (1-\beta^i)(1-\beta^{n-1-i}) \geq 0$. Therefore, $\max_{1\leq k\leq n} L_k = L_n$. Consequently, we have $m_1, \ldots, m_n > 0$ and $x_t = x_0 - \eta t$ for $t = 0, \ldots, n$.

Now, we have $F^n(x_0, m_0) = (-x_0, \beta^n m_0 + C)$, where $C := \eta(1+\beta^n)\delta$. Moreover, $F^k(-x_0, m_n) = -F^k(x_0, -m_n)$ for all $k = 1, \ldots, n$. Thus, to ensure $m_{n+1}, \ldots, m_{2n} > 0$, we need $-m_n > -\eta\delta\frac{1+\beta^{-n}}{1-\beta}$, following the same argument for the first half of the cycle. Since $-m_n = -(\beta^n m_0 + C)$, it suffices to have $m_0 < \eta\delta\frac{\beta^{-2n}-1}{1-\beta}$, which holds because $m_0 \in I$.

Since $m_K = m_{2n} = -\eta\delta\frac{1+\beta^n}{1-\beta} > -\eta\delta\frac{1+\beta^{-n}}{1-\beta}$ and $m_K < 0 < \eta\delta\frac{\beta^{-2n}-1}{1-\beta}$, so $m_K \in I$. This concludes the proof.

■

Lemma 38 *Let $\{\tilde{x}_t\}$ be the K-cycle sequence, as in (51). Suppose $n = K/2$ is odd. Then, for all $T \geq 1$,*

$$\sum_{t=0}^{T-1}|\tilde{x}_t| \geq \eta\left(\frac{1}{4}\left(n + \frac{1}{n}\right)T + \left(-\frac{n^2}{32} + \frac{n}{8} - \frac{1}{32} - \frac{3}{16n}\right)\right).$$

Moreover, assuming $T \geq 3n$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}|\tilde{x}_t| \geq \eta\left(\frac{15}{64}n + \frac{3}{32}\right). \tag{52}$$

Proof Let $z_t = \tilde{x}_t/\eta$, $n = 2m+1$ and $a = \frac{n^2+1}{4n}$. We define D_n as

$$D_0 = 0, \quad D_k = \sum_{t=0}^{k-1}(|z_t| - a) \quad \forall k \geq 1.$$

Since $\{|z_t|\}$ is periodic with period n, for any non-negative integer r, we have

$$D_{r+n} - D_r = \sum_{t=r}^{r+n-1}|z_t| - \frac{n^2+1}{4} = 0.$$

Thus, $\{D_k\}$ is also n-periodic and we only need to show that

$$\min_{0\le k<n} D_k \ge -\frac{n^2}{32}+\frac{n}{8}-\frac{1}{32}-\frac{3}{16n}.$$

Let $\Delta_k := D_{k+1}-D_k = |z_k|-a$. Since

$$|z_t| = \begin{cases} m+\frac{1}{2}-t, & \text{if } 0\le t\le m,\\ t-m-\frac{1}{2}, & \text{if } m+1\le t\le 2m,\end{cases}$$

we have

$$\Delta_k = \begin{cases} m+\frac{1}{2}-k-a, & \text{if } 0\le k\le m,\\ k-m-\frac{1}{2}-a, & \text{if } m+1\le k\le 2m.\end{cases}$$

Note that $\Delta_{k+1}-\Delta_k<0$ if $0\le k\le m$, and $\Delta_{k+1}-\Delta_k>0$ if $m+1\le k\le 2m$. Moreover, if $n\ge 5$, $\Delta_m=\frac{1}{2}-a<0$ and $\Delta_{n-1}=\frac{n}{2}-1-a>0$. Thus, if we define $k_\star := \min\{k:\Delta_k\ge 0\}$, then $m+1\le k_\star\le n-1$ and $\Delta_{k_\star-1}<0\le\Delta_{k_\star}$. Since $\Delta_{k-1}<0$ implies $D_k<D_{k-1}$ and $\Delta_k\ge 0$ implies $D_{k+1}\ge D_k$, $k_\star\in\arg\min_{0\le k<n}\{D_k\}$.

Using the fact that $\Delta_k\ge 0 \iff k\ge m+1/2+a$ for $m+1\le k\le 2m$, we have

$$k_\star = \left\lceil \frac{3n}{4}+\frac{1}{4n}\right\rceil.$$

Now, we first assume that $n\equiv 1 \pmod 4$. Then, for some $q\in\mathbb{Z}$, $n=4q+1$, $m=2q$ and $k_\star=3q+1$. We can compute

$$\begin{aligned} D_{k_\star} &= \sum_{t=0}^{k_\star-1}|z_t|-ak_\star\\ &= \sum_{t=0}^{2q}\left(2q+\frac{1}{2}-t\right)+\sum_{t=2q+1}^{3q}\left(t-2q-\frac{1}{2}\right)-\frac{n^2+1}{4n}k_\star\\ &= -\frac{n^3-4n^2+n+2}{32n}.\end{aligned}$$

Next, assume that $n\equiv 3 \pmod 4$. Then, for some $q\in\mathbb{Z}$, $n=4q+3$ and $k_\star=3q+3$. Similar to the previous case, we obtain

$$\begin{aligned} D_{k_\star} &= \sum_{t=0}^{k_\star-1}|z_t|-ak_\star\\ &= \sum_{t=0}^{2q+1}\left(2q+\frac{3}{2}-t\right)+\sum_{t=2q+2}^{3q+3}\left(t-2q-\frac{3}{2}\right)-\frac{n^2+1}{4n}k_\star\\ &= -\frac{n^3-4n^2+n+6}{32n}.\end{aligned}$$

Finally, if $n=1$, then $D_0=D_1=0$, so $\min_{0\le k<1}D_k=0$. If $n=3$, then $D_0=0$, $D_1=2/3$ and $D_2=1/3$, so $\min_{0\le k<1}D_k=0$. Combining those above, we have

$$\min_{0\le k<n} D_k \ge -\frac{n^3-4n^2+n+6}{32n} = -\frac{n^2}{32}+\frac{n}{8}-\frac{1}{32}-\frac{3}{16n},$$

when n is a positive odd number.

Next, assume that $T \geq 3n$. Let

$$R(n) := -\frac{n^2}{32} + \frac{n}{8} - \frac{1}{32} - \frac{3}{16n} = -\frac{(n-3)(n-2)(n+1)}{32n}.$$

Thus, for all $n \in \mathbb{N}$, $R(n) \leq 0$ and we have

$$\begin{aligned}\inf_{T \geq 2n} \frac{n^2+1}{4n} + \frac{R(n)}{T} &= \frac{n^2+1}{4n} + \frac{R(n)}{2n} \\ &= \frac{23n}{96} + \frac{1}{24} + \frac{23}{96n} - \frac{1}{16n^2}.\end{aligned}$$

Since

$$\begin{aligned}\frac{23n}{96} + \frac{1}{24} + \frac{23}{96n} - \frac{1}{16n^2} - \left(\frac{15}{64}n + \frac{3}{32}\right) &= \frac{n^3 - 10n^2 + 46n - 12}{192n^2} \\ &\geq 0\end{aligned}$$

for all $n \geq 1$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} |z_t| \geq \frac{15}{64}n + \frac{3}{32}.$$

■

Lemma 39 *If* $0.64 < \beta < 1$ *and* $0 < x < \frac{7}{3\sqrt{1-\beta}}$, *then* $\frac{1-\beta^x}{(1-\beta)(1+\beta^x)} - \frac{x}{2} > 0$.

Proof Let $a := -\log\beta > 0$, $X = \frac{7}{3\sqrt{1-\beta}}$ and $g(x) := \frac{1-\beta^x}{(1-\beta)(1+\beta^x)} - \frac{x}{2}$. Then, $\beta^x = e^{-ax}$ and

$$\frac{1-\beta^x}{1+\beta^x} = \tanh\left(\frac{ax}{2}\right).$$

Since $\tanh$ is concave on $(0,\infty)$, so is g. Moreover, $g(0) = 0$, so it suffices to show that $g(X) > 0$. Thus, it suffices to show that

$$\tanh\left(\frac{7}{6}\frac{a}{\sqrt{1-\beta}}\right) > \sqrt{1-\beta}.$$

Let $s = \sqrt{1-\beta} \in (0, 0.6)$. Then, $\beta = 1 - s^2$ and $a = -\log(1-s^2)$. Since $\tanh$ is increasing, we only need to show that

$$h(s) := -\frac{7}{6}\log(1-s^2) - s\tanh^{-1}\left(\frac{7}{6}s\right) > 0$$

We have

$$\begin{aligned}h'(s) &= \frac{42s}{49s^2-36} - \frac{7s}{3s^2-3} - \tanh^{-1}\left(\frac{7s}{6}\right) \\ h''(s) &= \frac{16807s^4 - 16961s^2 + 2520}{3(s^2-1)^2(49s^2-36)^2}s^2.\end{aligned}$$

Notice that the equation $16807s^4 - 16961s^2 + 2520 = 0$ has a unique solution in $(0, 0.6)$. Let $s_0 \approx 0.4255$ be the solution. Then, $h''(s)$ is negative on $(0, s_0)$ and positive on $(s_0, 0.6)$. Hence, $h'(s)$ is increasing from 0 to s_0, and then decreases. By evaluating h' at $s = 0.6$, we have $h'(0.6) \approx -0.052 < 0$. This implies that there exists a unique solution $s^\star$ of $h'(s) = 0$ in $(0, 0.6)$, because $h'(0) = 0$. Thus, it is sufficient to check $h(0.6) > 0$ to show that $h(s) > 0$ for all $s \in (0, 0.6)$. Since $h(0.6) \approx 0.00028 > 0$, we can conclude that $h(s) > 0$ for all $s \in (0, 0.6)$, as desired.

