BEYOND RATE OPTIMALITY: HOW POLYAK'S MOMENTUM SHAPES NON-CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Polyak's heavy-ball momentum is widely used in deep learning, where it often accelerates training, yet its advantage is not consistently reflected in standard smooth non-convex optimization theory. Since GD and SGD already attain optimal worst-case rates, any benefit of momentum within this framework must appear at a finer level. We therefore compare worst-case bounds beyond their asymptotic order, focusing on their dependence on the momentum parameter. Many existing convergence guarantees remain unfavorable to momentum, raising the question of whether this reflects loose analyses or genuine worst-case behavior. For SGD with heavy-ball momentum (SHB), SignGD with momentum (Signum), and deterministic Muon, we show that lower bounds on the respective averaged gradient norms can exceed the upper bounds for their non-momentum counterparts over commonly used ranges of β, even when each method uses its optimal constant step size. The gaps widen 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 show that the unfavorable momentum dependence in existing guarantees cannot in general be attributed solely to loose upper bounds, suggesting that the standard framework may miss key ingredients behind momentum's practical benefits, such as alternative convergence measures, structural assumptions, or a more refined treatment of stochasticity.

1 INTRODUCTION

The development of momentum-based acceleration traces back to the foundational work on Polyak's heavy-ball method (Polyak, 1964) and Nesterov's accelerated gradient (NAG) (Nesterov, 1983). Polyak's heavy-ball method enhances convergence on ill-conditioned quadratics, while NAG achieves the optimal $\mathcal{O}(T^{-2})$ rate for the class of smooth convex functions, where T denotes the number of iterations. 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 (Sutskever et al., 2013).

In practical large-scale applications, the heavy-ball-type update has become especially prevalent, due to its ease of implementation. Its use extends well beyond the standard stochastic gradient descent (SGD) framework and has been adopted across various methods, including Adam (Kingma & Ba, 2014), Signum (Bernstein et al., 2018), and Muon (Jordan et al., 2024), which represent adaptive, sign-based, and spectral optimizers, respectively.

Despite its empirical success, it remains unclear whether standard smooth non-convex theory predicts an advantage of momentum over its non-momentum counterparts. At the level of worst-case rates, there is no room for a uniform improvement: GD and SGD already achieve the optimal rates in the deterministic and stochastic settings, respectively (Carmon et al., 2020; Arjevani et al., 2023). Any advantage of momentum within this theory must therefore appear at a finer level, such as in the leading constants hidden by big-$\mathcal{O}$ notation or their dependence on the momentum parameter.

Yet existing convergence guarantees provide little evidence of such an improvement, and the difficulty of reconciling the empirical effectiveness of momentum with existing guarantees has been noted repeatedly in prior work (Liu et al., 2020; Fatkhullin et al., 2023). As we discuss in Section 3, the refined SHB bound of Liu et al. (2020) matches the rate of SGD while improving the momentum

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dependence of the stochastic noise term relative to earlier analyses. After minimizing the guarantees over the constant step size, however, its asymptotic leading constant still increases with β. In the deterministic setting, existing upper bounds for heavy-ball, Signum, and Muon diverge as $\beta \to 1$ (Liu et al., 2020; Sun et al., 2023; Shen et al., 2026). For momentum AdaGrad and Adam, Défossez et al. (2022) substantially sharpen earlier upper bounds, while noting that the resulting guarantees still deteriorate as the momentum parameter increases.

These upper-bound comparisons need not reflect a fundamental limitation of momentum, as the bounds themselves may be loose. Moreover, even under smooth non-convex assumptions, momentum has been shown to improve convergence for stochastic sign-based and normalized methods (Sun et al., 2023; Cutkosky & Mehta, 2020). This raises the question of whether sharper analyses can eliminate the unfavorable momentum dependence, or whether it reflects genuine worst-case behavior.

We address this question by deriving worst-case lower bounds for several methods with Polyak-style momentum. Under the same smooth non-convex assumptions and stationarity-based criteria, these lower bounds can exceed upper bounds for the corresponding non-momentum methods in specific regimes, even when the constant step size is optimized separately for each method. Thus, the unfavorable comparisons cannot in general be attributed solely to looseness in existing analyses, highlighting a limitation of this framework in accounting for the broad practical benefits of momentum.

1.1 Summary of Our Contributions

Our contributions center on worst-case lower bounds for methods with Polyak-style momentum in the stationarity-based smooth non-convex setting. We compare them with upper bounds for the corresponding non-momentum methods, allowing each method to use its own optimized constant step size. For SHB, Signum, and Muon, we study their standard averaged gradient norm criteria; for deterministic HB, we additionally consider the best-iterate squared gradient norm.

- In Section 4, we prove worst-case lower bounds for SHB. Under the averaged squared gradient norm, the SHB lower bound exceeds the corresponding SGD upper bound over a β-dependent range of stochasticity levels that expands with β. In the deterministic case, we prove an asymptotic best-iterate squared gradient norm lower bound scaling as $((1-\beta)T)^{-1}$ for every fixed $\beta \in [5/6, 1)$, yielding the same unfavorable comparison for HB throughout this range.
- In Section 5, we derive a deterministic worst-case lower bound for Signum. For values of the momentum parameter satisfying $\beta > 0.863$, this lower bound exceeds the corresponding SignGD upper bound, with a gap that grows as β approaches 1. We further show in Section 6 that the same construction can be realized by Muon on diagonal matrices, transferring the lower bound from the ℓ_1-norm to the nuclear norm and yielding an analogous separation between Muon and SpecGD.
- In Section 7, 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.

1.2 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 (Polyak, 1964). 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 (Pan et al., 2024) and to accelerate locally under the Polyak-Łojasiewicz condition around global minima (Kassing & Weissmann, 2024). More recently, Lu et al. (2026) show that heavy-ball momentum can accelerate optimization under a river-valley landscape model by stabilizing oscillations in sharp directions and permitting substantially larger learning rates. In stochastic difference-of-convex optimization, momentum can also be necessary for convergence under standard smoothness and bounded-variance assumptions with small batch sizes (Chayti & Jaggi, 2025). For normalized and sign-based methods, momentum has also been shown to improve convergence behavior (Cutkosky & Mehta, 2020; Sun et al., 2023; Yang et al., 2023). Momentum can likewise provide provable benefits for error-feedback methods in compressed stochastic optimization (Fatkhullin et al., 2023). These works demonstrate that momentum can be provably beneficial in a variety of settings, although the assumptions and mechanisms underlying these benefits differ substantially across results.

At the same time, existing first-order lower bounds for smooth non-convex optimization match the rates achieved by GD and SGD in the deterministic and stochastic settings, respec-

tively (Carmon et al., 2020; Arjevani et al., 2023). Beyond these general complexity results, several studies establish more specific limitations of momentum. For instance, Goujaud et al. (2025) show that HB does not yield an acceleration for the class of smooth strongly-convex functions. In stochastic regimes, Kidambi et al. (2018) demonstrate that for certain quadratic objectives, standard momentum methods fail to provide any acceleration over gradient descent. Ganesh et al. (2023) highlight that momentum can be insufficient to improve convergence rates in the presence of gradient noise. Wang et al. (2024) 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. (2024) show that SHB can be slower than SGD on strongly convex quadratic objectives in a regime where the momentum parameter approaches one jointly with the step size. Furthermore, Sahu et al. (2026) 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 momentum can provide provable benefits, but can also fail to accelerate and even slow down optimization in some regimes. Our work asks whether the unfavorable comparisons given by existing smooth non-convex upper bounds are merely artifacts of their analyses, and our lower bounds show that this need not be the case.

2 PRELIMINARIES

Notation. For vectors, $\|\cdot\|_p$ denotes the ℓ_p-norm; for matrices, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_*$ denote the Frobenius and nuclear norms. We apply $\mathrm{sign}(\cdot)$ element-wise, with $\mathrm{sign}(0) = 0$, and write $f^\star := \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$. For $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, let $\mathrm{SVD}(\boldsymbol{X}) = (\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V})$ be its compact SVD, with $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, $\boldsymbol{U} \in \mathbb{R}^{m\times k}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times k}$ having orthonormal columns, $\boldsymbol{S} \in \mathbb{R}^{k\times k}$ diagonal with non-negative entries, and $k = \mathrm{rank}(\boldsymbol{X})$. We write $\mathbf{0}_{m\times n}$ for the zero matrix in $\mathbb{R}^{m\times n}$. For $\boldsymbol{a} = (a_1, \ldots, a_d) \in \mathbb{R}^d$, $\mathrm{diag}(\boldsymbol{a}) = \mathrm{diag}(a_1, \ldots, a_d)$ denotes the $d \times d$ diagonal matrix with diagonal entries $a_1, \ldots, a_d$.

2.1 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 2.1 (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 2.2 (Function class)**.** For some $\Delta > 0$ and $L > 0$, we define

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

When the domain is a matrix space, we use the same notation $\mathcal{F}_L(\Delta)$, with $\mathbf{0}$ interpreted as the zero matrix and L-smoothness measured with respect to the Frobenius norm.

This function class is standard in lower bound analyses of first-order methods (Carmon et al., 2020; Arjevani et al., 2023). The choice of the origin as the initial point is without loss of generality. Indeed, any instance with objective $\tilde{f}$ and initial point $\boldsymbol{x}_0$ satisfying $\tilde{f}(\boldsymbol{x}_0) - \tilde{f}^\star \le \Delta$ can be translated to an instance in $\mathcal{F}_L(\Delta)$ by defining $f(\boldsymbol{x}) := \tilde{f}(\boldsymbol{x} + \boldsymbol{x}_0)$.

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

Assumption 2.4 (Bounded variance)**.** There exists $\sigma \ge 0$ such that $\mathbb{E}\left[\|g(\boldsymbol{x}; \xi) - \nabla f(\boldsymbol{x})\|_2^2\right] \le \sigma^2$.

At each iteration, ξ_t is sampled independently of the past.

2.2 ALGORITHMS

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] Note that setting $\beta = 0$ recovers the corresponding methods without momentum. In particular, SHB reduces to SGD, Signum reduces to SignSGD, and Muon reduces to Spectral Gradient Descent (SpecGD). In the deterministic setting ($\sigma = 0$), we refer to SHB and SignSGD as HB and SignGD, respectively. 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:

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

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.

Remark. If one considers an update rule using $\boldsymbol{m}'_{t+1} = \beta\boldsymbol{m}'_t + (1-\beta)\boldsymbol{g}_t$, then $\boldsymbol{m}'_{t+1} = (1-\beta)\boldsymbol{m}_{t+1}$. Since $1-\beta > 0$, the updates for Signum and Muon remain identical as the positive scalar factor is eliminated by the $\mathrm{sign}(\cdot)$ and $\mathrm{SVD}(\cdot)$ operations. For SHB, an update using $\boldsymbol{m}'_{t+1}$ with step size $\eta/(1-\beta)$ is equivalent to SHB in Algorithm 1.

Algorithm 1 SHB and Signum

Input: Initial point $\boldsymbol{x}_0$, step size $\eta > 0$, momentum parameter $\beta \in [0, 1)$, number of iterations T
Initialize $\boldsymbol{m}_0 = \boldsymbol{0}$
for $t = 0$ **to** $T-1$ **do**
  $\boldsymbol{g}_t = g(\boldsymbol{x}_t; \xi_t)$
  $\boldsymbol{m}_{t+1} = \beta\boldsymbol{m}_t + \boldsymbol{g}_t$
  if SHB **then**
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\boldsymbol{m}_{t+1}$
  else if Signum **then**
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\,\mathrm{sign}(\boldsymbol{m}_{t+1})$
  end if
end for

Algorithm 2 Muon

Input: Initial point $\boldsymbol{X}_0$, step size $\eta > 0$, momentum parameter $\beta \in [0, 1)$, number of iterations T
Initialize $\boldsymbol{M}_0 = \boldsymbol{0}$
for $t = 0$ **to** $T-1$ **do**
  $\boldsymbol{G}_t = g(\boldsymbol{X}_t; \xi_t)$
  $\boldsymbol{M}_{t+1} = \beta\boldsymbol{M}_t + \boldsymbol{G}_t$
  $(\boldsymbol{U}_{t+1}, \boldsymbol{S}_{t+1}, \boldsymbol{V}_{t+1}) = \mathrm{SVD}(\boldsymbol{M}_{t+1})$
  $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta\boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^\top$
end for

2.3 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 a function class $\mathcal{F}$, let $\mathcal{I}_\sigma(\mathcal{F})$ denote the class of objective-oracle pairs (f, g) such that $f \in \mathcal{F}$ and g satisfies Assumptions 2.3 and 2.4 with variance bound σ^2. Given an algorithm class $\mathcal{A}$, we measure performance by

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

where $\mathbb{M}(f, g, A)$ denotes a convergence metric for algorithm A on objective f with stochastic oracle g (*e.g.*, $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q]$, $\min_{0\le t<T}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|^q]$). The infimum and supremum represent optimal algorithmic parameter tuning and the worst-case objective-oracle pair, 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 held fixed across iterations within each run, although it may depend on the horizon T. The function class $\mathcal{F}$ is given by $\mathcal{F}_L(\Delta)$, and the corresponding instance class is $\mathcal{I}_\sigma(\mathcal{F}_L(\Delta))$. We compare worst-case bounds themselves, rather than only their asymptotic order in T. Our comparison focuses on contrasting the cases $\beta = 0$ and $\beta > 0$, corresponding to algorithms without and with momentum, respectively.

3 MOMENTUM DEPENDENCE IN EXISTING UPPER BOUNDS

Since GD and SGD already attain the optimal worst-case rates under the standard smooth non-convex assumptions, any advantage of momentum within this framework must appear at a finer level. We therefore look beyond rates and compare representative momentum upper bounds with their non-momentum counterparts, focusing on how the comparison depends on the momentum parameter.

3.1 HEAVY-BALL MOMENTUM FOR GD AND SGD

We begin with SGD and SHB under the standard stochastic smooth non-convex assumptions, using the averaged squared gradient norm. We first state the SGD upper bound.

[1] The original Muon (Jordan et al., 2024) 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.

Proposition 3.1 (Ghadimi & Lan (2013))**.** *Let $f \in \mathcal{F}_L(\Delta)$. Suppose that the stochastic first-order oracle g satisfies Assumptions 2.3 and 2.4. Consider the SGD iterates $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta g(\boldsymbol{x}_t; \xi_t)$ with a constant step size $\eta \in (0, 2/L)$. Then,*

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

The minimum of the right-hand side with respect to $\eta \in (0, 2/L)$ is $\frac{\Delta L}{T}\big(1 + \sqrt{1 + 2\sigma^2 T/(\Delta L)}\big)$.

Remark. There exists an L-smooth function and a stochastic first-order oracle for which the bound in Proposition 3.1, when stated with $f(\boldsymbol{x}_0) - \mathbb{E}[f(\boldsymbol{x}_T)]$ in place of Δ, holds with *equality*. See Appendix E.1 for a constructive example.

We next recall a representative upper bound for the SHB method under the same assumptions.

Proposition 3.2 (Liu et al. (2020), Theorem 1)**.** *Let $f \in \mathcal{F}_L(\Delta)$ and $0 < \eta \le (1-\beta)^2/(L(4-\beta+\beta^2))$.*[2] *Suppose that the stochastic first-order oracle g satisfies Assumptions 2.3 and 2.4. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ denote the iterates generated by SHB. Then,*

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

Remark. Note that Liu et al. (2020) 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.