■

Lemma 40 *Consider the problem*

$$\min_{x_1,\ldots,x_{T-1}} \sum_{t=0}^{T-1} |x_t| \quad \textit{subject to } |x_t - x_{t+1}| = \eta \quad \forall\, t = 0, \ldots, T-2, \tag{53}$$

with $x_0/\eta \notin \mathbb{Z}$ and $\eta > 0$. If $x_0 > 0$, then the sequence $\{\tilde{x}_t\}$ defined as

$$\tilde{x}_t = \begin{cases} x_0 - \eta t & \text{if } t \le t_0, \\ x_0 - \eta(t_0 - 1) & \text{if } t > t_0 \text{ and } t - t_0 \text{ is odd}, \\ x_0 - \eta t_0 & \text{if } t > t_0 \text{ and } t - t_0 \text{ is even}, \end{cases} \tag{54}$$

where $t_0 = \left\lfloor \frac{x_0}{\eta} \right\rfloor + 1$, is a solution of (53). If $x_0/\eta \notin \mathbb{Z}$ and $x_0 < 0$, then $\{-\tilde{x}_t\}$ is a solution of (53).

Proof Without loss of generality, assume that $x_0 > 0$ (because $x_0/\eta \notin \mathbb{Z}$ implies $x_0 \neq 0$). The constraint $|x_t - x_{t+1}| = \eta$ implies that $x_{k+1} = x_k + s_k\eta$ where $s_k \in \{-1, 1\}$. Then, $x_t = x_0 + \eta \sum_{k=0}^{t-1} s_k$. Let k_t be the number of times $s_k = +1$ for $k < t$. Then, the number of times $s_k = -1$ is $t - k_t$, so $x_t = x_0 + (2k_t - t)\eta$, and $k_t \in \{0, 1, \ldots, t\}$.

Thus, $x_t \in S_t = \{x_0 + (2k - t)\eta \mid k \in \{0, 1, \ldots, t\}\}$ for all $t \ge 0$. This implies

$$\sum_{t=0}^{T-1} |x_t| \ge \sum_{t=0}^{T-1} \min_{z \in S_t} |z|.$$

If we show that $|\tilde{x}_t| = \min_{z \in S_t} |z|$ for every t, then $\{\tilde{x}_t\}$ is a global minimizer.

We now consider the following two cases:

Case 1. $t < t_0$**.** Here $x_0 - t\eta \ge x_0 - (t_0 - 1)\eta \ge 0$. The set S_t contains $x_0 - t\eta$ (at $k = 0$). Any other element in S_t is $(x_0 - t\eta) + 2k\eta$ for $k \ge 1$, which is strictly larger in absolute value. Thus, $\min_{z \in S_t} |z| = x_0 - t\eta = \tilde{x}_t$.

Case 2. $t \ge t_0$**.** Let $u = x_0 - (t_0 - 1)\eta$ and $v = x_0 - t_0\eta$.

- If $t - t_0$ is even, t and t_0 have the same parity. Thus $x_0 - t_0\eta \in S_t$. Since $v \in (-\eta, 0)$ and the distance between two different points in S_t is at least 2η, it is the unique element in S_t closest to zero.
- If $t - t_0$ is odd, t and $t_0 - 1$ have the same parity. Thus $x_0 - (t_0 - 1)\eta \in S_t$. Since $u \in (0, \eta)$, it is the unique element in S_t closest to zero, following the same argument used in the case when $t - t_0$ is even.

In both cases, $|\tilde{x}_t| = \min_{z \in S_t} |z|$. Since $\{\tilde{x}_t\}$ satisfies the constraint $|\tilde{x}_t - \tilde{x}_{t+1}| = \eta$ for all $t = 0, \dots, T-2$, it is feasible. Thus, it is a global minimizer of the problem.

If $x_0 < 0$, let $y_t = -x_t$. Then, $y_0/\eta \notin \mathbb{Z}$, $y_0 > 0$ and the problem (53) becomes

$$\min_{y_1,\dots,y_{T-1}} \sum_{t=0}^{T-1} |y_t| \quad \text{subject to } |y_t - y_{t+1}| = \eta \quad \forall\, t = 0, \dots, T-2.$$

By the previous part of the proof, $\{y_t\} = \{\tilde{x}_t\}$ defined as (54) is a solution of the problem. Thus, if $x_0 < 0$, then $\{-\tilde{x}_t\}$ is a solution of the problem. ■

Lemma 41 *For any $\eta > 0$, $\beta \in [0, 1)$ and $T \in \mathbb{N}$, consider $\{x_t\}_{t=0}^{T-1}$ and $\{m_t\}_{t=0}^{T-1}$ generated by Signum on the function $\frac{L}{2}x^2$. Then, the set*

$$Z := \{x_0 \in \mathbb{R} \mid m_t = 0 \quad \textit{for some } t \in \{1, \dots, T-1\}\}$$

is finite.

Proof For any sign sequence $s = (s_1, \dots, s_{t-1}) \in \{\pm 1\}^{t-1}$, define the following sequences

$$\begin{aligned} x_0^{(s)} = 0, \quad m_0^{(s)} = 0 \\ x_k^{(s)} = x_{k-1}^{(s)} - \eta s_k \quad (k = 1, \dots, t-1) \\ m_{k+1}^{(s)} = \beta m_k^{(s)} + L x_k^{(s)} \quad (k = 1, \dots, t). \end{aligned}$$

We denote by $m_t(x_0)$ the value of m_t at time t obtained by running the recursion starting from that initial point x_0. Then, for each fixed s and t, we have

$$x_k^{(s)} = x_0 - \eta \sum_{i=1}^{k} s_i, \quad k = 0, \dots, t-1.$$

Solving the recursion, we have

$$m_t^{(s)} = \left(\sum_{i=0}^{t-1} \beta^{t-1-i} \right) L x_0 - \eta L \sum_{i=0}^{t-1} \beta^{t-1-i} \sum_{k=1}^{i} s_k.$$

For any fixed s and t, the solution of $m_t^{(s)}(x_0) = 0$ is a linear equation in x_0. Since $\sum_{i=0}^{t-1} \beta^{t-1-i} = \frac{1-\beta^t}{1-\beta} \neq 0$, the solution is unique:

$$x_0 = \eta \left(\sum_{i=0}^{t-1} \beta^{t-1-i} \right)^{-1} \sum_{i=0}^{t-1} \beta^{t-1-i} \sum_{k=1}^{i} s_k. \tag{55}$$

Now, suppose that $m_t(x_0) = 0$ for some $t \leq T$ for the first time. Then, $\operatorname{sign}(m_k(x_0)) \in \{\pm 1\}$ for $k = 1, \dots, t-1$, so the solution of $m_t(x_0) = 0$ is unique and can be written in the form of (55). Thus, Z is contained in

$$\bigcup_{t=1}^{T} \bigcup_{s \in \{\pm 1\}^{t-1}} \left\{ \eta \left(\sum_{i=0}^{t-1} \beta^{t-1-i} \right)^{-1} \sum_{i=0}^{t-1} \beta^{t-1-i} \sum_{k=1}^{i} s_k \right\},$$

which is a finite set. ■

Lemma 42 *Let $\{\boldsymbol{x}_t\}_{t\geq 0}$ be the iterates of Signum using step size η and momentum parameter $\beta \in [0,1)$, and let $\{\tilde{\boldsymbol{x}}_t\}_{t\geq 0}$ be the iterates of SignGD using the same step size η, where both are initialized at $\mathbf{0}$. Then, there exists a quadratic function f in $\mathcal{F}_L(\Delta)$ such that for any β,*

$$\tfrac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \geq \tfrac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\tilde{\boldsymbol{x}}_t)\|_1.$$

Moreover, we have

$$\sup_{f\in\mathcal{F}_L(\Delta)}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\tilde{\boldsymbol{x}}_t)\|_1 \geq \Lambda_1(\eta) := \begin{cases}\sqrt{2\Delta Ld} - \frac{Ld}{2}(T-1)\eta, & \text{if } \eta \leq \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1},\\ \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(1-\frac{1}{T}\right)\eta, & \text{if } \eta > \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}.\end{cases} \tag{56}$$

Proof Consider $\{x_t\}_{t=0}^{T-1}$ and $\{m_t\}_{t=0}^{T-1}$ generated by Signum on the function $\frac{L}{2}x^2$. Let

$$Z := \{x_0 \in \mathbb{R} \mid m_t = 0 \quad \text{for some } t \in \{1,\dots,T-1\}\}.$$

Then, by Lemma 41, Z is finite. Then, $\eta\mathbb{Z}\cup Z$ is countable, so $(x_0-\varepsilon, x_0+\varepsilon)\backslash(\eta\mathbb{Z}\cup Z)$ is non-empty for any $x_0\in\mathbb{R}$ and $\varepsilon>0$. Thus, there exists $x' \in (x_0-\varepsilon, x_0+\varepsilon)\backslash(\eta\mathbb{Z}\cup Z)$.

Now, for any $\varepsilon' \in (0,1)$, let $x_0 = \left(1-\frac{\varepsilon'}{2}\right)\sqrt{\frac{2\Delta}{Ld}}$ and $\varepsilon = \frac{\varepsilon'}{2}\sqrt{\frac{2\Delta}{Ld}}$. Then, there exists $x' \in \left((1-\varepsilon')\sqrt{\frac{2\Delta}{Ld}}, \sqrt{\frac{2\Delta}{Ld}}\right)$ such that $x' \notin (\eta\mathbb{Z}\cup Z)$.

Let $f(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x} + x'\mathbf{1}\|_2^2$. Then, $f\in\mathcal{F}_L(\Delta)$. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ and $\{\boldsymbol{m}_t\}_{t=0}^{T-1}$ be generated by Signum, and $\{\tilde{\boldsymbol{x}}_t\}_{t=0}^{T-1}$ be generated by SignGD. Then,

$$\boldsymbol{x}_t = x_t\mathbf{1}, \quad \boldsymbol{m}_t = m_t\mathbf{1}, \quad \tilde{\boldsymbol{x}}_t = \tilde{x}_t\mathbf{1},$$

where $\{x_t\}_{t=0}^{T-1}$ and $\{m_t\}_{t=0}^{T-1}$ are the output of Signum starting from $x=0$ on the function $\frac{L}{2}(x-x')^2$, and $\{\tilde{x}_t\}_{t=0}^{T-1}$ is the output of SignGD, with the same initialization and the function.

By Lemma 41, $m_t \neq 0$ for all $t\in\{1,\dots,T-1\}$, because $x'\notin Z$. This implies that $|x_{t+1}-x_t| = \eta$ for all $t\in\{1,\dots,T-1\}$. Moreover, $\tilde{x}_t \neq 0$ for all $t\in\{1,\dots,T-1\}$, because $x'/\eta \notin \mathbb{Z}$. Thus, we also have $|\tilde{x}_{t+1}-\tilde{x}_t| = \eta$ for all $t\in\{1,\dots,T-1\}$.

Note that $\{\tilde{x}_t + x'\}$ is the same as (54), with $x_0 = x'$. Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} L|x_t + x'| \geq \frac{1}{T}\sum_{t=0}^{T-1} L|\tilde{x}_t + x'|$$

by Lemma 40.

Using the fact that $\boldsymbol{x}_t = x_t\mathbf{1}$ and $\tilde{\boldsymbol{x}}_t = \tilde{x}_t\mathbf{1}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \geq \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\tilde{\boldsymbol{x}}_t)\|_1.$$

We first consider the case when $T=1$. In this case, the inequality

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \geq \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\tilde{\boldsymbol{x}}_t)\|_1$$

trivially holds, because $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0 = \mathbf{0}$. Furthermore,

$$\begin{aligned}\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\tilde{\boldsymbol{x}}_t)\|_1 &= \sup_{f\in\mathcal{F}_L(\Delta)} \|\nabla f(\mathbf{0})\|_1 \\ &\geq \sup_{\varepsilon'\in(0,1)} L\|x'\mathbf{1}\|_1 \\ &\geq \sqrt{2\Delta L d}.\end{aligned}$$

We now assume that $T > 1$. If $\eta \leq \frac{x'}{T-1}$, then

$$\begin{aligned}\sum_{t=0}^{T-1} |\tilde{x}_t + x'| &= \sum_{t=0}^{T-1} (x' - \eta t) \\ &= x'T - \frac{T(T-1)}{2}\eta\end{aligned}$$

and therefore

$$\frac{1}{T}\sum_{t=0}^{T-1} L|\tilde{x}_t + x'| = Lx' - \frac{\eta L}{2}(T-1).$$

Let $f(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x} + x'\mathbf{1}\|_2^2$. Then, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 = d \cdot \frac{1}{T}\sum_{t=0}^{T-1} L|\tilde{x}_t + x'|.$$

If $\eta > \frac{x'}{T-1}$,

$$\sum_{t=0}^{T-1} |\tilde{x}_t + x'| = \begin{cases} \displaystyle\sum_{t=0}^{t_0-1} (x' - \eta t) + \left(\frac{T - t_0 - 1}{2}\right)\eta + t_0\eta - x', & \text{if } T - t_0 \text{ is odd,} \\ \displaystyle\sum_{t=0}^{t_0-1} (x' - \eta t) + \left(\frac{T - t_0}{2}\right)\eta, & \text{if } T - t_0 \text{ is even.} \end{cases}$$

If $T - t_0$ is odd, we have

$$\begin{aligned}\sum_{t=0}^{T-1} |\tilde{x}_t + x'| &= x't_0 - \frac{t_0(t_0-1)}{2}\eta + \left(\frac{T - t_0 - 1}{2}\right)\eta + t_0\eta - x' \\ &= (x' + \eta)t_0 - \frac{\eta}{2}t_0^2 + \frac{T-1}{2}\eta - x' \\ &\geq \frac{x'^2}{2\eta} + \frac{T-1}{2}\eta.\end{aligned}$$

Here, the last inequality holds because $t_0 \in (x'/\eta,\, x'/\eta + 1]$ and $t \mapsto (x' + \eta)t - \frac{\eta}{2}t^2$ is increasing in that interval.

Similarly, if $T - t_0$ is even, we have

$$\begin{aligned}\sum_{t=0}^{T-1} |\tilde{x}_t + x'| &= x' t_0 - \frac{t_0(t_0-1)}{2}\eta + \left(\frac{T-t_0}{2}\right)\eta \\ &= x' t_0 - \frac{\eta}{2} t_0^2 + \frac{T}{2}\eta \\ &\geq \frac{x'^2}{2\eta} + \frac{T-1}{2}\eta.\end{aligned}$$

The last inequality holds because $t \mapsto x't - \frac{\eta}{2}t^2$ is decreasing in $(x'/\eta,\, x'/\eta + 1]$.

Consequently,

$$\frac{1}{T}\sum_{t=0}^{T-1} L|\tilde{x}_t + x'| \geq \frac{Lx'^2}{2\eta T} + \frac{L}{2}\left(1 - \frac{1}{T}\right)\eta.$$

Therefore, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \begin{cases} Lx' - \frac{\eta L}{2}(T-1), & \text{if } \eta \leq \frac{x'}{T-1}, \\ \frac{Lx'^2}{2\eta T} + \frac{L}{2}\left(1 - \frac{1}{T}\right)\eta, & \text{if } \eta > \frac{x'}{T-1}. \end{cases}$$

Thus,

$$\sup_{f \in \mathcal{Q}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 = \sup_{\varepsilon' \in (0,1)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \Lambda_1(\eta),$$

where

$$\Lambda_1(\eta) := \begin{cases} \sqrt{2L\Delta d} - \frac{Ld}{2}(T-1)\eta, & \text{if } \eta \leq \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}, \\ \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(1 - \frac{1}{T}\right)\eta, & \text{if } \eta > \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}. \end{cases}$$

■

G.3. Proof of Theorem 3

We now prove Theorem 3.

Proof Let $\mathcal{Q}_L(\Delta) \subset \mathcal{F}_L(\Delta)$ be a class of quadratic functions defined by

$$\mathcal{Q}_L(\Delta) := \left\{ \frac{L}{2}\|\boldsymbol{x} + \alpha \mathbf{1}\|_2^2 : \alpha^2 \leq \frac{2\Delta}{Ld} \right\}.$$

It suffices to establish a lower bound for

$$\sup_{f \in \mathcal{Q}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1.$$

Suppose $f(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x}\|_2^2$ and $g(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x} + \alpha\mathbf{1}\|_2^2$. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ be generated by Signum on f with $\boldsymbol{x}_0 = \alpha\mathbf{1}$ and $\{\tilde{\boldsymbol{x}}_t\}_{t=0}^{T-1}$ be generated by Signum on g with $\tilde{\boldsymbol{x}}_0 = 0$. Then,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 = \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g(\tilde{\boldsymbol{x}}_t)\|_1 = \frac{L}{T}\sum_{t=0}^{T-1}\|\boldsymbol{x}_t\|_1.$$

Thus, without loss of generality, we may assume that $f(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x}\|_2^2$ and we initialize at $\boldsymbol{x}_0 = \alpha\mathbf{1}$, where $\alpha^2 \leq \frac{2\Delta}{Ld}$.

Since $\nabla f(\boldsymbol{x}) = L\boldsymbol{x}$, the Signum updates are fully separable across coordinates. With the initialization $\boldsymbol{x}_0 = \alpha\mathbf{1}$, it follows that $\boldsymbol{x}_t = x_t\mathbf{1}$ for all $t \geq 0$, where $\{x_t\}$ is the one-dimensional Signum iterate applied to the function $\frac{L}{2}x^2$ with initialization $x_0 = \alpha$. Thus, we will consider the one-dimensional case, *i.e.*, $d = 1$.

Let n be the largest odd number such that

$$n \leq \min\left\{\frac{2}{\eta}\sqrt{\frac{2\Delta}{Ld}},\ \frac{7}{3\sqrt{1-\beta}}\right\}.$$

Then, if we use $x_0 = \eta\frac{n}{2}$, then

$$x_0^2 \leq \frac{2\Delta}{Ld},$$

which satisfies the initial condition $f(x_0) - f^\star \leq \Delta$.

Note that the largest odd number less than or equal to x is $2\left\lfloor\frac{x-1}{2}\right\rfloor + 1 > x - 2$. By the lemmas above, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}L|x_t| \geq \frac{L\eta}{2}\left(\frac{15}{32}\min\left\{\frac{2}{\eta}\sqrt{\frac{2\Delta}{Ld}},\ \frac{7}{3\sqrt{1-\beta}}\right\} - \frac{3}{4}\right).$$

For the function $\frac{L}{2}\|\boldsymbol{x}\|_2^2$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 &\geq d\cdot\frac{1}{T}\sum_{t=0}^{T-1}L|x_t| \\ &\geq \frac{L\eta d}{2}\left(\frac{15}{32}\min\left\{\frac{2}{\eta}\sqrt{\frac{2\Delta}{Ld}},\ \frac{7}{3\sqrt{1-\beta}}\right\} - \frac{3}{4}\right).\end{aligned}$$

Combining the above bounds, we finally obtain

$$\sup_{f\in\mathcal{F}_L(\Delta)}\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_1 \geq \max\left\{\Lambda_1(\eta),\,\Lambda_2(\eta)\right\},$$

where

$$\Lambda_2(\eta) := \begin{cases}\frac{Ld}{2}\left(\frac{35}{32\sqrt{1-\beta}} - \frac{3}{4}\right)\eta, & \text{if } \eta \leq \eta', \\ \frac{Ld}{2}\left(\frac{15}{16}\sqrt{\frac{2\Delta}{Ld}} - \frac{3}{4}\eta\right), & \text{if } \eta \geq \eta'\end{cases}$$

and $\eta' = \frac{6}{7}\sqrt{\frac{2\Delta}{Ld}}\sqrt{1-\beta}$. We also note that $\Lambda_2(\eta)$ is continuous.