For $\sigma > 0$ and sufficiently large T, minimizing the SGD and SHB upper bounds separately over the constant step size yields the same $\mathcal{O}(T^{-1/2})$ rate for both methods. However, the ratio of the SHB upper bound to the SGD upper bound converges to $\sqrt{\frac{5\beta^2+9\beta+8}{2(\beta+1)}}$ as $T \to \infty$. This ratio is strictly greater than 1 for all $\beta \in [0, 1)$, increasing from 2 at $\beta = 0$ to $\sqrt{5.5} \approx 2.34$ as $\beta \to 1$; see Appendix E.2 for a detailed derivation. Thus, despite having the same asymptotic rate, the upper-bound comparison is consistently unfavorable to SHB and becomes increasingly so as β grows. Yet the analysis of Liu et al. (2020) refines a sequence of earlier results (Yang et al., 2016; Yan et al., 2018; Yu et al., 2019), notably removing the additional $(1-\beta)^{-1}$ dependence in the stochastic term appearing in Yu et al. (2019).

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 3.2 reduces to $\frac{2\Delta L}{T}\frac{\beta^2-\beta+4}{1-\beta}$. Comparing the two bounds, their ratio scales as $\Theta((1-\beta)^{-1})$.

3.2 Signum and Muon

We turn to Signum and Muon and focus on deterministic guarantees. For stochastic sign-based methods, SignSGD can fail to converge whereas Signum admits convergence guarantees (Yang et al., 2023; Sun et al., 2023); see Section 5 for discussion. In the deterministic setting, however, existing upper bounds become increasingly unfavorable as β grows. We first state the SignGD upper bound.

Proposition 3.3. *Suppose that f is L-smooth. Consider the SignGD iterates with a constant step size $\eta > 0$. Then, with $\Delta = f(\boldsymbol{x}_0) - f^\star$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \le \frac{\Delta}{\eta T} + \frac{Ld}{2}\eta. \tag{3}$$

The minimum of the right-hand side with respect to $\eta > 0$ is $\sqrt{2\Delta L d/T}$.

The proof is given in Appendix E.3. The optimized upper bound in Proposition 3.3 is asymptotically tight; see Appendix E.4 for the proof.

Next, we derive a convergence guarantee for Signum, using the proof framework in Sun et al. (2023).

Proposition 3.4. *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 $\Delta = f(\boldsymbol{x}_0) - f^\star$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \le \frac{\Delta}{\eta T} + \frac{Ld}{2}\left(\frac{1+3\beta}{1-\beta}\right)\eta + \frac{2\sqrt{2\Delta L d}}{T}\frac{\beta}{1-\beta}, \tag{4}$$

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

[2] While the original step-size condition in Liu et al. (2020) involves a minimum of two terms, $(1-\beta)^2/(L(4-\beta+\beta^2))$ is the stronger constraint for all $\beta \in [0, 1)$, so it simplifies to $\eta \le (1-\beta)^2/(L(4-\beta+\beta^2))$.

See Appendix E.5 for the full proof. Thus, the optimized SignGD and Signum bounds have the same $\mathcal{O}(T^{-1/2})$ rate, while the ratio of their leading coefficients is $\sqrt{\frac{1+3\beta}{1-\beta}}$. Hence, the known upper-bound comparison becomes increasingly unfavorable to Signum in the large-momentum regime.

A similar comparison holds for SpecGD and Muon. For $m \times n$ matrix variables, let $r := \min\{m, n\}$. Building on the exact-SVD Muon analysis of Shen et al. (2026, Theorem 4.1), we obtain a Muon upper bound of the same form as Proposition 3.4, with d replaced by r.[3] The full statement and proof are given in Appendix E.6. At $\beta = 0$, the bound reduces to the SpecGD upper bound, and the ratio of the leading coefficients of the optimized Muon and SpecGD bounds scales as $\Theta((1-\beta)^{-1/2})$ as $\beta \to 1$. Thus, the upper-bound comparison is likewise increasingly unfavorable to Muon as β grows.

3.3 Adaptive Methods

Unfavorable momentum dependence has also appeared in convergence guarantees for adaptive methods. Writing β_1 for the first-moment momentum parameter, existing upper bounds exhibit an $\mathcal{O}((1-\beta_1)^{-3})$ dependence for momentum AdaGrad (Shen et al., 2024) and an $\mathcal{O}((1-\beta_1)^{-5})$ dependence for Adam (Zou et al., 2019). Défossez et al. (2022) improve this to $\mathcal{O}((1-\beta_1)^{-1})$ for both methods, although the resulting bounds still deteriorate as β_1 increases.

Together, these results show that unfavorable momentum dependence appears across several existing upper bounds but may be reduced by sharper analyses. Upper bounds alone therefore cannot determine whether the unfavorable comparisons above are intrinsic. We next turn to worst-case lower bounds to examine whether analogous deterioration is intrinsic rather than an artifact of the analysis.

4 SGD vs SHB

In this section, we study heavy-ball momentum under two standard gradient-based convergence criteria. We first derive a worst-case lower bound for SHB under the averaged squared gradient norm and compare it with the SGD upper bound in Section 3. We then ask whether the resulting lower/upper-bound separation is specific to averaging. In the deterministic setting, we show that an analogous separation persists for HB under the best-iterate squared gradient norm.

4.1 Averaged Squared Gradient Norm Lower Bound for SHB

Theorem 4.1 provides a lower bound for SHB, under the averaged squared gradient norm.

Theorem 4.1. *Suppose* $T \geq \max\left\{32,\ 56M\left[\log\big(400(1+M)\big)\right]^2,\ \frac{6}{1-\beta}\log\frac{112}{1-\beta}\right\}$, *where* $M := \sigma^2/(2\Delta L)$. *Then, the SHB iterates* $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ *with momentum parameter* β *and step size* $\eta > 0$ *satisfy*

$$\inf_{\eta>0}\sup_{(f,g)\in\mathcal{I}_\sigma(\mathcal{F}_L(\Delta))}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_t)\|_2^2] \geq \frac{\Delta L}{T}\left(1-\frac{1}{2T}\right)\left(\sqrt{\Psi}+\frac{1+\beta^2}{1-\beta^2}-\frac{M}{2}\right) \quad (5)$$

where $\Psi = 1 - 2M\frac{1+\beta}{1-\beta} + 4MT$. *Moreover, the right-hand side of (5) is positive.*

We defer the proof of Theorem 4.1 to Appendix A. To establish this worst-case lower bound, we use a simple hard instance: up to a translation, it is given by a one-dimensional quadratic function $Lx^2/2$ with additive gradient noise $\pm\sigma$ with probability $1/2$ each. This quadratic has condition number 1, so it lies outside the ill-conditioned quadratic regime underlying Polyak's classical heavy-ball acceleration. Thus, our lower bound does not contradict this classical acceleration result. Instead, it shows that under the averaged squared gradient norm, the standard smooth non-convex worst-case framework can yield a comparison unfavorable to SHB.

Figure 1 compares the SHB lower bound from Theorem 4.1 with the SGD upper bound from Proposition 3.1 across the stochasticity level $M = \sigma^2/(2\Delta L)$ and the momentum parameter β. For each fixed β, there is a threshold on M below which the comparison is unfavorable to SHB, and this threshold increases as $\beta \to 1$. When M is small, the separation appears for a broader range of momentum parameters; in particular, if $M = 0$, the ratio between the SHB lower bound and the SGD upper bound is approximately $(1-\beta^2)^{-1}$ for sufficiently large T.

[3] Shen et al. (2026) use $\boldsymbol{M}_0 = \nabla f(\boldsymbol{X}_0)$, whereas we use $\boldsymbol{M}_0 = \boldsymbol{0}_{m\times n}$. Under our initialization, an additional T^{-1} term appears, while the leading $T^{-1/2}$ term remains unchanged.

To understand the boundary, we note that

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

where $\Psi = 1 - 2M\frac{1+\beta}{1-\beta} + 4MT$. Thus, for sufficiently large T, we have $1-\frac{1}{2T} \approx 1$ and $\Psi \approx 1 + 4MT$. In this regime, the lower/upper-bound comparison 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 this asymptotic boundary for large T in the following corollary, with the proof deferred to Appendix B.

Corollary 4.2. *Let $M = \sigma^2/(2\Delta L) \geq 0$ 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 4.1 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.*

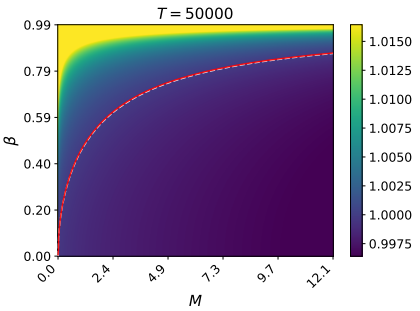

Figure 1: Ratio of the SHB lower bound to the SGD upper bound for $T = 50000$ (5–95th percentile clipping). The red line marks ratio 1; above this line, the SHB lower bound exceeds the SGD upper bound. The white dashed curve is the asymptotic boundary $\frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} = 1$.

4.2 Best-Iterate Lower Bound for HB

The comparison above is stated for the averaged squared gradient norm. Since this metric accumulates the gradient norms over all iterates, one may ask whether the unfavorable comparison for heavy-ball momentum is mainly caused by averaging. We show that this is not the case, at least in the deterministic setting: an analogous comparison also appears under the best-iterate metric

$$\min_{0\leq t<T}\|\nabla f(\boldsymbol{x}_t)\|_2^2.$$

For gradient descent, we recall the following *tight* upper bound under the best-iterate metric.

Proposition 4.3 (Rotaru et al. (2024))**.** *Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ denote the iterates generated by gradient descent with step size $\eta > 0$ and $\sigma = 0$. Then,*

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

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

Theorem 4.4. *Suppose $d \geq 4$ and fix $\beta \in [5/6, 1)$. Let $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ denote the iterates generated by HB with momentum parameter β and step size $\eta > 0$. Then,*

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

The full proof of Theorem 4.4 is provided in Appendix C. To complement this asymptotic result, Appendix F reports finite-horizon Performance Estimation Problem (PEP, Drori & Teboulle (2014)) computations. These results show that, after optimizing the step size, the worst-case best-iterate guarantee for HB deteriorates as β increases at fixed finite T.

Combining Proposition 4.3 and Theorem 4.4, GD admits a best-iterate upper bound with leading T^{-1} coefficient at most $3\sqrt{3}\Delta L/4$, whereas HB has a best-iterate lower bound with coefficient at least $\Delta L/(4(1-\beta))$. Hence, the ratio of the HB lower bound to the GD upper bound is asymptotically at least $1/(3\sqrt{3}(1-\beta))$. For every $\beta \in [5/6, 1)$, this ratio is at least $2/\sqrt{3} > 1$, and it diverges as $\beta \to 1$. Thus, in the deterministic setting, the unfavorable comparison for HB persists under the best-iterate squared gradient norm and it is not solely an artifact of using the averaged gradient norm.

5 SignGD vs Signum

The role of momentum in sign-based optimization changes substantially in the presence of stochasticity. Under standard smoothness and bounded-variance assumptions, Yang et al. (2023) prove a non-convergence result for normalized SGD and note that their one-dimensional construction also applies to SignSGD, whereas Sun et al. (2023) establish convergence of Signum. Thus, momentum can improve the convergence behavior of stochastic sign-based methods.

These results concern stochastic settings and therefore do not determine whether momentum provides a similar benefit in the deterministic setting considered below. In contrast, as discussed in Section 3, the existing upper-bound comparison between Signum and SignGD becomes increasingly unfavorable to Signum as β grows in the deterministic setting. In fact, the ratio between the two bounds diverges as $\beta \to 1$. We therefore ask whether this unfavorable upper-bound comparison is merely an artifact of the analysis or also arises in worst-case performance.

5.1 Lower Bound for Signum

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

Theorem 5.1. *Consider the iterates $\{\boldsymbol{x}_t\}_{t=0}^{T-1}$ generated by Signum with momentum parameter β, step size $\eta > 0$ and $\sigma = 0$. Suppose $T \geq 20 + 1/(1-\beta)$ and $\beta > 0.64$. Then,*

$$\inf_{\eta>0} \sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1 \geq \sqrt{\frac{\Delta L d}{T}\left(\frac{7}{8} + \frac{5}{12\sqrt{1-\beta}}\right)}.$$

The full proof of Theorem 5.1 is provided in Appendix D.2. We now compare the SignGD upper bound and the Signum lower bound under the optimal step sizes. We have

$$\text{SignGD UB} = \sqrt{\frac{2\Delta L d}{T}}, \quad \text{Signum LB} = \sqrt{\frac{2\Delta L d}{T}}\sqrt{\frac{7}{16} + \frac{5}{24\sqrt{1-\beta}}}.$$

Therefore, if $\beta > 0.863$, the Signum lower bound exceeds the corresponding SignGD upper bound, and the ratio between the two bounds scales as $\Theta((1-\beta)^{-1/4})$.

Together with prior work on stochastic sign-based methods, our result shows that the comparison of momentum and non-momentum methods can differ between stochastic and deterministic settings. In stochastic settings, SignSGD can fail to converge whereas Signum admits convergence guarantees; our lower bound shows the deterministic worst-case comparison can instead favor SignGD.

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 3.4. 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 Appendix D.3 for more discussion.

6 Implications for SpecGD vs Muon

In this section, we show that the Signum lower-bound construction can be realized within the Muon dynamics on diagonal matrix instances. Thus, the deterministic lower bound established for Signum directly applies to Muon. This is formalized in the following proposition, proved in Appendix E.7.

Proposition 6.1. *Let $\boldsymbol{x}_0, \boldsymbol{x}' \in \mathbb{R}^d$, $\boldsymbol{X}_0 = \operatorname{diag}(\boldsymbol{x}_0)$, and $\boldsymbol{X}' = \operatorname{diag}(\boldsymbol{x}')$. Let $\{\boldsymbol{X}_t, \boldsymbol{M}_t\}$ be generated by Muon with $\sigma = 0$ on $f(\boldsymbol{X}) = \frac{L}{2}\|\boldsymbol{X} + \boldsymbol{X}'\|_{\mathrm{F}}^2$, and let $\{\boldsymbol{x}_t, \boldsymbol{m}_t\}$ be generated by Signum on $\frac{L}{2}\|\boldsymbol{x} + \boldsymbol{x}'\|_2^2$ from the initialization $\boldsymbol{x}_0$. Then, for every t, $\boldsymbol{X}_t = \operatorname{diag}(\boldsymbol{x}_t)$ and $\boldsymbol{M}_t = \operatorname{diag}(\boldsymbol{m}_t)$.*

By Proposition 6.1, 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. Combining this equivalence with Theorem 5.1 yields the following lower bound for Muon.

Corollary 6.2. *Consider the iterates $\{\boldsymbol{X}_t\}_{t=0}^{T-1}$ generated by Muon with momentum parameter β, step size $\eta > 0$ and $\sigma = 0$. Suppose $T \geq 20 + 1/(1-\beta)$ and $\beta > 0.64$. Then,*

$$\inf_{\eta>0} \sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{X}_t)\|_* \geq \sqrt{\frac{\Delta L d}{T}\left(\frac{7}{8} + \frac{5}{12\sqrt{1-\beta}}\right)}.$$

For $d \times d$ matrices, setting $\beta = 0$ in Proposition E.1 gives the SpecGD upper bound $\sqrt{2\Delta L d/T}$. Thus, the lower bound in Corollary 6.2 exceeds the SpecGD upper bound whenever $\beta > 0.863$, yielding the same separation as for SignGD and Signum.

7 Discussion

7.1 Deep Learning Experiments

To compare our worst-case results with deep learning practice, Appendix G reports experiments in two settings: ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky et al., 2009), and a nanoGPT-style Transformer (Karpathy, 2022) on the character-level Tiny Shakespeare dataset (Karpathy, 2015). In these experiments, methods with Polyak-style momentum generally attain smaller gradient norms more rapidly than their non-momentum counterparts. Thus, gradient-norm-based criteria can still favor momentum in deep learning, motivating the question of what the standard smooth non-convex framework may be missing.

7.2 What Can and Cannot the Standard Framework Explain?

Convergence Metrics. A natural question is whether the unfavorable lower/upper-bound comparisons above are tied to how convergence is assessed in the standard smooth non-convex framework. Here,

the usual goal is to find a point with small gradient norm rather than a global minimum. Other convergence measures introduce different issues: the function-value gap targets global optimality, which is computationally hard even for smooth unconstrained non-convex objectives (Murty & Kabadi, 1987), whereas finite-time guarantees for the last-iterate gradient norm are impossible under the standard smooth non-convex assumptions alone (Drori & Shamir, 2020). Thus, stationarity-based convergence measures have been considered as the standard objects of study in this setting.

Our results yield unfavorable lower/upper-bound comparisons for SHB and deterministic Signum and Muon under averaged gradient norms, and for deterministic HB under the best-iterate squared gradient norm. Thus, at least for HB, switching from the averaged to the standard best-iterate criterion does not remove the unfavorable worst-case comparison. Together with Section 7.1, where the averaged squared gradient norm can favor SHB in deep learning, this suggests that the mismatch cannot be attributed simply to the use of averaged gradient norms.

Function Class. Another possible source of the mismatch is the function class. Deep learning objectives often violate global L-smoothness (Zhang et al., 2020), suggesting that the standard class $\mathcal{F}_L(\Delta)$ may miss key aspects of practical objectives. However, simply broadening the function class is not enough: any larger class containing our hard instances would inherit the same lower bounds. The role of these instances is especially transparent for the averaged gradient norm lower bounds. They use isotropic quadratics of the form $\frac{L}{2}\|\boldsymbol{x}\|_2^2$ (up to translation), with additive noise in the SHB construction. Their condition number is $\kappa = 1$, far from the ill-conditioned quadratic regime where Polyak's heavy-ball method is classically accelerated (Polyak, 1964). Thus, worst-case analysis over the standard L-smooth class can be driven by geometries lacking the anisotropic or curved structure where momentum is expected to help. Recent studies of deep learning optimization have further highlighted non-isotropic geometric structure, such as river-valley-like loss landscapes (Song et al., 2025a;b; Wen et al., 2025). Building on this line of work, Lu et al. (2026) analyze heavy-ball momentum on non-convex landscapes with flat and sharp directions, and show that momentum can damp oscillations in the sharp directions, allowing substantially larger learning rates and faster progress. These results suggest that explaining momentum's practical advantage may require assumptions distinguishing such geometries from the isotropic instances driving our lower bounds.

Stochasticity. Our results on SHB show that as the noise level increases, the range of β for which the SHB lower bound exceeds the SGD upper bound narrows, indicating that stochasticity affects the comparison between SHB and SGD. Relatedly, Defazio (2020) uses a Lyapunov analysis of SGD with momentum to study when momentum can be beneficial. The analysis shows that a negative iterate-difference term can partially cancel the stochastic-gradient contribution in the bound, and relates the resulting benefit of momentum to the alignment between the momentum buffer and subsequent gradients. Existing results for normalized and sign-based methods provide stronger evidence that stochasticity can qualitatively affect the role of momentum. In particular, momentum can enable convergence in settings where the corresponding normalized or sign-based method without momentum may fail to converge (Cutkosky & Mehta, 2020; Yang et al., 2023; Sun et al., 2023). These results suggest that stochasticity can make the theoretical comparison more favorable to momentum in some settings, although how broadly this extends remains unclear. Stochasticity alone also cannot fully account for the effectiveness of heavy-ball momentum in deep learning, as its advantage can persist even in full-batch settings; see Appendix G.

Remark. Another widely used variant of momentum is Nesterov accelerated gradient (NAG, Nesterov (1983)), which is incorporated into optimizers such as Nadam (Dozat, 2016). Our results concern heavy-ball updates, as used in SHB, Signum, and Muon, and do not imply an analogous lower/upper-bound separation for NAG. Extending our lower-bound analysis to NAG is an interesting direction.

8 Conclusion

We establish worst-case lower bounds for Polyak-style momentum methods in standard smooth non-convex settings. For SHB and deterministic versions of Signum and Muon, our results show that lower bounds on the corresponding averaged gradient norms can exceed upper bounds for their non-momentum counterparts, even when each method uses its own optimized constant step size. For HB, we further show that a similar separation holds under the best-iterate squared gradient norm. Thus, the unfavorable comparisons arising in this framework cannot in general be attributed solely to loose upper-bound analyses. Together with positive results in more structured or stochastic settings, our findings suggest that explaining when momentum is beneficial may require alternative convergence criteria, additional structural assumptions, or a more refined account of stochasticity.

AI USE STATEMENT

AI tools were used to improve the clarity and readability of the manuscript, to assist in identifying potentially relevant literature, and to help simplify some mathematical proofs. Literature suggested by AI tools was independently checked against the original sources before being cited in the paper. For proof-related assistance, the authors independently verified all mathematical arguments and retained only arguments that they confirmed to be correct. The authors reviewed all AI-assisted content and take full responsibility for the final content of the paper.

REFERENCES

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

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.

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

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

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

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

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

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPQhzTSWA7.

Xiaoge Deng, Tao Sun, Dongsheng Li, and Xicheng Lu. Exploring the inefficiency of heavy ball as momentum parameter approaches 1. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 3899–3907. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/431. URL https://doi.org/10.24963/ijcai.2024/431. Main Track.

Timothy Dozat. Incorporating nesterov momentum into adam. In *International Conference on Learning Representations Workshop*, 2016.

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

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

Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtarik. Momentum provably improves error feedback! In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 76444–76495. Curran Associates, Inc., 2023. doi: 10.52202/075280-3342. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f0b1515be276f6ba82b4f2b25e50bef0-Paper-Conference.pdf.

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 (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 602–612. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/ganesh23a.html.

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

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.

Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method. *Mathematical Programming*, pp. 1–59, 2025.

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*, pp. 770–778, 2016.

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/.

Andrej Karpathy. char-rnn. https://github.com/karpathy/char-rnn, 2015.

Andrej Karpathy. nanogpt. https://github.com/karpathy/nanoGPT, 2022.

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.

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)*, pp. 1–9. IEEE, 2018.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 45672–45718, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4c0129f1f7a036af5ed2409f1f9f8131-Paper-Conference.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18261–18271. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d3f5d4de09ea19461dab00590df91e4f-Paper.pdf.

Miao Lu, Zeyu Bian, Kaiyue Wen, Beining Wu, Siyu Chen, Tianhao Wang, and Zhiyuan Li. Towards understanding momentum acceleration in river-valley loss landscape. In *High-dimensional Learning Dynamics 2026*, 2026. URL https://openreview.net/forum?id=yRO8rLiOUL.

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.

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

Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 55989–56028, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f780a86b7145988ac219d49d8e37a58f-Paper-Conference.pdf.

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

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.

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.

Li Shen, Congliang Chen, Fangyu Zou, Zequn Jie, Ju Sun, and Wei Liu. A unified analysis of adagrad with weighted aggregation and momentum acceleration. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14482–14490, 2024. doi: 10.1109/TNNLS.2023.3279381.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=4nH4CulGaP. J2C Certification.

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

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*, 2025b. URL https://openreview.net/forum?id=CGx4XU9rCA.

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 (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33077–33099. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sun23l.html.

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 (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

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.

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.

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.

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, pp. 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.

Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned sgd and the power of adaptive methods. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74257–74288. Curran Associates, Inc., 2023. doi: 10.52202/075280-3248. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/eb1a323fa10d4102ff13422476a744ff-Paper-Conference.pdf.

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.

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*, pp. 7184–7193. PMLR, 2019.

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.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11119–11127, 2019. doi: 10.1109/CVPR.2019.01138.

A Proof of Theorem 4.1

A.1 Proof Sketch

The proof proceeds in four steps. First, we restrict attention to the quadratic $f(x) = Lx^2/2$ with an additive Rademacher noise oracle. We decompose each SHB iterate into the corresponding deterministic heavy-ball iterate and a linear combination of the past noise variables. Taking expectations then separates the objective into a deterministic term and a term coming from the accumulated noise.

Second, we rewrite the finite sums appearing in the objective using the corresponding infinite sums. These infinite sums can be computed explicitly, which gives a lower bound of the form

$$F(u) - \frac{1}{T}\sum_{t=T}^{\infty} \tilde{y}_t^2$$

for a reparameterization u of the step size. The explicit function $F(u)$ determines the main term.

Third, we lower bound $F(u)$ uniformly over the step size and then control the remaining sum $\sum_{t=T}^{\infty} \tilde{y}_t^2$. For sufficiently small step sizes, we instead lower bound the original finite sum directly. For the remaining step sizes, we divide the analysis according to the roots of the heavy-ball recurrence and bound the remaining sum in each case.

Finally, we show that the stated condition on T guarantees all the inequalities required in the preceding steps. Combining the resulting bounds gives a uniform lower bound over all step sizes and yields the claimed result.

A.2 Setup and Preliminary Reductions

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

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

Then,

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

Suppose $f(x) = \frac{L}{2}x^2$ and $\tilde{f}(x) = \frac{L}{2}(x+\alpha)^2$. 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 $\tilde{f}$ with $\tilde{x}_0 = 0$. Then,

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

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

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 2.3 and 2.4 with variance σ^2. Moreover, the random variables $\{\xi_t\}_{t\ge 0}$ are assumed to be independent across iterations.

For convenience, define the normalized step size $h := \eta L$.

Then,

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

with $x_{-1} = x_0$.

Now, we define y_k as

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

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 A.1.

$$x_k = y_k - \eta \sum_{i=0}^{k-1} d_{k-i}\xi_i, \tag{6}$$

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

Proof. We use induction.

Base case.

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

Inductive step. We now assume that (6) holds for all $k \leq n$. Then,

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

Therefore, by induction, (6) is proved.

□

By the independence of ξ_i, we have

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

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,

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

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

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

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 h, β, and M.

Throughout the remainder of the proof, we assume that

$$T \ge \max\left\{32,\ 56M\left[\log(400(1+M))\right]^2,\ \frac{6}{1-\beta}\log\frac{112}{1-\beta}\right\}. \tag{7}$$

Let $L_T := \log(8T)$. We first show that (7) implies

$$T \ge \frac{23}{4}L_T, \quad \sqrt{T} \ge \frac{23}{5}\sqrt{M}\,L_T, \quad (1-\beta)T \ge \frac{51}{10}L_T. \tag{8}$$

For the first inequality, T/L_T is increasing for $T \ge 32$, and

$$\frac{32}{\log 256} > \frac{23}{4}.$$

For the M-dependent inequality, define $L_M := \log(400(1+M))$ and $T_M := 56ML_M^2$. If $T_M \ge 32$, then

$$\log(8T_M) = \log(448ML_M^2) \le L_M + 2\log L_M + \log\frac{28}{25} \le \frac{13}{8}L_M.$$

Since

$$\frac{23}{5}\cdot\frac{13}{8} = \frac{299}{40} < \sqrt{56},$$

we get $\sqrt{T_M} \ge (23/5)\sqrt{M}\log(8T_M)$. If $T_M < 32$, then

$$\sqrt{M} < \frac{\sqrt{32/56}}{L_M} \le \frac{\sqrt{32/56}}{\log 400},$$

and hence directly

$$\frac{23}{5}\sqrt{M}\log 256 < \sqrt{32}.$$

The result for every larger T follows because $\sqrt{T}/\log(8T)$ is increasing for $T \ge 32$.

For the β-dependent inequality, write $\delta := 1-\beta$, $L_\beta := \log(112/\delta)$, and $T_\beta := 6L_\beta/\delta$. At $T = T_\beta$,

$$\log(8T_\beta) = L_\beta + \log\frac{3L_\beta}{7}.$$

For all $L > 0$, we have

$$\log\frac{3L}{7} \le \frac{3}{17}L.$$

Therefore

$$\log(8T_\beta) \le \frac{20}{17}L_\beta,$$

so

$$\delta T_\beta = 6L_\beta \ge \frac{51}{10}\log(8T_\beta).$$

If $T_\beta < 32$, then $32\delta > 6L_\beta \ge 6\log 112$, while

$$6\log 112 > \frac{51}{10}\log 256.$$

Again use monotonicity of $T/\log(8T)$ for larger T.

A.3 Infinite Sums

Throughout this subsection, suppose $0 < h < 2(1+\beta)$. Let

$$Y_T = \sum_{k=0}^{T-1}\tilde{y}_k^2, \quad D_T = \sum_{k=0}^{T-1} d_k^2, \quad E_T = \sum_{k=0}^{T-1} k d_k^2$$

$$Y_\infty = \sum_{k=0}^{\infty}\tilde{y}_k^2, \quad D_\infty = \sum_{k=0}^{\infty} d_k^2, \quad E_\infty = \sum_{k=0}^{\infty} k d_k^2.$$

Our goal is to obtain a lower bound on the following objective, uniformly over h:

$$\begin{aligned}\mathcal{J}_T(h) &= \frac{1}{T}\sum_{k=0}^{T-1}|\tilde{y}_k|^2 + \frac{h^2M}{T}\sum_{j=0}^{T-1}(T-j)|d_j|^2 \\ &= \frac{Y_T}{T} + h^2M\left(D_T - \frac{E_T}{T}\right). \end{aligned} \tag{9}$$

Since

$$\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)d_j^2 = D_\infty - \frac{E_\infty}{T} + \sum_{j=T}^{\infty}\left(\frac{j}{T}-1\right)d_j^2$$

and the last term is non-negative, we have

$$\mathcal{J}_T(h) \ge \frac{Y_\infty}{T} + h^2M\left(D_\infty - \frac{E_\infty}{T}\right) - \frac{1}{T}\sum_{t=T}^{\infty}\tilde{y}_t^2. \tag{10}$$

Therefore, by finding appropriate upper bounds for the tail sum $\frac{1}{T}\sum_{t=T}^{\infty}\tilde{y}_t^2$, we can obtain a lower bound for $\mathcal{J}_T(h)$.

Define

$$u := \frac{h}{2(1+\beta)-h}.$$

For the remainder of the proof, define

$$r := \frac{1-\beta}{1+\beta}, \quad c := \frac{1+\beta^2}{1-\beta^2}.$$

We first prove the following lemma:

Lemma A.2. *Suppose $\beta \in [0, 1)$ and $0 < h < 2(1 + \beta)$. Then,*

$$\begin{aligned} Y_\infty &= \frac{1}{4}\left(\frac{r}{u} + 2c + \frac{u}{r}\right) \\ h^2 D_\infty &= \frac{u}{r} \\ h^2 E_\infty &= \frac{1}{4}\left(1 + \frac{2u}{r^2} + u^2\right). \end{aligned}$$

Proof. Set $a := 1 - h + \beta$. Since $0 < h < 2(1 + \beta)$, both roots of

$$\lambda^2 - a\lambda + \beta = 0$$

lie strictly inside the unit disk. Hence all infinite sums below converge.