Since $\max\{a,b\} \geq \frac{3}{4}a + \frac{1}{4}b$ and $\sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1} \leq \frac{6}{7}\sqrt{\frac{2\Delta}{Ld}}\sqrt{1-\beta}$, we obtain

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \\
&\geq \frac{3}{4}\Lambda_1(\eta) + \frac{1}{4}\Lambda_2(\eta) \\
&= \begin{cases} \frac{3\sqrt{2\Delta Ld}}{4} + \frac{Ld}{2}\left(-\frac{3T}{4} + \frac{9}{16} + \frac{35}{128\sqrt{1-\beta}}\right)\eta & \text{if } \eta \leq \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}, \\ \frac{3\Delta}{4T\eta} + \frac{Ld}{2}\left(\frac{35}{128\sqrt{1-\beta}} + \frac{9}{16} - \frac{3}{4T}\right)\eta & \text{if } \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1} \leq \eta \leq \eta', \\ \frac{3\Delta}{4T\eta} + \frac{Ld}{2}\left(\frac{15}{64}\sqrt{\frac{2\Delta}{Ld}} + \left(\frac{9}{16} - \frac{3}{4T}\right)\eta\right) & \text{if } \eta \geq \eta'. \end{cases}
\end{aligned}
$$

Notice that $\frac{3}{4}\Lambda_1(\eta) + \frac{1}{4}\Lambda_2(\eta)$ is decreasing on $\left(0, \sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}\right)$ and increasing on (η', ∞), implying that $\frac{3}{4}\Lambda_1(\eta) + \frac{1}{4}\Lambda_2(\eta)$ has a global minimizer in the interval $\left[\sqrt{\frac{2\Delta}{Ld}}\frac{1}{T-1}, \eta'\right]$. Moreover, the minimum of

$$\frac{3\Delta}{4T\eta} + \frac{Ld}{2}\left(\frac{35}{128\sqrt{1-\beta}} + \frac{9}{16} - \frac{3}{4T}\right)\eta$$

is attained at

$$\eta_\star = 4\sqrt{6}\sqrt{\frac{2\Delta}{Ld}}\frac{(1-\beta)^{1/4}}{\sqrt{(72\sqrt{1-\beta}+35)T - 96\sqrt{1-\beta}}}.$$

Meanwhile,

$$\left(\frac{\eta_\star}{\eta'}\right)^2 = \frac{392}{(216T-288)(1-\beta) + 105T\sqrt{1-\beta}} \leq 1$$

because $T \geq 20 + \frac{1}{1-\beta}$.

Therefore, the minimum of $\frac{3}{4}\Lambda_1(\eta) + \frac{1}{4}\Lambda_2(\eta)$ is attained at $\eta_\star$, and the minimum value is

$$\sqrt{\frac{3\Delta Ld}{2T}\left(\frac{9}{16} + \frac{35}{128\sqrt{1-\beta}} - \frac{3}{4T}\right)} \geq \sqrt{\frac{3\Delta Ld}{2T}\left(\frac{21}{40} + \frac{35}{128\sqrt{1-\beta}}\right)},$$

because $T \geq 20$.

■

G.4. Additional Illustration: Step Size Sensitivity of Signum

We note that the behavior of Signum is considerably more sensitive to the step size η than that of SHB. Even in the one-dimensional quadratic setting, small changes in η can result in drastically different trajectories. This phenomenon is illustrated in Figure 2. We fix the initialization at $x_0 = 1$ and run Signum on $f(x) = \frac{1}{2}x^2$ for a range of step sizes η, plotting the average gradient ℓ_1-norm as a function of η.

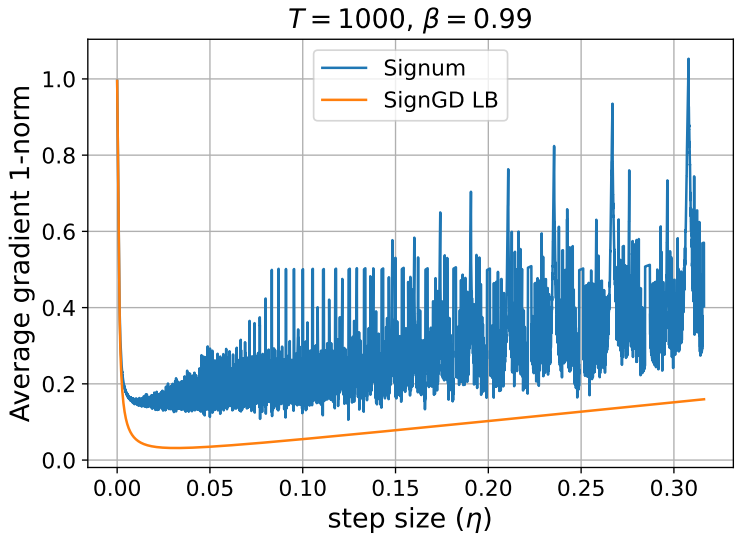

Figure 2: η vs average gradient ℓ_1-norm on $f(x) = \frac{1}{2}x^2$ with $x_0 = 1$, $T = 1000$, $\beta = 0.99$. The SignGD lower bound from (56) is shown for comparison.

While our analysis establishes a worst-case lower-bound separation that scales as $\Omega((1-\beta)^{-1/4})$, it does not match the $\mathcal{O}((1-\beta)^{-1/2})$ dependency suggested by the optimized upper bound in Proposition 9. Identifying the precise exponent of $1-\beta$ remains an open question, and given the chaotic behavior of Signum, we expect that establishing matching worst-case lower and upper bounds may be very challenging.

This sensitivity also makes stochastic extensions nontrivial. Our proof of Theorem 3 relies heavily on a deterministic limit-cycle behavior of the Signum iterates, and stochastic perturbations may disrupt the exact cycling structure used in the construction. Extending the lower bound argument to stochastic settings therefore appears to require additional ideas.

Appendix H. Muon Lower Bound via Diagonal Reduction to Signum

We prove the diagonal reduction used in the main text: on diagonal quadratic instances, the Muon dynamics preserve diagonality and coincide with Signum applied to the diagonal entries.

Proposition 43 *Let $f(\boldsymbol{X}) = \frac{L}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 : \mathbb{R}^{d\times d} \to \mathbb{R}$ and $\boldsymbol{X}_0 = \mathrm{diag}(x_1^{(0)}, \dots, x_d^{(0)}) \in \mathbb{R}^{d\times d}$ be a diagonal matrix with non-zero diagonal entries. Then, $\{\boldsymbol{X}_t\}_{t=0}^{T-1}$ and $\{\boldsymbol{M}_t\}_{t=0}^{T-1}$ from Muon are all diagonal. Moreover, $\boldsymbol{X}_{T-1} = \mathrm{diag}(\boldsymbol{x}_{T-1})$, where $\boldsymbol{x}_{T-1}$ is from the output of Signum on the function $\frac{L}{2}\|\boldsymbol{x}\|_2^2$, with initialization $(x_1^{(0)}, \dots, x_d^{(0)}) \in \mathbb{R}^d$.*

Proof We prove that both $\boldsymbol{X}_t$ and $\boldsymbol{M}_{t+1}$ are diagonal, for all $t \in \{0, \dots, T-1\}$, by induction. Trivially, $\boldsymbol{M}_0$ is diagonal because it is initialized as $\boldsymbol{0}_{d\times d}$.

Base case. If $t = 0$, $\boldsymbol{X}_0$ is diagonal by the initialization assumption. Since $\nabla f(\boldsymbol{X}) = L\boldsymbol{X}$, $\boldsymbol{G}_0$ is also diagonal, so $\boldsymbol{M}_1$ is diagonal.

Inductive step. Since $\boldsymbol{X}_t$ and $\boldsymbol{M}_t$ are diagonal by the inductive hypothesis, so is $\boldsymbol{G}_t = L\boldsymbol{X}_t$. Thus, $\boldsymbol{M}_{t+1}$ is diagonal. Let $\boldsymbol{M}_{t+1} = \mathrm{diag}(m_1^{(t+1)}, \dots, m_d^{(t+1)})$. Then, there exists an SVD of the form

$$\begin{aligned}\boldsymbol{U}_{t+1} &= \boldsymbol{I}_d \\ \boldsymbol{S}_{t+1} &= \mathrm{diag}\left(|m_1^{(t+1)}|, \dots, |m_d^{(t+1)}|\right) \\ \boldsymbol{V}_{t+1} &= \mathrm{diag}\left(\mathrm{sign}(m_1^{(t+1)}), \dots, \mathrm{sign}(m_d^{(t+1)})\right).\end{aligned}$$

Consequently,

$$\boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^\top = \mathrm{diag}\left(\mathrm{sign}(m_1^{(t+1)}), \dots, \mathrm{sign}(m_d^{(t+1)})\right).$$

Thus, $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta\boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^\top$ is also diagonal.

Now, let

$$\boldsymbol{M}_t = \mathrm{diag}(m_1^{(t)}, \dots, m_d^{(t)}), \quad \boldsymbol{X}_t = \mathrm{diag}(x_1^{(t)}, \dots, x_d^{(t)}).$$

Then, for any $i \in \{1, \dots, d\}$ and $t \in \{0, \dots, T-1\}$, we have the following update rule:

$$m_i^{(t+1)} = \beta m_i^{(t)} + Lx_i^{(t)}, \quad x_i^{(t+1)} = x_i^{(t)} - \eta\,\mathrm{sign}(m_i^{(t)}),$$

which is exactly the same as that of Signum, with initial point $(x_1^{(0)}, \dots, x_d^{(0)})$ on $\frac{L}{2}\|\boldsymbol{x}\|_2^2$. ■

This equivalence is invariant to translations of the objective, *i.e.*, replacing $f(\boldsymbol{X})$ with $f(\boldsymbol{X}+\boldsymbol{X}')$ for some diagonal $\boldsymbol{X}' \in \mathbb{R}^{d\times d}$.

For the purpose of proving lower bounds, Proposition 43 allows us to realize the Signum lower bound construction within the Muon dynamics. Under this reduction, identifying $\boldsymbol{X}_t = \mathrm{diag}(\boldsymbol{x}_t)$, we have $\|\nabla f(\boldsymbol{X}_t)\|_* = \|\nabla f(\boldsymbol{x}_t)\|_1$, because $\nabla f(\boldsymbol{X}_t)$ is diagonal and the nuclear norm of a diagonal matrix equals the sum of the absolute values of its diagonal entries. We formalize this implication in the following corollary.

Corollary 44 *Consider the iterates $\{\boldsymbol{X}_t\}_{t=0}^{T-1}$ generated by Muon, with $\sigma = 0$. Suppose $T \geq 20 + \frac{1}{1-\beta}$ and $\beta > 0.64$. Then,*

$$\inf_{\eta>0}\sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{X}_t)\|_* \geq \sqrt{\frac{3\Delta Ld}{2T}\left(\frac{21}{40} + \frac{35}{128\sqrt{1-\beta}}\right)}.$$

Appendix I. Performance Estimation Problem (PEP) Setup and Results

PEP Formulation. We evaluate algorithmic performance using the best-iterate gradient norm $\min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2$. In this section, we provide additional details on the worst-case analysis setup.

To this end, we employ the PEP framework [8], which enables the numerical computation of worst-case performance over a prescribed class of functions. For fixed values of the step size η, momentum parameter β, and the number of iterations T, the PEP computes

$$P(\eta, \beta, T) := \sup_{f\in\mathcal{F}_L(\Delta)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2,$$

where $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ is generated by the heavy-ball method with step size η and momentum parameter β.