First, define

$$Q_\infty := \sum_{k=1}^{\infty} d_k d_{k-1}.$$

Multiplying the recurrence $d_{k+1} = ad_k - \beta d_{k-1}$ by d_k and summing over $k \geq 1$ gives

$$(1 + \beta)Q_\infty = aD_\infty.$$

Squaring the recurrence and summing over $k \geq 1$ gives

$$D_\infty - 1 = (a^2 + \beta^2)D_\infty - 2a\beta Q_\infty.$$

Eliminating Q_∞, we obtain

$$\begin{aligned} D_\infty &= \frac{1 + \beta}{(1 - \beta)\left((1 + \beta)^2 - a^2\right)} \\ &= \frac{1 + \beta}{(1 - \beta)h\left(2(1 + \beta) - h\right)}. \end{aligned}$$

Since

$$u = \frac{h}{2(1 + \beta) - h},$$

it follows that

$$h^2 D_\infty = \frac{u}{r}.$$

Next, the sequences $\tilde{y}_k$ and $d_{k+1} - \beta d_k$ satisfy the same recurrence and have the same initial values. Therefore,

$$\tilde{y}_k = d_{k+1} - \beta d_k.$$

Consequently,

$$\begin{aligned} Y_\infty &= \sum_{k=0}^{\infty} (d_{k+1} - \beta d_k)^2 \\ &= (1 + \beta^2)D_\infty - 2\beta Q_\infty. \end{aligned}$$

Substituting the expressions for D_∞ and Q_∞, and using

$$h = \frac{2(1 + \beta)u}{1 + u},$$

gives

$$Y_\infty = \frac{1}{4}\left(\frac{r}{u} + 2c + \frac{u}{r}\right).$$

It remains to compute E_∞. Define

$$R_\infty := \sum_{k=1}^{\infty} k\, d_k d_{k-1}.$$

Multiplying the recurrence by kd_k and summing over $k \geq 1$ yields

$$(1+\beta)R_\infty = aE_\infty + Q_\infty.$$

On the other hand, multiplying the squared recurrence by $k+1$ and summing over $k \geq 1$ gives

$$E_\infty - 1 = a^2(E_\infty + D_\infty) + \beta^2(E_\infty + 2D_\infty) - 2a\beta(R_\infty + Q_\infty).$$

Eliminating R_∞ from the last two identities and substituting

$$Q_\infty = \frac{a}{1+\beta} D_\infty, \quad D_\infty = \frac{1+\beta}{(1-\beta)h\,(2(1+\beta)-h)}$$

gives, after simplification,

$$h^2 E_\infty = \frac{1}{4}\left(1 + \frac{2u}{r^2} + u^2\right).$$

□

A.4 Chebyshev Polynomials and Exclusion of Large Step Sizes

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

For $\beta > 0$, define

$$z := \frac{1-h+\beta}{2\sqrt{\beta}}.$$

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:

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

Lemma A.3. *Suppose $\beta > 0$. Then $\tilde{y}_k$ and d_k admit the representations*

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

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

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

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

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 with $\tilde{y}_{-1} = \tilde{y}_0 = 1$.
- 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 $d_1 = 1$.

□

We next use this representation to exclude large step sizes from the minimization of $\mathcal{J}_T(h)$.

Lemma A.4. *For every $h \geq (1 + \sqrt{\beta})^2$,*

$$\mathcal{J}_T(h) \geq \mathcal{J}_T\left((1 - \sqrt{\beta})^2\right). \tag{11}$$

Consequently,

$$\inf_{h>0} \mathcal{J}_T(h) = \inf_{0<h\leq(1+\sqrt{\beta})^2} \mathcal{J}_T(h).$$

Proof. We write $\tilde{y}_k(h)$ and $d_k(h)$ to make their dependence on h explicit. First, suppose $\beta > 0$, and let

$$h_- := (1 - \sqrt{\beta})^2, \quad h_+ := (1 + \sqrt{\beta})^2.$$

If $h \geq h_+$, then $z = -q$ for some $q \geq 1$, whereas $h = h_-$ corresponds to $z = 1$. Using $U_k(-q) = (-1)^k U_k(q)$, we obtain

$$|\tilde{y}_k(h)| = \sqrt{\beta}^k \left(U_k(q) + \sqrt{\beta}U_{k-1}(q)\right),$$
$$|\tilde{y}_k(h_-)| = \sqrt{\beta}^k \left((k+1) - k\sqrt{\beta}\right).$$

Since $U_k(q) \geq U_k(1) = k + 1$ and $U_{k-1}(q) \geq U_{k-1}(1) = k$ for $q \geq 1$, it follows that

$$|\tilde{y}_k(h)| \geq |\tilde{y}_k(h_-)|.$$

Likewise, for every $k \geq 1$,

$$|d_k(h)| = \sqrt{\beta}^{k-1} U_{k-1}(q) \geq \sqrt{\beta}^{k-1} k = |d_k(h_-)|.$$

Moreover, $h \geq h_+ \geq h_-$. Since $M \geq 0$ and $1 - k/T \geq 0$ for $1 \leq k \leq T - 1$, these termwise comparisons imply (11).

If $\beta = 0$, then

$$\tilde{y}_k(h) = (1 - h)^k, \quad d_k(h) = (1 - h)^{k-1} \quad (k \geq 1).$$

Hence every $h \geq 1$ is dominated by $h = 1$, which again gives (11). Finally, $\mathcal{J}_T(h)$ is continuous in h, so the relation yields the stated restriction of the minimization domain.

□

Suppose $\beta > 0$. Let $\lambda_{1,2}$ be the roots of characteristic polynomial of $\tilde{y}_k$ (and d_k). Then,

- $|z| < 1$ if and only if λ_1 and λ_2 are nonreal complex conjugates.
- $z = \pm 1$ if and only if $\lambda_1 = \lambda_2 = \pm\sqrt{\beta}$.

A.5 Minimizing the Infinite-Horizon Lower Bound

By Lemma A.2 and (10), we obtain

$$\mathcal{J}_T(h) \geq \frac{r}{4Tu} + \frac{c}{2T} - \frac{M}{4T} + \frac{\Psi}{4Tr}u - \frac{M}{4T}u^2 - \frac{1}{T}\sum_{t=T}^{\infty} \tilde{y}_t^2,$$

where

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

By (7), we have $T \geq \frac{2}{1-\beta} \geq \frac{1}{2r}$ and this implies $M\,(4T - 2/r) \geq 0$. Thus,

$$\Psi = 1 + M\left(4T - \frac{2}{r}\right) \geq 1. \tag{12}$$

Also, by Lemma A.4, we can only consider the case $0 < u \leq \left(\frac{1+\sqrt{\beta}}{1-\sqrt{\beta}}\right)^2 =: u_{\max}$.

Let

$$G(u) := \frac{r}{4Tu} + \frac{c}{2T} - \frac{M}{4T} + \frac{\Psi}{4Tr}u.$$
$$F(u) := G(u) - \frac{M}{4T}u^2.$$

Then, $G(u)$ is minimized at $u_g := \frac{r}{\sqrt{\Psi}}$.

We will show that F has a unique minimizer u_f on $(0, u_{\max}]$, when $T \geq \frac{5}{1-\beta}$. We have

$$4TF'(u_{\max}) = \left(\frac{1}{r} - \frac{r}{u_{\max}^2}\right) + M\left(\frac{4T}{r} - \frac{2}{r^2} - 2u_{\max}\right),$$

For the first term, since $u_{\max} \geq 1$ and $r \leq 1$, we have $u_{\max}^2 \geq r^2$ and consequently $\frac{1}{r} - \frac{r}{u_{\max}^2} \geq 0$. For the second term, it is enough that

$$T \geq \frac{1}{2r} + \frac{u_{\max} r}{2}. \tag{13}$$

Since $r = \frac{1-\beta}{1+\beta}$ and $u_{\max} = \left(\frac{1+\sqrt{\beta}}{1-\sqrt{\beta}}\right)^2$, we have

$$\frac{1}{2r} + \frac{u_{\max} r}{2} = \frac{\beta^2 + 2\beta^{3/2} + 4\beta + 2\sqrt{\beta} + 1}{(1-\beta)(1+\beta)} \leq \frac{5}{1-\beta}.$$

Thus, under (7), we have (13) and the second term is also non-negative, which concludes that $4TF'(u_{\max})$ is non-negative.

Since $F''(u) = \frac{r}{2Tu^3} - \frac{M}{2T}$ is strictly decreasing and $F''(0+) = \infty$, F' is either increasing throughout $(0, u_{\max}]$ or increases and then decreases. Thus, under (13), $F'(0+) = -\infty$ and $F'(u_{\max}) \geq 0$ imply that F' has a unique zero in $(0, u_{\max}]$. Hence, F has a unique minimizer u_f on $(0, u_{\max}]$.

We next compare u_f with u_g. Since

$$F'(u) = \frac{1}{4T}\left(-\frac{r}{u^2} + \frac{\Psi}{r} - 2Mu\right),$$

for any $\rho \geq 1$, direct substitution of $u_g = r/\sqrt{\Psi}$ gives

$$\frac{F'(\rho u_g)}{\Psi/(4Tr)} = 1 - \frac{1}{\rho^2} - \frac{2Mr^2}{\Psi^{3/2}}\rho.$$

In particular, $F'(u_g) \leq 0$, because $\Psi \geq 1$ by (12).

If

$$\frac{2Mr^2}{\Psi^{3/2}} < \frac{\rho^2 - 1}{\rho^3}, \tag{14}$$

then $F'(\rho u_g) > 0$.

Let

$$\chi := 4T - \frac{2}{r},$$

so that $\Psi = 1 + \chi M$. Since $\chi > 0$, which is already proved when showing (12), we have

$$\sup_{M \geq 0} \frac{M}{(1+\chi M)^{3/2}} = \frac{2}{3\sqrt{3}\,\chi},$$

and hence

$$\frac{2Mr^2}{\Psi^{3/2}} \le \frac{4r^2}{3\sqrt{3}\,\chi}.$$

Therefore, it suffices to choose $\rho > 1$ such that

$$\frac{4r^2}{3\sqrt{3}\,\chi} < \frac{\rho^2 - 1}{\rho^3}. \tag{15}$$

We now set

$$\rho := 1 + \frac{1}{T}.$$

By (7), $T \ge 2/(1-\beta)$, and hence $\chi = 4T - \frac{2}{r} \ge 2T$. Since $r \le 1$,

$$\frac{4r^2}{3\sqrt{3}\,\chi} \le \frac{2}{3\sqrt{3}\,T} < \frac{1}{2T}.$$

Moreover, $T \ge 2$ gives

$$\frac{\rho^2 - 1}{\rho^3} = \frac{T(2T+1)}{(T+1)^3} \ge \frac{1}{2T}.$$

Thus, (15) holds, and $u_g \le r \le 1 \le u_{\max}$. The uniqueness of u_f, together with $F'(u_g) \le 0$ and $F'(\rho u_g) > 0$, then implies

$$u_g \le u_f \le \rho u_g$$

when $\rho u_g \le u_{\max}$. If $\rho u_g > u_{\max}$, the same conclusion follows from $u_f \le u_{\max} < \rho u_g$. Consequently,

$$\begin{aligned}
\min_{0<u\le u_{\max}} F(u) &= F(u_f) \\
&= G(u_f) - \frac{M}{4T}u_f^2 \\
&\ge G(u_g) - \frac{M}{4T}\rho^2 u_g^2 \\
&= \frac{\sqrt{\Psi} + c - M/2}{2T} - \frac{\rho^2}{4}\frac{Mr^2}{T\Psi}.
\end{aligned} \tag{16}$$

Define

$$A_\rho := \sqrt{\Psi} + c - \frac{M}{2} - \frac{\rho^2 M r^2}{2\Psi}.$$

Then, (16) can be written as

$$\min_{0<u\le u_{\max}} F(u) \ge \frac{A_\rho}{2T}.$$

A.6 Tail Bounds

Before considering the three root regimes, define

$$K := \sqrt{\Psi} + c.$$

We first establish several bounds that will be used throughout the proof.

Lemma A.5. *Under (7),*

$$\frac{T}{K} \ge \frac{3}{2}L_T, \quad A_\rho > \frac{31}{32}K, \quad (1-\beta)A_\rho > \frac{31}{40}.$$

Proof. For the first condition, we use the fact that

$$K = \sqrt{\Psi} + c \le \sqrt{1 + 4MT} + \frac{1}{1-\beta}.$$

Using (7), we obtain

$$\frac{3L_T}{2T}K \le \frac{3}{2}\sqrt{\left(\frac{L_T}{T}\right)^2 + \frac{4ML_T^2}{T}} + \frac{3L_T}{2(1-\beta)T} \le \frac{3\sqrt{116}}{46} + \frac{5}{17} < 1.$$

For the second condition, we use the fact that $(1-\beta)T > 5L_T$ and $L_T \ge \log 256 \ge 16/3$, which follow from (8). Then, we obtain

$$\frac{2}{r} \le \frac{4}{1-\beta} \le \frac{4T}{5L_T} \le \frac{3T}{20}.$$

Thus,

$$\chi = 4T - \frac{2}{r} > \frac{77}{20}T, \quad \Psi = 1 + \chi M \ge 1 + \frac{77}{20}MT.$$

In particular, $\Psi \ge 1$.

Since $K = \sqrt{\Psi} + c \ge \sqrt{\Psi}$, we have

$$K^2 \ge \Psi \ge \frac{77}{20}MT.$$

Also, since $T \ge 3L_T K/2$ and $L_T \ge \log 256$, we have

$$M \le \frac{20}{77}\frac{K}{T}K \le \frac{40}{231L_T}K \le \frac{40}{231\log 256}K \le \frac{1}{30}K.$$

Using $\Psi = 1 + \chi M \ge \chi M$, $\chi = 4T - 2/r \ge 2T$, $K \ge 2$ and $\rho = 1 + 1/T \le 33/32$, we obtain

$$\frac{\rho^2 M r^2}{2\Psi K} \le \frac{\rho^2}{2\chi K} \le \frac{(33/32)^2}{8T} < \frac{1}{128}.$$

Then,

$$\frac{A_\rho}{K} = 1 - \frac{M}{2K} - \frac{\rho^2 M r^2}{2\Psi K} > 1 - \frac{1}{60} - \frac{1}{128} > \frac{31}{32}.$$

Finally, since $K = \sqrt{\Psi} + c \ge c$ and $(1-\beta)c = (1+\beta^2)/(1+\beta) \ge 4/5$, we have

$$(1-\beta)A_\rho > \frac{31}{32}(1-\beta)K \ge \frac{31}{32}(1-\beta)c \ge \frac{31}{40}.$$

□

A.6.1 Small Positive-Real-Root Region

For any $\gamma > 0$, consider

$$0 < h \le (1-\sqrt{\beta})^2, \quad 0 < u \le \frac{\gamma r}{2K}. \tag{17}$$

Since $h = \frac{2(1+\beta)u}{1+u}$ and $(1+\beta)r = 1-\beta$, the second condition in (17) implies

$$h \le 2(1+\beta)u \le \frac{\gamma(1-\beta)}{K}. \tag{18}$$

Let $\lambda_+ \ge \lambda_-$ be the two roots of

$$\lambda^2 - (1-h+\beta)\lambda + \beta = 0.$$

The first condition in (17) ensures that the roots are real and non-negative. Moreover, the root formula gives

$$1 - \lambda_+ = \frac{1-\beta+h-\sqrt{(1-\beta+h)^2 - 4h}}{2} > 0,$$

where strict positivity follows from $h > 0$. Thus, $\lambda_+ < 1$. Define

$$\omega := 1 - \lambda_+ \in (0, 1].$$

Substituting $\lambda_+ = 1 - \omega$ into the characteristic equation gives

$$\omega^2 - (1 - \beta + h)\omega + h = 0,$$

and hence

$$\omega = \frac{2h}{1 - \beta + h + \sqrt{(1 - \beta + h)^2 - 4h}}.$$

Because $h \le (1 - \sqrt{\beta})^2 \le 1 - \beta$, we have

$$0 \le \sqrt{(1 - \beta + h)^2 - 4h} \le 1 - \beta - h.$$

It follows that

$$\frac{h}{1 - \beta} \le \omega \le \frac{2h}{1 - \beta}. \tag{19}$$

When $\lambda_+ > \lambda_-$, solving the recurrence for $\tilde{y}_t$ gives

$$\tilde{y}_t = \frac{\lambda_+^{t+1}(1 - \lambda_-) - \lambda_-^{t+1}(1 - \lambda_+)}{\lambda_+ - \lambda_-}.$$

Consequently,

$$\tilde{y}_t - \lambda_+^t = \frac{\lambda_-(1 - \lambda_+)(\lambda_+^t - \lambda_-^t)}{\lambda_+ - \lambda_-} \ge 0.$$

The same inequality follows for the repeated-root case by continuity. Thus,

$$\tilde{y}_t \ge \lambda_+^t = (1 - \omega)^t. \tag{20}$$

Dropping the noise term from $\mathcal{J}_T(h)$, we obtain

$$\begin{aligned}
\mathcal{J}_T(h) &\ge \frac{1}{T} \sum_{t=0}^{T-1} \tilde{y}_t^2 \\
&\ge \frac{1}{T} \sum_{t=0}^{T-1} (1 - \omega)^{2t} \\
&= \frac{1 - (1 - \omega)^{2T}}{T\omega(2 - \omega)}.
\end{aligned}$$

Using (19), we have

$$0 \le \omega(2 - \omega) \le 2\omega \le \frac{4h}{1 - \beta},$$

and combining with $(1 - \omega)^{2T} \le e^{-2\omega T}$, we have

$$\mathcal{J}_T(h) \ge \frac{1 - \beta}{4hT} \left(1 - e^{-2hT/(1-\beta)}\right).$$

The function $\phi(x) := (1 - e^{-x})/x$ is decreasing on $(0, \infty)$ because

$$\phi'(x) = \frac{(x + 1)e^{-x} - 1}{x^2} \le 0.$$

Applying this fact with $x = \frac{2hT}{1-\beta}$, and then using (18) gives

$$\mathcal{J}_T(h) \ge \frac{K}{4\gamma T} \left(1 - e^{-2\gamma T/K}\right).$$

Therefore, this region satisfies the same target lower bound as (16) whenever

$$\frac{1 - e^{-2\gamma T/K}}{2\gamma} \ge \frac{A_\rho}{K}. \tag{21}$$

Now, let $\gamma = 4/9$. Then,

$$\frac{1 - e^{-2\gamma T/K}}{2\gamma} = \frac{9}{8}\left(1 - e^{-8T/(9K)}\right).$$

Since $A_\rho \le K$, it suffices to show that

$$\frac{T}{K} \ge \frac{9}{8}\log 9.$$

Using the first condition in Lemma A.5, we have $T/K \ge 3L_T/2 \ge \frac{3}{2}\log 256$. Since $\frac{3}{2}\log 256 > \frac{9}{8}\log 9$, we have the desired result.

Therefore, in this region,

$$\mathcal{J}_T(h) \ge \frac{A_\rho}{2T}.$$

A.6.2 Remaining Positive-Real-Root Region

We now consider the remaining part of the positive-real-root region,

$$0 < h \le (1 - \sqrt{\beta})^2, \quad u \ge \frac{\gamma r}{2K}. \tag{22}$$

The sequences $\{\tilde{y}_t\}$ and $\{d_t\}$ satisfy

$$d_{t+1} = \sum_{i=0}^{t} \beta^{t-i}\tilde{y}_i, \quad \tilde{y}_{t+1} = \tilde{y}_t - h d_{t+1}. \tag{23}$$

Indeed, the recurrence and initial conditions give

$$\tilde{y}_t = d_{t+1} - \beta d_t,$$

so $d_{t+1} = \tilde{y}_t + \beta d_t$; iterating this identity gives the first equality in (23), while substituting it into the recurrence for d_t gives the second.

In the positive-real-root region, (20) shows that $\tilde{y}_t \ge 0$. Also,

$$d_{t+1} = \frac{\lambda_+^{t+1} - \lambda_-^{t+1}}{\lambda_+ - \lambda_-} \ge 0$$

when the roots are distinct, with the repeated-root case following by continuity. The second identity in (23) therefore shows that $\{\tilde{y}_t\}$ is non-increasing. Consequently, for every $0 \le i \le t$, $\tilde{y}_i \ge \tilde{y}_t$, and hence

$$\begin{aligned} d_{t+1} &= \sum_{i=0}^{t} \beta^{t-i}\tilde{y}_i \\ &\ge \tilde{y}_t \sum_{j=0}^{t} \beta^j = \frac{1 - \beta^{t+1}}{1 - \beta}\tilde{y}_t. \end{aligned}$$

Using this bound in the update for $\tilde{y}_{t+1}$ and the inequality $1 - x \le e^{-x}$ gives

$$\tilde{y}_{t+1} \le \tilde{y}_t \exp\left(-\frac{h}{1-\beta}(1 - \beta^{t+1})\right).$$

Iterating from 0 to $t - 1$ yields

$$\begin{aligned} \tilde{y}_t &\le \exp\left(-\frac{h}{1-\beta}\sum_{j=0}^{t-1}(1 - \beta^{j+1})\right) \\ &= \exp\left[-\frac{h}{1-\beta}\left(t - \frac{\beta(1-\beta^t)}{1-\beta}\right)\right] \\ &\le \exp\left[-\frac{h}{1-\beta}\left(t - \frac{\beta}{1-\beta}\right)\right]. \end{aligned} \tag{24}$$

The second condition in (22) gives a uniform lower bound on the exponential rate. Specifically,

$$\frac{2h}{1-\beta} = \frac{4u}{r(1+u)} \geq \frac{4\gamma}{2K+\gamma r} =: \theta_\gamma. \tag{25}$$

We now show that $T - \frac{\beta}{1-\beta} > 0$ to upper bound $\sum_{t=T}^{\infty} \tilde{y}_t^2$. Using the third condition in (8), we have

$$(1-\beta)T \geq 5L_T \geq 5\log 256 \geq 25$$

and therefore

$$T - \frac{\beta}{1-\beta} \geq T - \frac{1}{1-\beta} \geq \frac{24}{25}T > 0.$$

From (24), we obtain the bound

$$\sum_{t=T}^{\infty} \tilde{y}_t^2 \leq \frac{\exp\left[-\theta_\gamma\left(T - \frac{\beta}{1-\beta}\right)\right]}{1-\exp(-\theta_\gamma)}. \tag{26}$$

Combining the above gives the explicit lower bound

$$\mathcal{J}_T(h) \geq \frac{A_\rho}{2T} - \frac{1}{T}\frac{\exp\left[-\theta_\gamma\left(T - \frac{\beta}{1-\beta}\right)\right]}{1-\exp(-\theta_\gamma)}.$$

Equivalently, for any $\varepsilon_{\mathrm{r}} \in (0,1)$, if $A_\rho > 0$ and

$$\frac{2}{A_\rho}\frac{\exp\left[-\theta_\gamma\left(T - \frac{\beta}{1-\beta}\right)\right]}{1-\exp(-\theta_\gamma)} \leq \varepsilon_{\mathrm{r}}, \tag{27}$$

then

$$\mathcal{J}_T(h) \geq \frac{A_\rho}{2T}(1-\varepsilon_{\mathrm{r}}).$$

Now, we show that $\varepsilon_{\mathrm{r}} = 1/(3T)$ and $\gamma = 4/9$ satisfy (27).

Since $K = \sqrt{\Psi} + c \geq 2$ and $r \leq 1$, we have $2r \leq K$ and therefore

$$\theta_\gamma = \frac{8}{2r+9K} \geq \frac{4}{5K}.$$

Also, since $K \geq 1$, we have $\frac{8}{2r+9K} \leq 1$ and therefore $\theta_\gamma \in (0,1)$. Thus,

$$1 - e^{-\theta_\gamma} \geq \frac{\theta_\gamma}{2}.$$

Then,

$$\begin{aligned}\frac{\exp\left[-\theta_\gamma\left(T - \frac{\beta}{1-\beta}\right)\right]}{1-e^{-\theta_\gamma}} &\leq \frac{2}{\theta_\gamma}\exp\left(-\frac{24}{25}\theta_\gamma T\right) \\ &\leq \frac{5K}{2}\exp\left(-\frac{96T}{125K}\right).\end{aligned}$$

From the first condition in Lemma A.5, we have $T/K \geq 3L_T/2$, and

$$\exp\left(-\frac{96T}{125K}\right) \leq e^{-(144/125)L_T} = (8T)^{-144/125}.$$

Also, from the second condition in Lemma A.5, we have

$$\begin{aligned}\frac{2}{A_\rho}\frac{\exp\left[-\theta_\gamma\left(T - \frac{\beta}{1-\beta}\right)\right]}{1-\exp(-\theta_\gamma)} &\leq \frac{2}{A_\rho}\frac{5K}{2}(8T)^{-144/125} \\ &\leq \frac{160}{31}(8T)^{-144/125} < \frac{1}{3T},\end{aligned}$$

where the last inequality holds for $T \geq 32$ by (7). Therefore, (27) holds with $\varepsilon_{\mathrm{r}} = 1/(3T)$ and $\gamma = 4/9$.

A.6.3 COMPLEX-ROOT REGION

It remains to consider

$$(1-\sqrt{\beta})^2 \le h \le (1+\sqrt{\beta})^2. \tag{28}$$

Suppose first that $\beta > 0$. Then $z \in [-1,1]$, and the Chebyshev representation gives

$$\tilde{y}_t = \beta^{t/2}\left(U_t(z) - \sqrt{\beta}U_{t-1}(z)\right).$$

For completeness, writing $z = \cos\theta$ and using

$$U_t(\cos\theta) = \frac{\sin((t+1)\theta)}{\sin\theta}, \quad |\sin((t+1)\theta)| \le (t+1)|\sin\theta|,$$

shows that $|U_t(z)| \le t+1$ for $z \in (-1,1)$; the same bound holds at $z = \pm 1$ by continuity. Therefore,

$$|\tilde{y}_t| \le \beta^{t/2}\left((t+1) + \sqrt{\beta}\,t\right) \le 2(t+1)\beta^{t/2}.$$

Let

$$a_t := (t+1)^2\beta^t.$$

Then, for $t \ge T$, we have

$$\frac{a_{t+1}}{a_t} = \beta\left(\frac{t+2}{t+1}\right)^2 \le \beta\left(\frac{T+2}{T+1}\right)^2.$$

Moreover, by the third condition in (8), we have $(1-\beta)T \ge 51L_T/10 \ge 51\log 256/10 > 4$, and hence $2/(T+1) \le (1-\beta)/2$. Therefore,

$$\begin{aligned}\log\left(\beta\left(\frac{T+2}{T+1}\right)^2\right) &= \log\beta + 2\log\left(1+\frac{1}{T+1}\right)\\ &\le -(1-\beta) + \frac{2}{T+1}\\ &\le -\frac{1-\beta}{2}.\end{aligned}$$

In particular,

$$\beta\left(\frac{T+2}{T+1}\right)^2 \le e^{-(1-\beta)/2} < 1.$$

Thus, we have

$$\begin{aligned}\sum_{t=T}^{\infty}\tilde{y}_t^2 &\le 4\sum_{t=T}^{\infty}(t+1)^2\beta^t\\ &\le \frac{4(T+1)^2\beta^T}{1-\beta\left(\frac{T+2}{T+1}\right)^2}.\end{aligned} \tag{29}$$

When $\beta = 0$, the interval in (28) consists only of $h = 1$. In that case, $\tilde{y}_0 = 1$ and $\tilde{y}_t = 0$ for every $t \ge 1$, so (29) continues to hold for $T \ge 1$ with a zero right-hand side.

Combining the above yields

$$\mathcal{J}_T(h) \ge \frac{A_\rho}{2T} - \frac{4(T+1)^2\beta^T}{T\left[1-\beta\left(\frac{T+2}{T+1}\right)^2\right]}.$$

In particular, for any $\varepsilon_c \in (0,1)$, if $A_\rho > 0$ and

$$\frac{8(T+1)^2\beta^T}{A_\rho\left[1-\beta\left(\frac{T+2}{T+1}\right)^2\right]} \le \varepsilon_c, \tag{30}$$

then

$$\mathcal{J}_T(h) \geq \frac{A_\rho}{2T}(1-\varepsilon_{\mathrm{c}}).$$

Now, we show that $\varepsilon_{\mathrm{c}} = 1/(3T)$ satisfies (30). When $\beta = 0$, the condition holds immediately since the tail is zero. Suppose that $\beta > 0$. From the bound established above,

$$1-\beta\left(\frac{T+2}{T+1}\right)^2 \geq 1-e^{-(1-\beta)/2} \geq \frac{1-\beta}{4}.$$

Also,

$$\beta^T \leq e^{-(1-\beta)T} \leq e^{-5L_T} = (8T)^{-5}.$$

Combining the above and using $(1-\beta)A_\rho > 31/40$ from Lemma A.5, we have

$$\begin{aligned}\frac{8(T+1)^2\beta^T}{A_\rho\left[1-\beta\left(\frac{T+2}{T+1}\right)^2\right]} &\leq \frac{32(T+1)^2(8T)^{-5}}{(1-\beta)A_\rho} \\ &< \frac{1280}{31}\frac{(T+1)^2}{(8T)^5} \leq \frac{1}{3T},\end{aligned}$$

where the last inequality holds for $T \geq 1$.