To implement this, we adopt the necessary and sufficient interpolation conditions for L-smooth functions, given by Theorem 7 in Drori and Shamir [7].

All PEP instances are implemented using the PEPit toolbox [11], and the resulting semidefinite programs are solved using the MOSEK solver [1].

Normalization. We set the smoothness constant to $L = 1$ and the initial function value gap to $\Delta = 1/2$. This is without loss of generality. For any $g \in \mathcal{F}_{L_g}(\Delta_g)$, the rescaled function $f(\boldsymbol{x}) = \frac{1}{2\Delta_g} g\left(\sqrt{\frac{2\Delta_g}{L_g}}\,\boldsymbol{x}\right)$ is in $\mathcal{F}_1(1/2)$. The heavy-ball iterates $\boldsymbol{x}_t$ on g (with step size η) and $\boldsymbol{y}_t$ on f (with step size $L_g\eta$) satisfy $\boldsymbol{x}_t = \sqrt{\frac{2\Delta_g}{L_g}}\,\boldsymbol{y}_t$. Consequently, $\|\nabla g(\boldsymbol{x}_t)\|_2^2 = 2\Delta_g L_g \|\nabla f(\boldsymbol{y}_t)\|_2^2$, meaning that worst-case bounds for $\mathcal{F}_L(\Delta)$ are obtained by multiplying the results for $\mathcal{F}_1(1/2)$ by $2\Delta L$. Since the heavy-ball method is invariant under this transformation, the worst-case performance on $\mathcal{F}_1(1/2)$ characterizes the performance on the general class $\mathcal{F}_L(\Delta)$ up to the constant factor $2\Delta L$.

Worst-Case Values. We report PEP results for $T \in \{5, 10, 15, 20, 25, 30\}$. For each T, we visualize $P(\eta, \beta, T)$ on an (η, β) grid as heatmaps in Figure 3. We fix $\beta \in \{0.01, 0.02, \ldots, 0.99\}$ and minimize the PEP objective over the step size η to plot $\inf_{\eta>0} P(\eta, \beta, T)$ as a function of β in Figure 4. For numerical search near $\beta = 1$, where the range of convergent step sizes becomes narrow, we minimize over $\frac{\eta}{1-\beta}$ and then convert back to η. This shows the best achievable finite-horizon worst-case guarantee after T iterations for each momentum value. Across all T, the minimized worst-case gradient norm increases monotonically in β, with the best guarantee at $\beta = 0$.

Worst-Case Instances. Beyond the convergence rates, we examine representative worst-case instances: the functions and iterate sequences that approximately attain the supremum in the PEP objective. Note that the worst-case instance produced by the PEP can have dimension up to $(T + 2)$ [35]. We analyze specific worst-case instances for $T = 50$ and $\beta \in \{0.9, 0.95\}$. For each β, we visualize the instance obtained at the optimized step size $\eta^\star$ from the scalar minimization above, together with one representative larger step size, $\eta = 1.5$, to illustrate how the worst-case geometry changes away from the optimized step size. These instances are visualized in Figure 5 through heatmaps of the iterate matrix $\boldsymbol{X}$ and the gradient matrix $\boldsymbol{G}$, where the iterates are translated so that $\boldsymbol{x}_0 = \boldsymbol{0}$. Here, $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_0 & \cdots & \boldsymbol{x}_{T-1}\end{bmatrix}^\top$ and $\boldsymbol{G} = \begin{bmatrix}\nabla f(\boldsymbol{x}_0) & \cdots & \nabla f(\boldsymbol{x}_{T-1})\end{bmatrix}^\top$ denote the iterate and gradient sequences returned by the PEP worst-case instance. The heatmaps display the projections of $\boldsymbol{X}$ and $\boldsymbol{G}$ onto the leading 10 right singular directions of the joint matrix $\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{G}\end{bmatrix}$. When extracting instances for visualization, we use PEPit's dimension-reduction heuristic with `logdet10`, which often yields a lower-dimensional representative solution. This heuristic is used only for

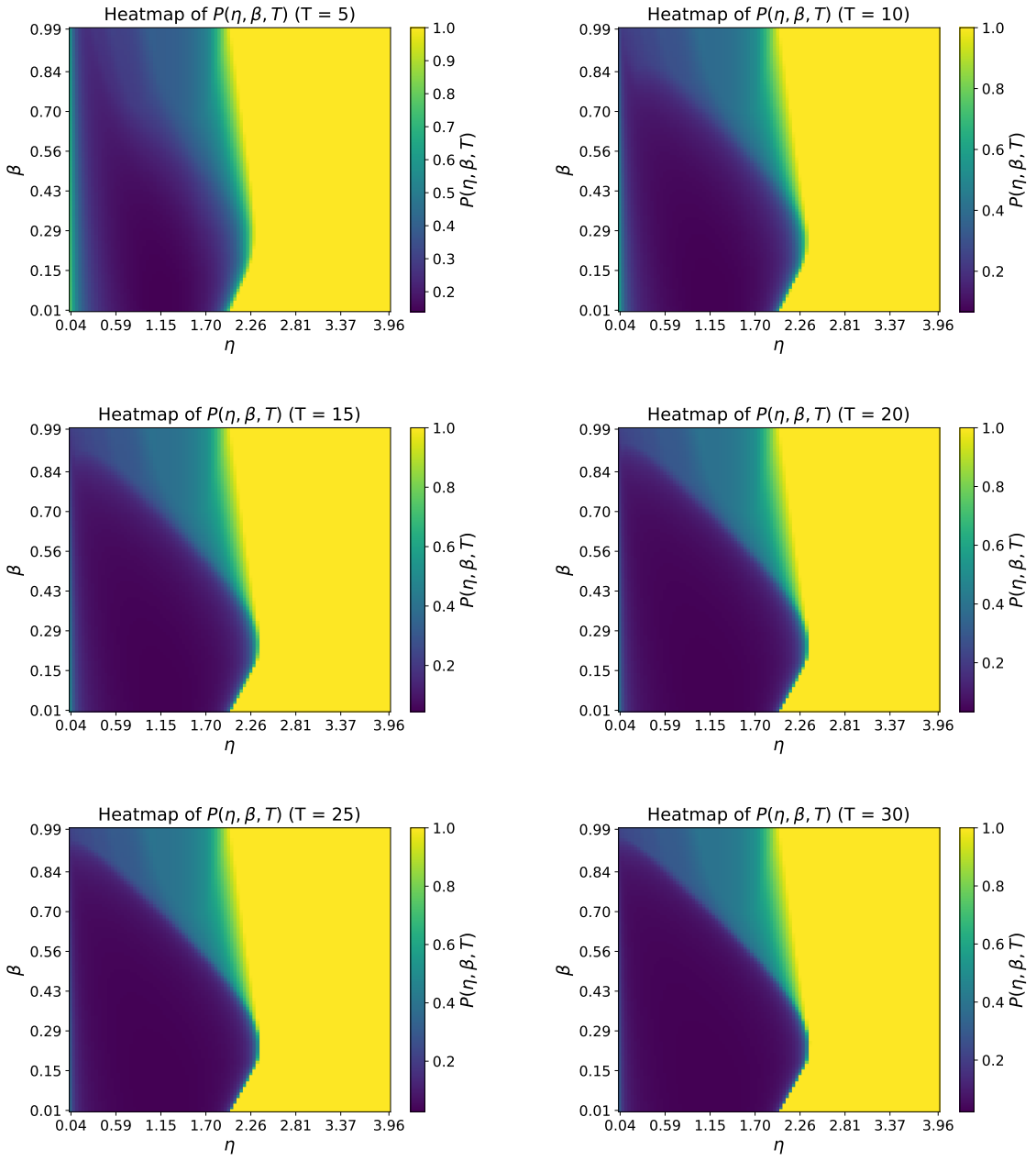

Figure 3: PEP results in terms of best-iterate squared gradient ℓ_2-norm ($T \in \{5, 10, 15, 20, 25, 30\}$).

visualization; the reported worst-case values are always taken from the original SDP objective values. Therefore, the plotted heatmaps should be interpreted as representative nearly worst-case instances, rather than unique canonical worst-case instances. In Figure 6, we additionally plot the corresponding iterate trajectories projected onto the leading 3 right singular directions of the same joint matrix, together with the projected negative gradients. The trajectory visually resembles the helical trajectories used in the proof of Theorem 2.

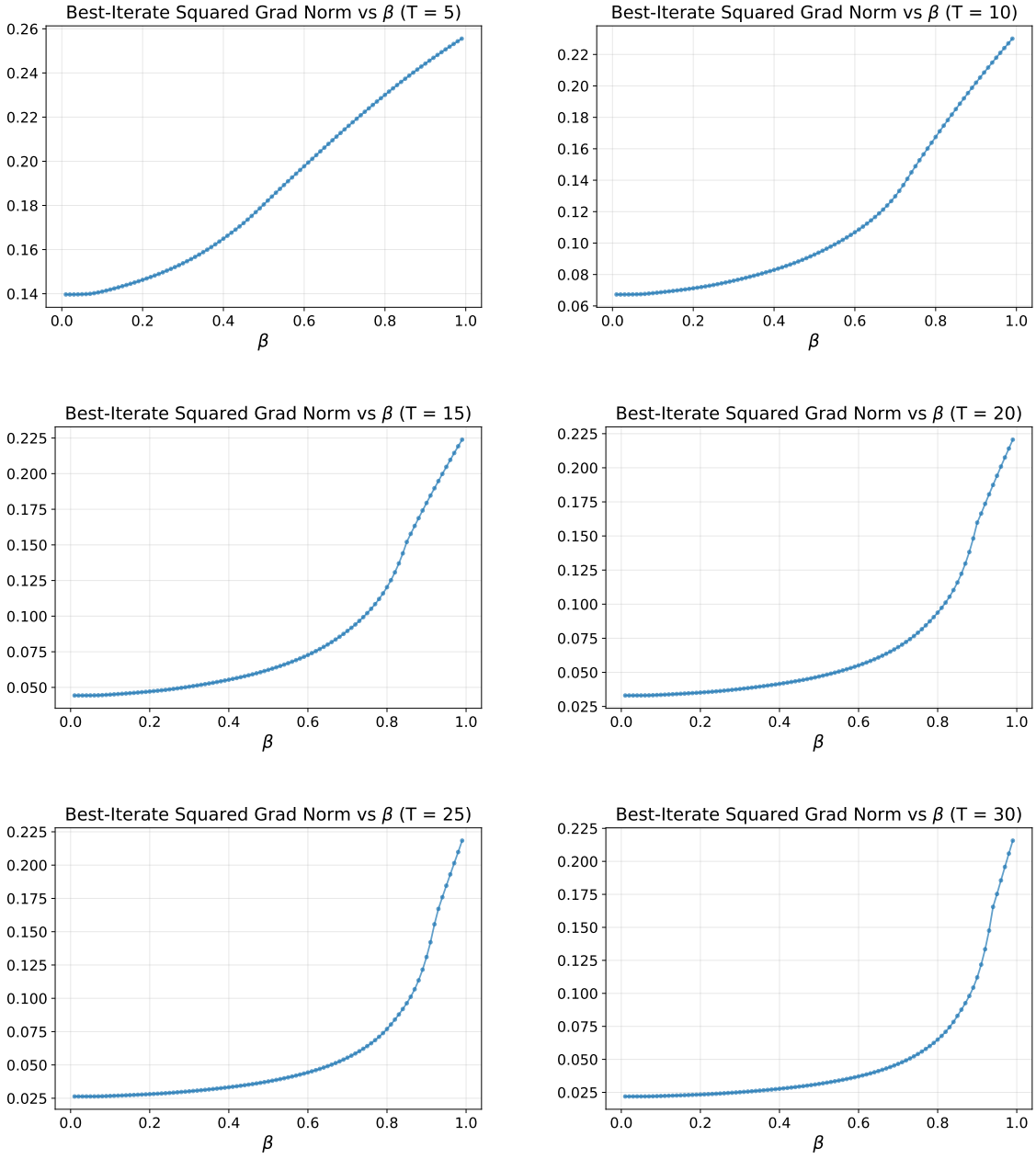

Figure 4: Minimized worst-case value $\inf_\eta P(\eta, \beta, T)$ as a function of the momentum parameter β for different values of T.

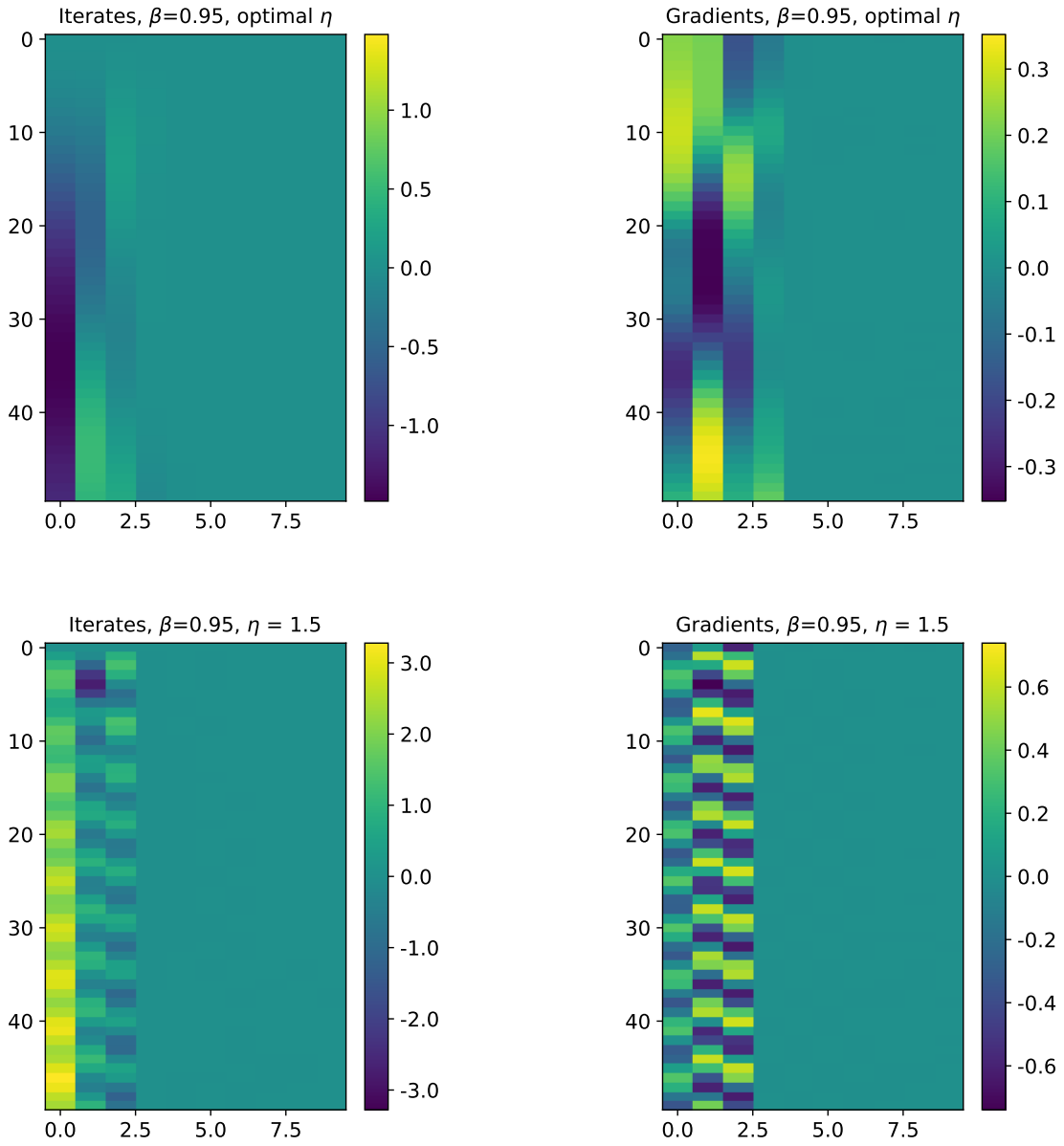

Figure 5: Heatmaps of representative PEP worst-case instances for $T = 50$ and $\beta = 0.95$. We compare the instance obtained at the optimized step size $\eta^\star \approx 0.02348$ with the instance obtained at the larger step size $\eta = 1.5$. The plots show the iterate matrix $\boldsymbol{X}$, translated so that $\boldsymbol{x}_0 = \mathbf{0}$, and the gradient matrix $\boldsymbol{G}$, after rotating both matrices onto the leading 10 right singular directions of the joint matrix $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{G} \end{bmatrix}$. Each row corresponds to one iteration $t = 0, \ldots, T-1$, with rows given by $\boldsymbol{x}_t$ and $\nabla f(\boldsymbol{x}_t)$, respectively.

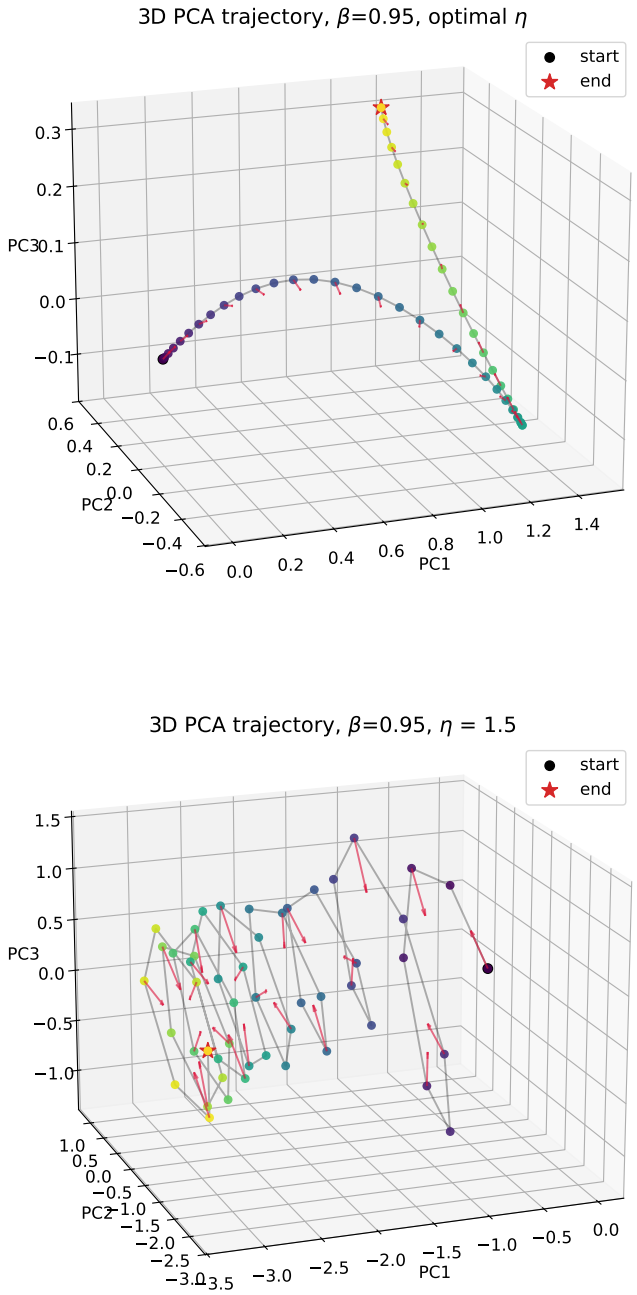

Figure 6: Trajectories of the worst-case iterates for $T = 50$ and $\beta = 0.95$, projected onto the first three principal components for $\eta = \eta^\star \approx 0.02348$ and $\eta = 1.5$, with arrows indicating the *negative* gradient directions at each iterate.

Appendix J. Details and Results of Deep Learning Experiments

To provide a practical perspective on the performance of momentum-based methods, we conducted experiments on standard image classification tasks. All models are trained from scratch using the cross-entropy loss for multi-class classification. All experiments are conducted over 5 independent runs with different random seeds, and we use NVIDIA RTX 4090 GPUs.