A.7 Completion of the Proof

We first relate A_ρ to the coefficient appearing in the desired lower bound. Recall that

$$A_\rho = K - \frac{M}{2} - \frac{\rho^2 M r^2}{2\Psi}.$$

As shown above, $4T - 2/r \geq 2T$, and hence

$$\Psi = 1 + M\left(4T - \frac{2}{r}\right) \geq 1 + 2MT.$$

Therefore, $M/\Psi \leq 1/(2T)$. Moreover, by Lemma A.5,

$$K - \frac{M}{2} \geq A_\rho > \frac{31}{32}K \geq \frac{31}{16},$$

where we used $K \geq 2$. Since $\rho = 1 + 1/T \leq 33/32$ and $r \leq 1$, we obtain

$$\begin{aligned}\frac{(K-M/2)-A_\rho}{K-M/2} &= \frac{\rho^2 M r^2}{2\Psi(K-M/2)} \\ &\leq \frac{\rho^2}{4T(K-M/2)} \\ &< \frac{4(33/32)^2}{31T} < \frac{1}{7T}.\end{aligned}$$

Thus,

$$A_\rho \geq \left(1-\frac{1}{7T}\right)\left(K-\frac{M}{2}\right).$$

We now combine the three regions. In the small positive-real-root region,

$$\mathcal{J}_T(h) \geq \frac{A_\rho}{2T},$$

whereas in the remaining positive-real-root and complex-root regions,

$$\mathcal{J}_T(h) \geq \frac{A_\rho}{2T}\left(1-\frac{1}{3T}\right).$$

Therefore, for every $0 < h \leq (1+\sqrt{\beta})^2$,

$$\mathcal{J}_T(h) \geq \frac{1}{2T}\left(1-\frac{1}{7T}\right)\left(1-\frac{1}{3T}\right)\left(K-\frac{M}{2}\right).$$

Since

$$\left(1-\frac{1}{7T}\right)\left(1-\frac{1}{3T}\right) = 1-\frac{10}{21T}+\frac{1}{21T^2} \geq 1-\frac{1}{2T},$$

we obtain

$$\mathcal{J}_T(h) \geq \frac{1}{2T}\left(1-\frac{1}{2T}\right)\left(K-\frac{M}{2}\right).$$

Finally, Lemma A.4 extends this bound to every $h > 0$. Substituting $K = \sqrt{\Psi} + c$ and $c = (1+\beta^2)/(1-\beta^2)$, and restoring the normalization factor $2\Delta L$, gives

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

B Proof of Corollary 4.2

Proof. Define

$$\zeta := \frac{1+\beta^2}{1-\beta^2} - \frac{M}{2} - 1 > 0.$$

After dividing both bounds by $\Delta L/T$, their difference is

$$\begin{aligned}\frac{T}{\Delta L}\big(\text{SHB LB} - \text{SGD UB}\big) &= \left(1 - \frac{1}{2T}\right)\left(1 + \zeta + \sqrt{\Psi}\right) - \left(1 + \sqrt{1+4MT}\right) \\ &= \zeta - X_T,\end{aligned}$$

where

$$X_T := \sqrt{1+4MT} - \sqrt{\Psi} + \frac{1}{2T}\left(1 + \zeta + \sqrt{\Psi}\right).$$

Since $\Psi = 1 + 4MT - 2M\frac{1+\beta}{1-\beta}$, we have

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

If $T \ge \frac{1+\beta}{1-\beta}$, then $\Psi \ge 1 + 2MT$, and therefore

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

Moreover, $\Psi \le 1 + 4MT$ gives

$$\frac{1}{2T}\left(1 + \zeta + \sqrt{\Psi}\right) \le \frac{2+\zeta}{2T} + \frac{\sqrt{M}}{\sqrt{T}}.$$

Consequently,

$$X_T \le \frac{1+\beta}{1-\beta}\sqrt{\frac{M}{2T}} + \frac{2+\zeta}{2T} + \frac{\sqrt{M}}{\sqrt{T}}.$$

Set

$$\begin{aligned}T_0 := \max\Bigg\{&32,\ 56M\left[\log\big(400(1+M)\big)\right]^2,\ \frac{6}{1-\beta}\log\frac{112}{1-\beta},\ \frac{1+\beta}{1-\beta}, \\ &\frac{8M}{\zeta^2}\left(\frac{1+\beta}{1-\beta}\right)^2,\ \frac{64M}{\zeta^2},\ \frac{4(2+\zeta)}{\zeta}\Bigg\}.\end{aligned}$$

For every $T > T_0$, the three terms in the preceding upper bound for X_T are smaller than $\zeta/4$, $\zeta/8$, and $\zeta/8$, respectively. Hence $X_T < \zeta/2$, and thus

$$\text{SHB LB} - \text{SGD UB} > \frac{\Delta L}{T}\frac{\zeta}{2} > 0.$$

□

C Proof of Theorem 4.4

C.1 Proof Sketch

The proof proceeds in four steps. First, we reduce the lower bound to the construction of a finite heavy-ball trajectory $\{(\boldsymbol{x}_t, \boldsymbol{g}_t, f_t)\}$ that can be interpolated by a smooth function. The resulting lower bound depends on the gradient norms along the trajectory and the corresponding initial function gap. Thus, we seek trajectories along which the gradients remain large while the function values decrease sufficiently slowly.

Second, we divide the step size into three regimes. For sufficiently small and sufficiently large step sizes, simple constant-gradient constructions give the desired lower bound. The main difficulty is therefore the intermediate regime, where η ranges from order ε^2 to order ε, with $\varepsilon := 1 - \beta$.

Third, in the intermediate regime, we use a helical construction whose primary coordinates rotate by a fixed angle and advance by a fixed amount. For all iterations after initialization, the gradient norm is constant, the function value decreases by a fixed amount at each step, and the required interpolation conditions are satisfied. We then add an auxiliary scalar coordinate to handle the interpolation conditions involving the initial iterate.

Finally, in the intermediate regime, the constructed trajectory has gradient norms of constant size, while its function values decrease only at a rate proportional to ε. Over T iterations, this yields the desired lower bound of order $1/(\varepsilon T)$. We then choose explicit parameters satisfying all conditions of the construction and combine the three step-size regimes to obtain the stated result.

C.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 C.1. *(Rotaru et al., 2024) $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, f_i)\}_{i\in\mathcal{I}}$ 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*

$$\frac{1}{2L}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|_2^2 - \frac{L}{4}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j - \frac{1}{L}(\boldsymbol{g}_i - \boldsymbol{g}_j)\right\|_2^2 \leq f_i - f_j - \langle \boldsymbol{g}_j, \boldsymbol{x}_i - \boldsymbol{x}_j\rangle \tag{31}$$

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

Proposition C.2. *(Drori & Shamir, 2020) Suppose that $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{g}_i, f_i)\}_{i\in\mathcal{I}}$ 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 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

$$\boldsymbol{g}_t = \frac{(1+\beta)\boldsymbol{x}_t - \boldsymbol{x}_{t+1} - \beta\boldsymbol{x}_{t-1}}{\eta}. \tag{32}$$

Lemma C.3. *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 C.2, 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 $f \in \mathcal{F}_1(1/2)$. 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 $f \in \mathcal{F}_1(1/2)$ gives the claim. □

Lemma C.4 (Pairing form of the interpolation conditions)**.** *Let $\{(\boldsymbol{x}_t, \boldsymbol{g}_t)\}_{t=0}^N$ be an arbitrary finite tuple. Define*

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

For $0 \leq i, j \leq N$, 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

$$f_{t+1} - f_t = \langle \boldsymbol{g}_{t+1}, \Delta \boldsymbol{x}_t \rangle - Q_{t,t+1} \tag{33}$$

for $0 \leq t < N$. 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 $0 \leq k < \ell \leq N$,

$$S_{k,\ell}^- := \frac{1}{2} \sum_{k \leq t < s \leq \ell - 1} \langle \Delta \boldsymbol{x}_t - \Delta \boldsymbol{g}_t, \Delta \boldsymbol{x}_s + \Delta \boldsymbol{g}_s \rangle \geq 0 \tag{34}$$

$$S_{k,\ell}^+ := \frac{1}{2} \sum_{k \leq t \leq s \leq \ell - 1} \langle \Delta \boldsymbol{x}_t + \Delta \boldsymbol{g}_t, \Delta \boldsymbol{x}_s - \Delta \boldsymbol{g}_s \rangle \geq 0. \tag{35}$$

Proof. For $0 \le i, j \le N$, the 1-smooth interpolation condition from Proposition C.1 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 $k < \ell$. □

C.3 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 three 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 $0 < \varepsilon \le \varepsilon_{\max}$. The three ranges are

$$0 < \eta \le a\varepsilon^2, \quad a\varepsilon^2 \le \eta \le b\varepsilon, \quad \eta \ge b\varepsilon.$$

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 $a\varepsilon^2 \le \eta \le b\varepsilon$, the primary coordinates rotate and advance by fixed amounts, and an auxiliary scalar coordinate handles the initial interpolation inequalities. This is the only regime that determines the final finite constant.

Regime III. For $\eta \ge 4\varepsilon$, we use a one-dimensional constant-gradient construction whose total negative function-value drift is summable. In the final check, $b \ge 4$, so this proposition covers the entire range $\eta \ge b\varepsilon$.

Combining the three regimes gives a uniform lower bound over all step sizes. The constant-gradient Regimes I and III are handled first, followed by Regime II.

C.4 Regimes I and III

In this subsection, we handle the two constant-gradient regimes.

Proposition C.5 (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 C.3,

$$\sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \le t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{1}{1 + 2a\varepsilon T}.$$

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 C.6 (Regime III)**.** *Fix $0 < \varepsilon < 1$, and let $\beta = 1 - \varepsilon$. Then*

$$\liminf_{T \to \infty} \varepsilon T \inf_{\eta \ge 4\varepsilon} \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \le t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 = +\infty.$$

Proof. Fix $T \ge 1$ and $\eta \ge 4\varepsilon$. Set $g_t \equiv -1$ for every $t \ge 0$ and initialize

$$x_{-1} = x_0 = 0,$$

then generate $\{x_t\}$ by the heavy-ball recursion. For every $j \ge 0$, define

$$d_j := x_j - x_{j-1}.$$

Since $d_0 = 0$, the recursion gives

$$d_{j+1} = \eta + \beta d_j,$$

and hence, for every $j \geq 1$,

$$d_j = \eta \sum_{i=0}^{j-1} \beta^i = \frac{\eta}{\varepsilon}(1-\beta^j) \geq 4(1-\beta^j) > 0. \tag{36}$$

Set $f_0 = 0$ and define, for $1 \leq j < T$,

$$f_j - f_{j-1} = -d_j + \frac{d_j^2}{4}. \tag{37}$$

We first verify interpolation conditions. Since $\Delta g_t = 0$ and $\Delta x_t = d_{t+1} > 0$, Lemma C.4 gives, for every $0 \leq k < \ell < T$,

$$S^-_{k,\ell} = \frac{1}{2} \sum_{k \leq t < s \leq \ell - 1} d_{t+1} d_{s+1} \geq 0,$$
$$S^+_{k,\ell} = \frac{1}{2} \sum_{k \leq t \leq s \leq \ell - 1} d_{t+1} d_{s+1} \geq 0.$$

Indeed, (37) is precisely the recursion (33), because $Q_{j-1,j} = -d_j^2/4$. Thus the tuple is $\mathcal{F}_1$-interpolable.

It remains to control its function values. Let

$$q(d) := -d + \frac{d^2}{4}, \quad q_-(d) := [-q(d)]_+.$$

For $0 \leq d \leq 4$, one has $q_-(d) = d(4-d)/4$, whereas $q_-(d) = 0$ for $d \geq 4$. Consequently, $0 \leq q_-(d) \leq 1$, and q_- is decreasing on $[2, 4]$.

We claim that, for every $j \geq 1$,

$$q_-(d_j) \leq 4\beta^j. \tag{38}$$

Fix $j \geq 1$ and write $r := \beta^j$. If $r \geq 1/2$, then $q_-(d_j) \leq 1 \leq 4r$. If $r < 1/2$, then (36) yields $d_j \geq 4(1-r) > 2$. The claim is immediate if $d_j \geq 4$; otherwise $2 < d_j < 4$, and monotonicity on $[2, 4]$ gives

$$q_-(d_j) \leq q_-(4(1-r)) = 4r(1-r) \leq 4r.$$

This proves (38).

Using (37) and (38), for every $0 \leq t < T$ we obtain

$$f_t = \sum_{j=1}^{t} q(d_j) \geq -\sum_{j=1}^{t} q_-(d_j) \geq -4 \sum_{j=1}^{t} \beta^j \geq -\frac{4\beta}{\varepsilon}.$$

Since $|g_t|^2 = 1$, it follows that

$$G_T^2 = 1, \quad \Delta_T \leq \frac{4\beta}{\varepsilon} + \frac{1}{2}.$$

Moreover, $\Delta_T \geq 1/2$ by considering $t = 0$. Therefore, Lemma C.3 applies and yields, uniformly over $\eta \geq 4\varepsilon$,

$$\sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{1}{8\beta/\varepsilon + 1} = \frac{\varepsilon}{8 - 7\varepsilon} \geq \frac{\varepsilon}{8}.$$

Consequently,

$$\varepsilon T \inf_{\eta \geq 4\varepsilon} \sup_{f \in \mathcal{F}_1(1/2)} \min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq \frac{\varepsilon^2 T}{8},$$

which diverges as $T \to \infty$. □

C.5 Regime II

We now prove the lower bound in Regime II.

Helical trajectory and basic formulas. We identify $\mathbb{R}^2$ with $\mathbb{C}$, and use the real inner product

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

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 (32). Since

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

and $\phi_{t+1} - \phi_t = \delta_{t+1}$, $\phi_t - \phi_{t-1} = \delta_t$, direct substitution gives

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

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 the z-increments are

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

Consequently, their contributions to the 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

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

Indeed,

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

so

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

The xy-contributions are therefore

$$S_{k,\ell}^{-,xy} = \frac{1}{2} \sum_{k\le t<s\le \ell-1} \mathrm{Re}\left(\overline{U_t^-} U_s^+ e^{i(\phi_s - \phi_t)}\right) \tag{39}$$

$$S_{k,\ell}^{+,xy} = \frac{1}{2} \sum_{k\le t\le s\le \ell-1} \mathrm{Re}\left(\overline{U_t^+} U_s^- e^{i(\phi_s - \phi_t)}\right). \tag{40}$$

In particular,

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

We now specialize to constant phase increments. Let $0 < \theta < 2\pi$, $\lambda > 0$, and

$$\boldsymbol{x}_t := (e^{it\theta}, \lambda t\theta) \in \mathbb{C} \times \mathbb{R}. \tag{41}$$

Write $\varepsilon := 1 - \beta$ and define the heavy-ball gradients by (32). The preceding identities with $\delta_t = \theta$ give

$$\boldsymbol{g}_t = \left(Ce^{it\theta}, -\frac{\lambda\varepsilon\theta}{\eta} \right), \quad C := \frac{(1+\beta)(1-\cos\theta) - i\varepsilon\sin\theta}{\eta}. \tag{42}$$

Writing $C = C_r + iC_i$ and $\alpha := 1 - \cos\theta$, we have

$$C_i = -\frac{\varepsilon\sin\theta}{\eta}, \quad G^2 := \|\boldsymbol{g}_t\|_2^2 = |C|^2 + \frac{\lambda^2\varepsilon^2\theta^2}{\eta^2}. \tag{43}$$

The norm is independent of t.

When

$$K := \frac{2}{1+\beta}, \quad \theta^2 = K\eta, \tag{44}$$

we use the notation

$$H := \frac{2(1-\cos\theta)}{\theta^2}, \quad R := \frac{\sin\theta}{\theta}, \quad s := \frac{\varepsilon\theta}{\eta}. \tag{45}$$

Indeed, $1 - \cos\theta = \theta^2 H/2$ and $\sin\theta = \theta R$, so substitution into (42)–(43) gives

$$C_r = H, \quad C_i = -sR, \quad |C|^2 = H^2 + s^2R^2, \quad G^2 = H^2 + s^2(R^2 + \lambda^2). \tag{46}$$

Lemma C.7. *For the trajectory in* (41)*, the ordered sums over an interval of length $n = \ell - k$ depend only on n. 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

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

Proof. Since $\delta_t = \theta$, we have $h_t = C$ 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 the decomposition into xy- and z-components above gives the intermediate forms

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

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

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 C.8. *For the trajectory in* (41)*, 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*

$$M = \frac{1-\cos\theta}{2}(1-|C|^2) - \frac{\varepsilon\sin^2\theta}{\eta} + \lambda^2\theta^2\left(\frac{1}{4} - \frac{\varepsilon}{\eta}\right).$$

Under the scaling (44) *and notation* (45)*, this becomes*

$$M = A_{xy}(\eta,\theta) + K\left(\frac{\eta}{4} - \varepsilon\right)\lambda^2,$$

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.

Under (44), since

$$1 - \cos\theta = \frac{\theta^2}{2}H, \quad \sin\theta = \theta R,$$

substitution of (46) gives the remaining formulas for M.