J.1. Setup

We evaluate the empirical performance of momentum-based methods against their non-momentum counterparts. All experiments were conducted using the ResNet-18 [13] architecture on the CIFAR-10 [19] dataset. The models were implemented in PyTorch [24]. The specific settings used in the experiments are as follows:

- **Algorithms.** We compared **(i)** SGD vs SHB, **(ii)** SignSGD vs Signum, and **(iii)** SpecGD vs Muon.
- **Momentum (β).** We compared $\beta \in \{0.0, 0.5, 0.9\}$.
- **Batch size.** We use batch sizes of $B \in \{256, 512\}$. Additionally, we conducted experiments using a full-batch setting on a subset of 5,000 training samples.
- **Learning rate (η).** For each configuration, we searched over the following grids to find the optimal η:
  - SGD / SHB ($B = 256$): $\{0.003, 0.01, 0.03, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$
  - SGD / SHB ($B = 512$): $\{0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$
  - SignSGD / Signum ($B = 256$): $\{0.001, 0.002, 0.003, 0.005, 0.007, 0.01\}$
  - SignSGD / Signum ($B = 512$): $\{0.0002, 0.0005, 0.001, 0.002, 0.005\}$
  - SpecGD / Muon ($B = 256$): $\{0.001, 0.003, 0.007, 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.1, 0.2, 0.3, 0.5, 0.8\}$
  - SpecGD / Muon ($B = 512$): $\{0.001, 0.002, 0.003, 0.005, 0.007, 0.01\}$
  - Full-batch SGD: $\{0.00001, 0.0003, 0.001, 0.003, 0.006, 0.01, 0.03, 0.06, 0.1, 0.15, 0.2, 0.25, 0.3\}$

We disable dropout and weight decay in all experiments. SGD/SHB and SignSGD/Signum were trained for 100 epochs, while SpecGD/Muon were trained for 40 epochs. For Muon, the spectral update was applied only to parameters that are at least two-dimensional (*e.g.*, convolutional kernels and internal weight matrices). One-dimensional parameters, such as biases, and the weights of the final fully-connected layer were updated using SGD with momentum.

J.2. Results

We recorded the training loss and the ℓ_2-norm of the gradient at every training step for all optimizers and configurations.

	Optimizer								
	SGD/SHB			SignSGD/Signum			SpecGD/Muon		
	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$
$B = 256$	0.3	0.3	0.1	0.002	0.002	0.002	0.05	0.01	0.003
$B = 512$	0.3	0.4	0.1	0.001	0.001	0.0005	0.01	0.01	0.003

Table 1: Optimal learning rate η with respect to training loss, for different optimizers, batch sizes, and momentum β.

	Optimizer								
	SGD/SHB			SignSGD/Signum			SpecGD/Muon		
	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$
$B = 256$	0.7	0.6	0.3	0.003	0.001	0.003	0.1	0.07	0.05
$B = 512$	0.5	0.4	0.2	0.005	0.01	0.005	0.01	0.01	0.01

Table 2: Optimal learning rate η with respect to average gradient norm, for different optimizers, batch sizes, and momentum β.

J.2.1. TRAINING LOSS

For each algorithm and configuration (β and B), we identified the optimal learning rate η by minimizing the tail-averaged training loss, calculated over the final 10% of the training steps. These selected η values are summarized in Table 1. Using these optimal learning rates, we compare the training progress across different momentum parameters in Figure 7. The training loss curves of Muon are less favorable to momentum in this setup, whereas the gradient-norm results give a different qualitative picture. We therefore do not view these experiments as evidence of a uniform empirical advantage of momentum across all optimizer classes or all performance measures.

J.2.2. AVERAGED GRADIENT NORM

We computed the gradient norms specific to each optimizer class: the ℓ_2-norm for SGD/SHB, the ℓ_1-norm for SignSGD/Signum. For SpecGD/Muon, we calculated the Frobenius norm of the gradient specifically for the parameters which are the targets of the Muon update. The learning rates that minimize the corresponding metric are listed in Table 2. The resulting comparisons are shown in Figure 8. For Muon, this gradient norm comparison gives a different qualitative picture from the training loss comparison in Figure 7, illustrating that the empirical comparison can depend on the performance measure.

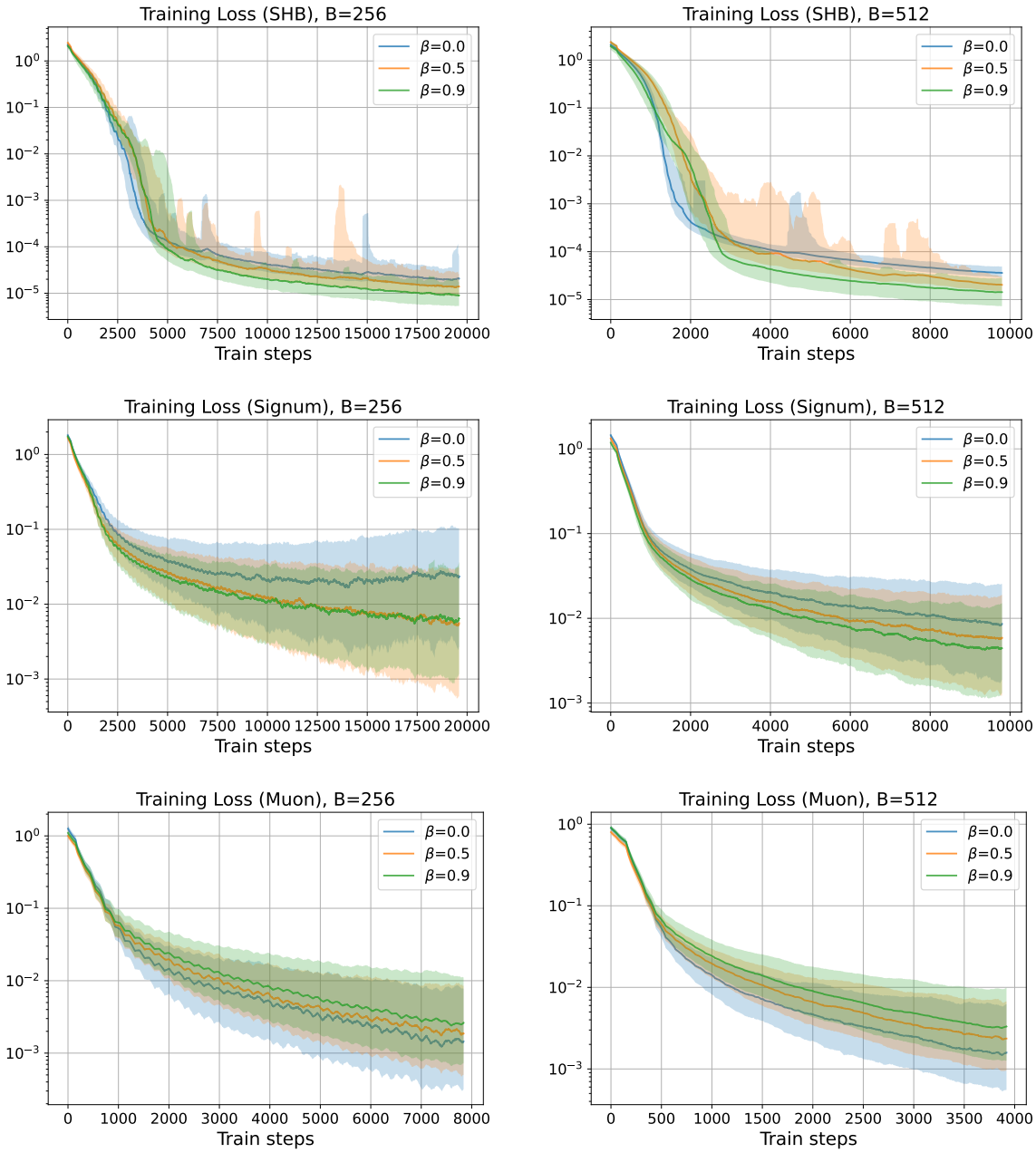

Figure 7: Training loss trajectories under different momentum parameters β across optimizers and batch sizes. Solid lines and shaded regions represent the median and min-max range over 5 seeds, respectively. Smoothed using a moving average with window size 300.

J.2.3. FULL-BATCH EXPERIMENTS

We conducted experiments using a subset of 5,000 training samples from CIFAR-10. This subset was constructed by uniformly selecting 500 samples from each of the 10 classes. We selected the

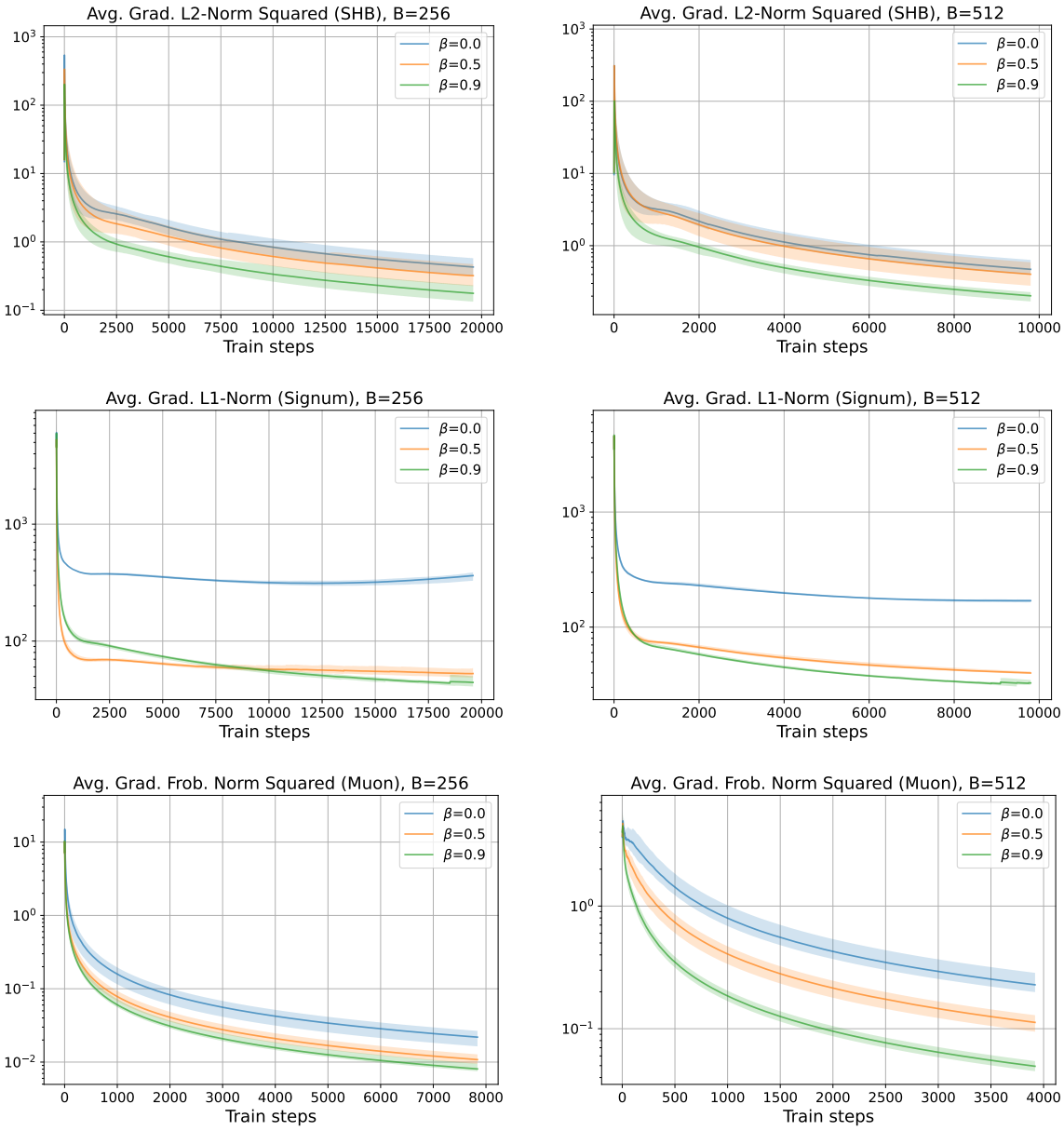

Figure 8: Average gradient norm under different momentum parameters β across optimizers and batch sizes. Solid lines and shaded regions represent the median and min-max range over 5 seeds, respectively.

optimal learning rate in terms of training loss and average gradient ℓ_2-norm. The selected learning rates are summarized as follows:

- **Training loss.** $\eta = 0.001$ for $\beta = 0.0$, $\eta = 0.0003$ for $\beta = 0.5$, and $\eta = 0.0003$ for $\beta = 0.9$.
- **Average gradient ℓ_2-norm squared.** $\eta = 0.3$ for $\beta = 0.0$, $\eta = 0.2$ for $\beta = 0.5$, and $\eta = 0.1$ for $\beta = 0.9$.

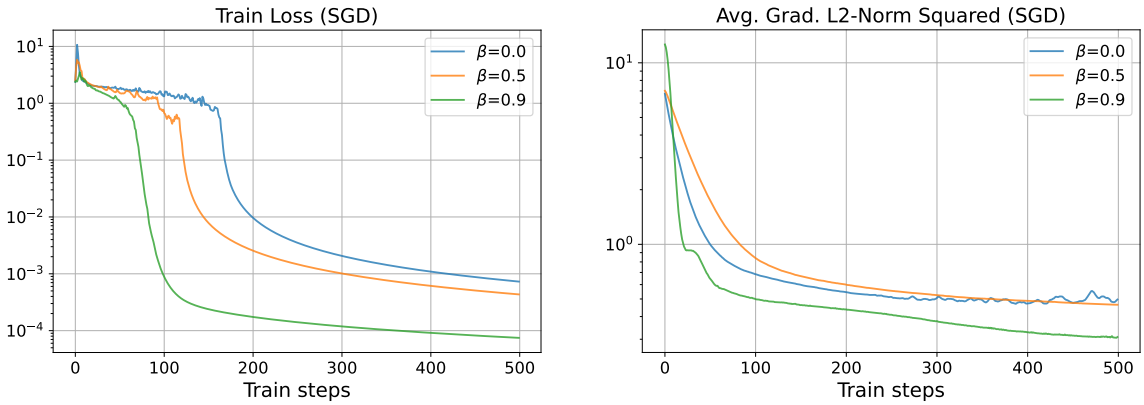

Figure 9: Training loss and average gradient ℓ_2-norm squared for full-batch SGD.

The results of this comparison are presented in Figure 9.

Appendix K. Details and Results of Rosenbrock Experiments

K.1. Setup

We consider the Rosenbrock function [26]

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2,$$

which is a standard non-convex test function with a narrow curved valley.

We use initial points of the form

$$\boldsymbol{x}_0(\delta) = (-1, 1 + \delta).$$

We use the offsets

$$\delta \in \{-0.1, -0.075, -0.05, -0.025, 0, 0.025, 0.05, 0.075, 0.1\}.$$

We compare heavy-ball momentum with gradient descent. We sweep

$$\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.92, 0.94, 0.95, 0.96, 0.97, 0.98\}.$$

For each value of β, each offset δ, and each metric, the step size is tuned separately. We evaluate the averaged squared gradient norm, the best-iterate squared gradient norm, and the final function value. The final function value is included only as an additional diagnostic.

K.2. Step Size Tuning

We parameterize the step size as $\eta = 10^s$ and search over $s \in [-6.5, 1.5]$. For each configuration, we evaluate a uniform coarse grid in s, then refine selected candidate intervals using scalar optimization. The same tuning protocol is applied separately to heavy-ball momentum and gradient descent for each metric. Runs that diverge receive infinite metric value and are not selected by the tuning procedure.

For each metric $\mathbb{M}$, we view $\mathbb{M}(A; \boldsymbol{x}_0)$ as the value of the chosen performance criterion obtained by running algorithm A from the initial point $\boldsymbol{x}_0$. In our experiments, $\mathbb{M}$ is one of the averaged squared gradient norm, the best-iterate squared gradient norm, or the final function value. We report the comparison ratio between heavy-ball momentum and gradient descent

$$R_{\mathbb{M}}(\beta, \delta) = \frac{\min_{\eta>0} \mathbb{M}(\mathrm{HB}_{\beta,\eta}; \boldsymbol{x}_0(\delta))}{\min_{\eta>0} \mathbb{M}(\mathrm{GD}_{\eta}; \boldsymbol{x}_0(\delta))}.$$

Thus, $R_{\mathbb{M}}(\beta, \delta) < 1$ means that the tuned heavy-ball run attains a smaller value of the metric than the tuned gradient descent run. In the plots below, the ratio is shown on a logarithmic scale.

K.3. Results

Figure 10 shows the comparison ratios between heavy-ball momentum and gradient descent for selected offsets δ. The qualitative comparison depends on the metric. For the averaged squared gradient norm, heavy-ball momentum can attain a smaller tuned value for some offsets, but this behavior is not uniform across the initializations. At the same time, the best-iterate squared gradient norm and the final function value provide different qualitative comparisons.

These observations support the point made in Section 5: the convergence metric and the function class should not necessarily be viewed in isolation. Even for the same Rosenbrock objective, the comparison between heavy-ball momentum and gradient descent can depend on how performance is measured.

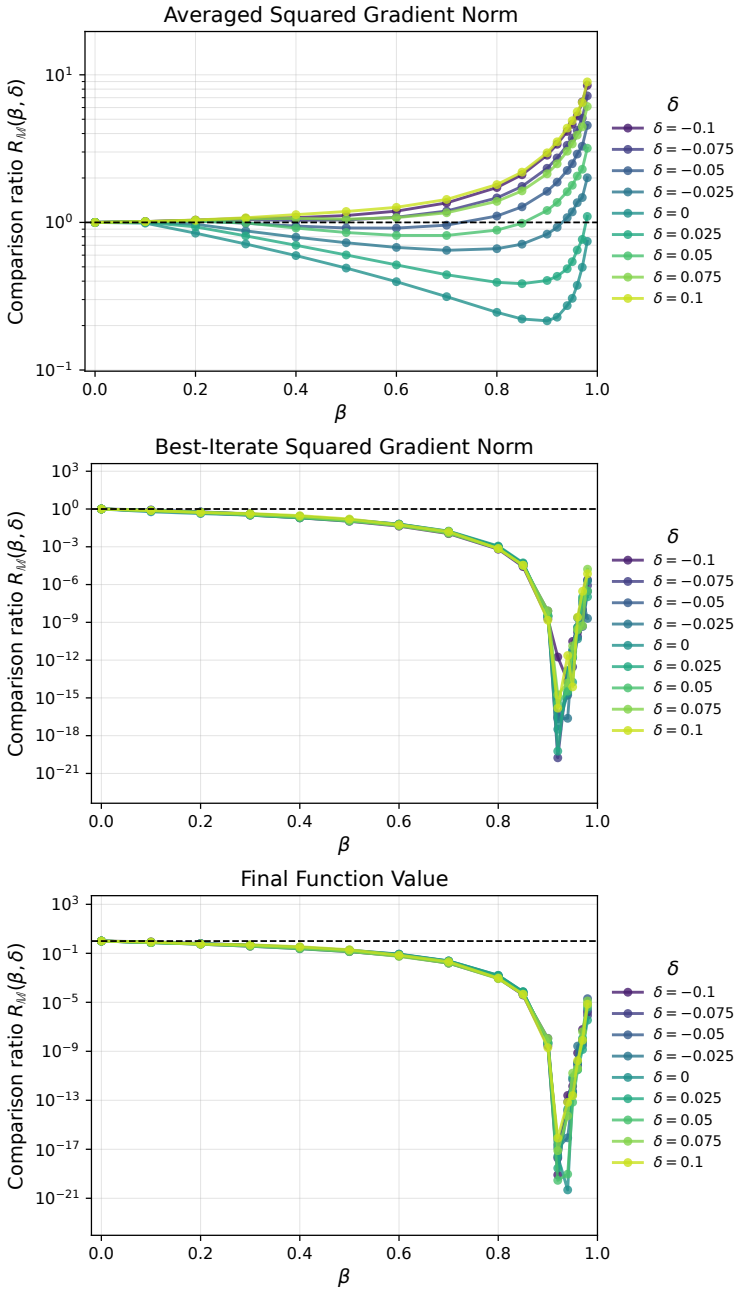

Figure 10: Comparison ratios between heavy-ball momentum and gradient descent on the Rosenbrock function for selected offsets δ in the initialization $\boldsymbol{x}_0(\delta) = (-1, 1 + \delta)$. For each fixed β, offset δ, and metric, the step size is tuned separately. The black dashed line indicates 1. Values below 1 indicate that heavy-ball momentum attains a smaller tuned value of the corresponding metric.