□

Auxiliary-coordinate lemma. The auxiliary coordinate is built from a scalar sequence that follows a trigonometric formula for finitely many indices and then joins a geometric tail. The following lemma records the identities needed below.

Lemma C.9. *Fix $0 < \beta < 1$, write $\varepsilon := 1 - \beta$, and let $\eta > 0$, $N \geq 1$, and $r \in \{1, \beta\}$. If $r = 1$, assume*

$$N < \frac{\sqrt{\beta}}{1 - \sqrt{\beta}}.$$

Then there exists $\omega \in (0, \pi/(N+1))$ such that

$$\sqrt{\beta}\sin((N+1)\omega) = r\sin(N\omega). \tag{47}$$

For any $A > 0$, define

$$u_t := \begin{cases} A\beta^{(t-N)/2}\dfrac{\sin(t\omega)}{\sin(N\omega)}, & 0 \leq t \leq N, \\ Ar^{t-N}, & t \geq N. \end{cases}$$

Evaluating the first branch at $t = N+1$ gives $u_{N+1} = rA$, which agrees with the second branch, and $u_t > 0$ for all $t \geq 1$. Moreover, with

$$m_t := \frac{u_{t+1} - \beta u_t}{\eta}, \quad p := \frac{(1-\sqrt{\beta})^2 + 2\sqrt{\beta}(1-\cos\omega)}{\eta},$$

and $p_t := p$ for $t < N$ and $p_t := 0$ for $t \geq N$, we have

$$m_{t+1} - m_t = -p_t u_{t+1}. \tag{48}$$

In particular, $m_t \geq m_N$ for all $t \geq 0$, where

$$m_N = \begin{cases} \varepsilon A/\eta, & r = 1, \\ 0, & r = \beta. \end{cases}$$

If $N + 1 \geq \tau/\varepsilon$ for some $\tau > 0$, then

$$0 \leq p \leq \frac{\varepsilon^2}{\eta}\left(\frac{1}{(1+\sqrt{\beta})^2} + \frac{\pi^2}{\tau^2}\right). \tag{49}$$

Finally, if $\Delta x_t = u_{t+1}$ and $g_t := (\beta u_t - u_{t+1})/\eta = -m_t$, then

$$\Delta g_t = p_t u_{t+1}, \quad \Delta x_t \pm \Delta g_t = (1 \pm p_t)u_{t+1}. \tag{50}$$

Proof. For $r \in \{1, \beta\}$, let

$$F_r(\omega) := \sqrt{\beta}\sin((N+1)\omega) - r\sin(N\omega).$$

At the origin,

$$\lim_{\omega\downarrow 0}\frac{F_r(\omega)}{\omega} = \sqrt{\beta}(N+1) - rN.$$

This is positive automatically for $r = \beta$, and for $r = 1$ it is positive by the assumed bound on N. At the other endpoint,

$$F_r\left(\frac{\pi}{N+1}\right) = -r\sin\left(\frac{N\pi}{N+1}\right) < 0.$$

Thus (47) has a root in the stated interval. The boundary equation gives $u_{N+1} = rA$, so the two branches agree at $t = N+1$; positivity follows from $0 < t\omega < \pi$ for $1 \le t \le N$.

For $0 \le t < N$, the trigonometric branch satisfies

$$u_{t+2} - 2\sqrt{\beta}\cos\omega\, u_{t+1} + \beta u_t = 0.$$

Consequently,

$$m_{t+1} - m_t = \frac{u_{t+2} - (1+\beta)u_{t+1} + \beta u_t}{\eta} = -pu_{t+1}.$$

For either choice of r, m_t is constant for $t \ge N$ with the stated value, proving (48). Since $p \ge 0$ and $u_{t+1} > 0$, m_t is non-increasing up to N and constant thereafter, so $m_t \ge m_N$.

Finally,

$$p\eta \le \frac{\varepsilon^2}{(1+\sqrt{\beta})^2} + \sqrt{\beta}\,\omega^2 \le \varepsilon^2\left(\frac{1}{(1+\sqrt{\beta})^2} + \frac{\pi^2}{\tau^2}\right),$$

where we used $2(1-\cos\omega) \le \omega^2$ and $\omega < \pi/(N+1) \le \pi\varepsilon/\tau$. This proves (49); (50) follows immediately from $g_t = -m_t$. □

We now specify the parameter conditions under which these identities will be used.

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 $n \ge 1$,

$$H(u) := \frac{2(1-\cos u)}{u^2} \ge h, \tag{51a}$$

$$R(u) := \frac{\sin u}{u} \ge r, \tag{51b}$$

$$0 \le n\sin u - \sin(nu) \le du^2 n(n+1), \tag{51c}$$

$$1 - H(u)^2 \le \rho u^2. \tag{51d}$$

We note that $H \le 1$ and $R \le 1$ always hold.

We group the remaining assumptions by their role. The range condition is

$$\Theta \le 1. \tag{52a}$$

The interpolation conditions for the primary coordinates are

$$\frac{1-S^2}{2} - 2Sd > 0, \tag{52b}$$

$$S^2 + \rho\Theta^2 < 1. \tag{52c}$$

The function-value slope and coefficient conditions are

$$\frac{b}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) < 1 + r^2, \tag{52d}$$

$$\frac{h^2}{4K_{\max}} \ge c_0. \tag{52e}$$

The auxiliary-coordinate conditions are

$$\frac{1}{a}\left(\frac{1}{(1+\sqrt{1-\varepsilon_{\max}})^2} + \frac{\pi^2}{\tau^2}\right) < 1, \tag{52f}$$

$$\left\lceil \frac{\tau}{\varepsilon_{\max}} \right\rceil \ge 2. \tag{52g}$$

Proposition C.10 (Regime II)**.** *Assume the admissible parameters* (51) *and* (52)*. 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. Under these conditions, we now construct the interpolable tuple.

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 (33), with $\{(\boldsymbol{x}_t^h, \boldsymbol{g}_t^h)\}$.

Let $P_{k,\ell}^{\pm}$ denote the expressions in (34) and (35) computed using only $\{(\boldsymbol{x}_t^h, \boldsymbol{g}_t^h)\}$.

Lemma C.11. *Under* (52a)*,* (52b) *and* (52c)*, 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 C.7, with $\lambda = 1$. Thus, we denote $P_{k,\ell}^{\pm} = S_n^{\pm}$. Then,

$$2S_n^- = \theta^2 \frac{n(n-1)}{2} + 2(1-\cos\theta)(1-|C|^2)A_n^- - 2C_i D_n, \tag{53}$$

$$2S_n^+ = \theta^2 \frac{n(n+1)}{2} + 2(1-\cos\theta)(1-|C|^2)A_n^+ + 2C_i D_n, \tag{54}$$

where

$$C_i = -sR,$$

and A_n^-, A_n^+, and D_n are as in Lemma C.7. 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 (52a), we have $D_n \geq 0$. Also, by (51c),

$$D_n \leq d\theta^2 n(n+1).$$

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

$$2S_n^+ \geq \theta^2 n(n+1)\left(\frac{1-S^2}{2} - 2Sd\right), \tag{55}$$

from (54). By (52b), 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 (53), we get

$$2S_n^- \geq \theta^2 \frac{n(n-1)}{2}\left(1 - S^2 - \rho\Theta^2\right). \tag{56}$$

By (52c), this is nonnegative for $n \geq 2$. For $n = 1$, the sum S_n^- is 0. □

Lemma C.12. *Under* (52d) *and* (52e)*, 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 C.8 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 $0 < H = \frac{2(1-\cos u)}{u^2} \le 1$,

$$0 \le H(1-H^2) \le \frac{2}{3\sqrt{3}}.$$

Also,

$$-Hs^2R^2 \le 0.$$

Thus,

$$B \le 1 + \frac{2}{3\sqrt{3}}.$$

Since $\eta \le b\varepsilon$ and $R \ge r$ by (51b),

$$\frac{M}{K\varepsilon} \le -1 - r^2 + \frac{b}{4}\left(1 + \frac{2}{3\sqrt{3}}\right) = -\tilde{\delta}.$$

Hence $M \le -K\varepsilon\tilde{\delta} < 0$.

For the coefficient, first note that

$$G^2 = H^2 + s^2(R^2+1) \ge H^2 \ge h^2$$

by (51a). Moreover, (52c) implies $S < 1$. Using $H^2, R^2 \le 1$ and $s^2 \le S^2$, we have

$$\begin{aligned} B &= 1 + H(1 - H^2 - s^2R^2) \\ &= 1 + H(1-H^2) - Hs^2R^2 \\ &\ge 1 - Hs^2R^2 \\ &\ge 1 - S^2 > 0. \end{aligned}$$

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 (52e). □

We keep the primary helical coordinate and add one auxiliary scalar coordinate.

Let

$$N := \left\lceil \frac{\tau}{\varepsilon} \right\rceil.$$

By (52g), for every $0 < \varepsilon \le \varepsilon_{\max}$, $N \ge 2$.

Choose

$$\omega \in \left(0, \frac{\pi}{N+1}\right)$$

such that

$$\sin((N+1)\omega) = \sqrt{\beta}\sin(N\omega). \tag{57}$$

The existence of this root follows from Lemma C.9 with $r = \beta$; in this case (47) is equivalent to (57).

Define $\{\gamma_t\}$ by

$$\gamma_t := \begin{cases} \beta^{(t-N)/2}\frac{\sin(t\omega)}{\sin(N\omega)}, & 0 \le t \le N, \\ \beta^{t-N}, & t \ge N. \end{cases}$$

By Lemma C.9, the two branches in the definition agree at $t = N+1$, and $\gamma_t > 0$ for $t \ge 1$.

Define

$$x_t^c := \begin{cases} 0, & t \in \{-1, 0\}, \\ \sum_{j=1}^{t} \gamma_j, & t \geq 1. \end{cases}$$

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 (32). 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

$$m_t^c := \frac{\gamma_{t+1} - \beta\gamma_t}{\eta}.$$

Let

$$p_{\max} := \frac{1}{a}\left(\frac{1}{(1+\sqrt{1-\varepsilon_{\max}})^2} + \frac{\pi^2}{\tau^2}\right).$$

Applying Lemma C.9 with $u_t = \gamma_t$, $A = 1$, and $r = \beta$ gives a sequence p_t, with $p_t = p$ for $t < N$ and $p_t = 0$ for $t \geq N$, such that $m_{t+1}^c - m_t^c = -p_t\gamma_{t+1}$ and $m_t^c = 0$ for $t \geq N$. Since $N + 1 \geq \tau/\varepsilon$, $\eta \geq a\varepsilon^2$, and $\beta \geq 1 - \varepsilon_{\max}$, (49) and (52f) give

$$0 \leq p_t \leq p_{\max} < 1.$$

For the auxiliary scalar coordinate with Γ,

$$\Delta x_t^c = \gamma_{t+1}, \quad g_t^c = -m_t^c, \quad \Delta g_t^c = p_t\gamma_{t+1}.$$

By (50),

$$\Delta x_t^c \pm \Delta g_t^c = (1 \pm p_t)\gamma_{t+1} > 0.$$

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 $\Gamma(1 \pm p_t)\gamma_{t+1} \geq 0$ for all $t \geq 0$, we have

$$C_{k,\ell}^{\pm} \geq 0.$$

For $k = 0$, we also note that

$$C_{0,\ell}^{+} \geq c_+, \quad \ell \geq 1$$
$$C_{0,\ell}^{-} \geq c_-, \quad \ell \geq 2.$$

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 $k \geq 1$, by Lemma C.11,

$$P_{k,\ell}^{\pm} \geq 0.$$

Since $C_{k,\ell}^{\pm} \geq 0$, all pairs of (k, ℓ) with $k \geq 1$ 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 $t \geq 1$, 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

$$|C| \leq H + sR \leq 1 + S,$$

we have

$$|U^-| \leq (2+S)\theta, \quad |U^+| \leq (2+S)\theta.$$

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 (52a),

$$|1 - e^{i\theta}| = 2\sin(\theta/2) \geq \frac{\theta}{2}.$$

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 $\|A_0^h\|_2$ and $\|B_0^h\|_2$. 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 $n := \ell - 1$. From the right-hand sides of (55) and (56), 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 (52c) and (52b).

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

$$P^\pm_{0,\ell} \geq q_\pm n(n \pm 1) - C_\varepsilon (n+1).$$

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,

$$\begin{aligned} [-P^+_{0,\ell}]_+ &\leq [C_\varepsilon (n+1) - q_+ n(n+1)]_+ \\ &\leq [C_\varepsilon (n+1) - q_+ n^2]_+ \\ &= [C_\varepsilon + C_\varepsilon n - q_+ n^2]_+ \\ &\leq C_\varepsilon + \frac{C^2_\varepsilon}{4q_+}. \end{aligned}$$

Define

$$B_\varepsilon := D_\varepsilon^2 + C_\varepsilon + \max\left\{\frac{(C_\varepsilon + q_-)^2}{4q_-}, \frac{C_\varepsilon^2}{4q_+}\right\} < \infty.$$

For $\ell = 1$, we have $P_{0,1}^- = 0$ and

$$[-P_{0,1}^+]_+ \le \frac{1}{2}\|B_0^h\|_2\|A_0^h\|_2 \le \frac{1}{2}D_\varepsilon^2 \le B_\varepsilon.$$

Then, for every $\eta \in [a\varepsilon^2, b\varepsilon]$ and $\ell \ge 1$, we have

$$[-P_{0,\ell}^\pm]_+ \le B_\varepsilon.$$

Choose Γ so that

$$\Gamma^2 \ge \frac{B_\varepsilon}{\min\{c_+, c_-\}}.$$

Then, for every $\ell \ge 1$,

$$S_{0,\ell}^+ = P_{0,\ell}^+ + \Gamma^2 C_{0,\ell}^+ \ge 0,$$

because $C_{0,\ell}^+ \ge c_+$ and $[-P_{0,\ell}^+]_+ \le B_\varepsilon$. Similarly, for every $\ell \ge 2$,

$$S_{0,\ell}^- = P_{0,\ell}^- + \Gamma^2 C_{0,\ell}^- \ge 0,$$

because $C_{0,\ell}^- \ge c_-$ and $[-P_{0,\ell}^-]_+ \le B_\varepsilon$. For $\ell = 1$, we have $S_{0,1}^- = 0$. Hence every interpolation inequality holds.

Gradient norm lower bound. For every $t \ge 1$, we have $\boldsymbol{g}_t = (\boldsymbol{g}_t^h, \Gamma g_t^c)$, so

$$\|\boldsymbol{g}_t\|_2^2 \ge \|\boldsymbol{g}_t^h\|_2^2 = G^2.$$

At $t = 0$, the auxiliary scalar coordinate gives

$$g_0^c = -\frac{\gamma_1}{\eta}.$$

Thus

$$\|\boldsymbol{g}_0\|_2^2 \ge \Gamma^2 \frac{\gamma_1^2}{\eta^2}.$$

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

$$\Gamma^2 \ge \max\left\{\frac{B_\varepsilon}{\min\{c_+, c_-\}}, \frac{G_{\max,\varepsilon}^2 b^2 \varepsilon^2}{\gamma_1^2}\right\}.$$

Then

$$\|\boldsymbol{g}_0\|_2^2 \ge G^2.$$

Therefore, for every T,

$$G_T^2 := \min_{0 \le t < T} \|\boldsymbol{g}_t\|_2^2 \ge G^2.$$

Function values. We define f_t^h and f_t^c as (33), with $f_0^h = f_0^c = 0$, 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 (33) 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 $t \ge 1$, 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 of Proposition C.10. 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. Since $f_0 = 0$ and $G_T^2 \geq G^2 > 0$, we also have $\Delta_T > 0$.

Then,

$$\sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{G^2}{2(C_{\Delta,\varepsilon}+(-M)T)}.$$

Multiplying by εT, we obtain

$$\varepsilon T \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge \frac{G^2}{-2M/\varepsilon}\cdot\frac{1}{1+C_{\Delta,\varepsilon}/((-M)T)}.$$

By Lemma C.12,

$$-M \ge K\varepsilon\tilde{\delta} \ge \varepsilon\tilde{\delta}.$$

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 C.12,

$$\frac{G^2}{-2M/\varepsilon} \ge c_0.$$

Taking the infimum over $a\varepsilon^2\le\eta\le b\varepsilon$ and then taking $\liminf_{T\to\infty}$ gives

$$\liminf_{T\to\infty} \varepsilon T \inf_{a\varepsilon^2\le\eta\le b\varepsilon} \sup_{f\in\mathcal{F}_1(1/2)} \min_{0\le t<T} \|\nabla f(\boldsymbol{x}_t)\|_2^2 \ge c_0.$$

□

C.6 Final Admissibility Check

We choose

$$\varepsilon_{\max}=\frac{1}{6},\quad a=4,\quad b=4,\quad \tau=\frac{5}{3},\quad c_0=\frac{1}{8}.$$

Then

$$K_{\max}=\frac{12}{11},\quad \Theta^2=K_{\max}b\varepsilon_{\max}=\frac{8}{11},\quad S^2=\frac{K_{\max}}{a}=\frac{3}{11}.$$

We take

$$h=\frac{31}{33},\quad r=\frac{29}{33},\quad d=\frac{1}{3},\quad \rho=\frac{1}{6}.$$

We first verify the envelope conditions in (51). For $0<u\le\Theta\le 1$,

$$\frac{\sin u}{u}\ge 1-\frac{u^2}{6},\quad \frac{2(1-\cos u)}{u^2}\ge 1-\frac{u^2}{12},$$

and hence

$$H(u)\ge 1-\frac{\Theta^2}{12}=\frac{31}{33}=h,\quad R(u)\ge 1-\frac{\Theta^2}{6}=\frac{29}{33}=r.$$

Moreover, since $0\le H(u)\le 1$,

$$1-H(u)^2\le 2(1-H(u))\le\frac{u^2}{6}=\rho u^2.$$

For every integer $n\ge 1$,

$$0\le n\sin u-\sin(nu)\le nu-\sin(nu)\le\frac{n^2u^2}{3}\le\frac{u^2n(n+1)}{3}=du^2n(n+1).$$

Thus (51) holds.

The range condition follows from

$$\Theta^2=\frac{8}{11}<1.$$

The interpolation conditions for the primary coordinates follow from

$$\frac{1-S^2}{2}-2Sd=\frac{12-2\sqrt{33}}{33}>0,\quad S^2+\rho\Theta^2=\frac{13}{33}<1.$$

The function-value slope and coefficient conditions follow from

$$1+\frac{2}{3\sqrt{3}}<1+\left(\frac{29}{33}\right)^2, \quad \frac{h^2}{4K_{\max}}=\frac{961}{4752}>\frac{1}{8}=c_0.$$

Finally, the auxiliary-coordinate conditions follow from

$$\frac{1}{4}\left(\frac{1}{(1+\sqrt{5/6})^2}+\frac{9\pi^2}{25}\right)<1, \quad \lceil 10\rceil \geq 2.$$

Therefore, all assumptions of Proposition C.10 are satisfied.

The three regimes meet at their endpoints. Indeed, for every $0<\varepsilon\leq 1/6$,

$$a\varepsilon^2\leq b\varepsilon=4\varepsilon,$$

and Proposition C.6 applies throughout $\eta\geq 4\varepsilon$. Hence

$$0<\eta\leq a\varepsilon^2, \quad a\varepsilon^2\leq\eta\leq b\varepsilon, \quad \eta\geq 4\varepsilon$$

cover all $\eta>0$.

Regime I gives $1/(2a)=1/8$, Regime II gives $c_0=1/8$, and Regime III gives a divergent εT-scaled lower bound. Combining them yields, for every $0<\varepsilon\leq 1/6$,

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

Equivalently, the bound holds for every $\beta\geq 5/6$.

D Proofs of Theorems in Section 5

D.1 Proof Sketch

The proof combines two complementary lower-bound constructions. First, on a translated quadratic, every coordinate moves by exactly η at each iteration. The cases where the momentum becomes zero can be avoided by an arbitrarily small perturbation of the initialization. This yields the first lower bound $\Lambda_1(\eta)$.

Second, we show that, for suitable cycle lengths n, Signum on another quadratic follows a $2n$-periodic trajectory. Such cycles can be constructed for n up to order $(1-\beta)^{-1/2}$, which yields a second lower bound $\Lambda_2(\eta)$ that captures the dependence on β.

Finally, we combine $\Lambda_1(\eta)$ and $\Lambda_2(\eta)$ and minimize the resulting bound uniformly over η. Balancing the relevant terms gives the $(1-\beta)^{-1/4}/\sqrt{T}$ dependence in the theorem. The remaining calculations verify the parameter and constant conditions needed for this argument.

D.2 Proof of Theorem 5.1

Set

$$R := \sqrt{\frac{2\Delta}{Ld}}, \quad s := \sqrt{1-\beta}, \quad \mathcal{P}_T(\eta) := \sup_{f\in\mathcal{F}_L(\Delta)} \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\boldsymbol{x}_t)\|_1.$$

We first show the following lemma.

Lemma D.1. *For every $\eta > 0$,*

$$\mathcal{P}_T(\eta) \geq \Lambda_1(\eta) := Ld \begin{cases} R - \dfrac{T-1}{2}\eta, & \eta \leq \dfrac{R}{T-1}, \\ \dfrac{R^2}{2T\eta} + \dfrac{1}{2}\left(1-\dfrac{1}{T}\right)\eta, & \eta \geq \dfrac{R}{T-1}. \end{cases} \tag{58}$$

Proof. Fix $u > 0$ with $u/\eta \notin \mathbb{Z}$, and let $y_0, \dots, y_{T-1} \in \mathbb{R}$ satisfy

$$y_0 = u, \quad |y_{t+1} - y_t| = \eta, \quad 0 \leq t \leq T-2.$$

For each t, necessarily

$$y_t \in S_t := \{u + (2j-t)\eta : 0 \leq j \leq t\}.$$

Write $u/\eta = k + r$, where $k \in \mathbb{Z}_{\geq 0}$ and $r \in (0,1)$. For $t \leq k$, the element of S_t closest to zero is $u - t\eta$. For $t > k$, it is ηr when $t-k$ is even and $-\eta(1-r)$ when $t-k$ is odd. In each case the stated element has absolute value less than η, whereas every other element differs from it by a nonzero even multiple of η. Thus

$$|y_t| \geq \begin{cases} u - t\eta, & 0 \leq t \leq k, \\ \eta r, & t > k \text{ and } t-k \text{ is even}, \\ \eta(1-r), & t > k \text{ and } t-k \text{ is odd}. \end{cases}$$

If $\eta \leq u/(T-1)$, then $T-1 \leq k$, and summing the first case gives

$$\sum_{t=0}^{T-1} |y_t| \geq Tu - \frac{T(T-1)}{2}\eta.$$

Suppose instead that $\eta > u/(T-1)$, and let $N := T-k$. From time k onward, the lower bounds for $|y_t|$ alternate between ηr and $\eta(1-r)$. Their sum, divided by η, is $\geq (N-1+r^2)/2$. When N is even, the sum is $N/2$; when N is odd, it is $(N-1)/2 + r \geq (N-1+r^2)/2$. Hence

$$\begin{aligned} \frac{1}{\eta}\sum_{t=0}^{T-1} |y_t| &\geq \frac{k(k+1)}{2} + kr + \frac{T-k-1+r^2}{2} \\ &= \frac{(k+r)^2}{2} + \frac{T-1}{2}, \end{aligned}$$

and therefore every such sequence satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}|y_t| \geq \begin{cases} u - \dfrac{T-1}{2}\eta, & \eta \leq \dfrac{u}{T-1}, \\ \dfrac{u^2}{2T\eta} + \dfrac{1}{2}\left(1 - \dfrac{1}{T}\right)\eta, & \eta > \dfrac{u}{T-1}. \end{cases} \tag{59}$$

Now consider the scalar Signum recursion on $q(y) = Ly^2/2$,

$$m_{t+1} = \beta m_t + Ly_t, \quad y_{t+1} = y_t - \eta\,\mathrm{sign}(m_{t+1}), \quad y_0 = u, \quad m_0 = 0.$$

For fixed η, β, and T, let

$$Z := \{u \in \mathbb{R} : m_t = 0 \text{ for some } 1 \leq t \leq T-1\}.$$

The set Z is finite. To see this, suppose that t is the first index at which $m_t = 0$, and fix the preceding signs $(\sigma_1, \ldots, \sigma_{t-1}) \in \{-1, 1\}^{t-1}$. Then

$$y_i = u - \eta\sum_{j=1}^{i}\sigma_j, \quad m_t = L\sum_{i=0}^{t-1}\beta^{t-1-i}y_i.$$

As a function of u, m_t is affine with slope

$$L\sum_{i=0}^{t-1}\beta^{t-1-i} = L\frac{1-\beta^t}{1-\beta},$$

which is strictly positive. Hence each choice of t and its preceding signs determines at most one u for which $m_t = 0$. There are only finitely many such choices.

Fix any $u \in (0, R)$ such that

$$u \notin Z \cup \eta\mathbb{Z},$$

and define

$$f_u(\boldsymbol{x}) := \frac{L}{2}\|\boldsymbol{x} + u\mathbf{1}\|_2^2,$$

where $\mathbf{1} \in \mathbb{R}^d$ denotes the all-ones vector. Since

$$f_u(\mathbf{0}) - f_u^\star = \frac{Ldu^2}{2} < \Delta,$$

we have $f_u \in \mathcal{F}_L(\Delta)$. By translation, the Signum iterates on f_u, initialized at $\boldsymbol{x}_0 = \mathbf{0}$, satisfy

$$\boldsymbol{x}_t + u\mathbf{1} = y_t\mathbf{1}, \quad \|\nabla f_u(\boldsymbol{x}_t)\|_1 = dL|y_t|,$$

where $\{y_t\}$ follows the scalar recursion above. The choice $u \notin Z$ gives

$$|y_{t+1} - y_t| = \eta, \quad 0 \leq t \leq T-2.$$

Applying (59) yields

$$\mathcal{P}_T(\eta) \geq dL \begin{cases} u - \dfrac{T-1}{2}\eta, & \eta \leq \dfrac{u}{T-1}, \\ \dfrac{u^2}{2T\eta} + \dfrac{1}{2}\left(1 - \dfrac{1}{T}\right)\eta, & \eta > \dfrac{u}{T-1}. \end{cases}$$

Such choices of u exist arbitrarily close to R. If $\eta < R/(T-1)$, the first case applies for all $u < R$ sufficiently close to R. If $\eta \geq R/(T-1)$, the second case applies for every $u < R$. Letting $u \uparrow R$ gives (58). □

Lemma D.2. *Let $L, \eta > 0$, $\beta \in (0, 1)$, and $n \in \mathbb{N}$. Define*

$$\delta_n := \frac{1-\beta^n}{(1-\beta)(1+\beta^n)} - \frac{n}{2}.$$

Consider Signum on $q(x) = Lx^2/2$ with

$$x_0 = \frac{\eta n}{2}, \quad m_0 = 0.$$

If $\delta_n > 0$, then for $r \in \{0, \dots, 2n-1\}$ satisfying $t \equiv r \pmod{2n}$,

$$x_t = \begin{cases} \eta\left(\frac{n}{2} - r\right), & 0 \le r \le n, \\ \eta\left(r - \frac{3n}{2}\right), & n < r < 2n. \end{cases} \tag{60}$$

In particular, $x_{t+n} = -x_t$. Moreover, for every $\tau > 1/8$, if $T \ge \tau n$, then

$$\frac{1}{T}\sum_{t=0}^{T-1} |x_t| \ge \left(\frac{1}{4} - \frac{1}{32\tau}\right)\eta n. \tag{61}$$

Proof. Set

$$c_n := \frac{L\eta\delta_n}{1-\beta}, \quad I_n := \left(-c_n(1+\beta^{-n}),\, c_n(\beta^{-2n}-1)\right).$$

We show that if $x_0 = \eta n/2$ and $m_0 \in I_n$, then $m_1, \dots, m_n > 0$, $m_{n+1}, \dots, m_{2n} < 0$, and $m_{2n} \in I_n$.

Suppose that $m_1, \dots, m_{k-1} > 0$ for some $1 \le k \le n$. Then $x_i = \eta(n/2 - i)$ for $0 \le i \le k-1$, and therefore

$$m_k = \beta^k m_0 + L\eta \sum_{i=0}^{k-1} \beta^{k-1-i}\left(\frac{n}{2} - i\right).$$

Evaluating the sum gives

$$m_k = \beta^k m_0 + \frac{L\eta}{1-\beta}\left[\frac{n}{2}(1-\beta^k) - k + \frac{1-\beta^k}{1-\beta}\right].$$

Define

$$\ell_k := -\frac{L\eta\beta^{-k}}{1-\beta}\left[\frac{n}{2}(1-\beta^k) - k + \frac{1-\beta^k}{1-\beta}\right].$$

Then $m_k > 0$ is equivalent to $m_0 > \ell_k$, and

$$\ell_{k+1} - \ell_k = L\eta\beta^{-(k+1)}\left(k - \frac{n}{2}\right), \quad 1 \le k < n.$$

Hence $\max_{1\le k\le n} \ell_k$ is attained at $k = 1$ or $k = n$. Moreover,

$$\begin{aligned} \frac{(1-\beta)(\ell_n - \ell_1)}{L\eta} &= \frac{\beta^{-n}}{2}\left[n(1+\beta^{n-1}) - \sum_{i=0}^{n-1}(\beta^i + \beta^{n-1-i})\right] \\ &= \frac{\beta^{-n}}{2}\sum_{i=0}^{n-1}(1-\beta^i)(1-\beta^{n-1-i}) \ge 0. \end{aligned}$$

Since

$$\frac{n}{2}(1-\beta^n) - n + \frac{1-\beta^n}{1-\beta} = (1+\beta^n)\delta_n,$$

we have

$$\ell_n = -c_n(1+\beta^{-n}).$$

Thus every $m_0 \in I_n$ satisfies $m_0 > \ell_k$ for $1 \le k \le n$. By induction,

$$m_1, \dots, m_n > 0, \quad x_k = \eta\left(\frac{n}{2} - k\right), \quad 0 \le k \le n,$$

and

$$m_n = \beta^n m_0 + c_n(1+\beta^n).$$

Let

$$F(x, m) := (x - \eta\,\mathrm{sign}(\beta m + Lx),\, \beta m + Lx)$$

denote one Signum step. Since $F(-x,-m) = -F(x,m)$ and

$$(x_n, m_n) = \Big(-\frac{\eta n}{2}, m_n\Big) = -\Big(\frac{\eta n}{2}, -m_n\Big),$$

the values $m_{n+1}, \dots, m_{2n}$ are negative if the values of m over the first n steps starting from $(\eta n/2, -m_n)$ are positive. By the argument above, it suffices that

$$-m_n > \ell_n.$$

Using the expression for m_n, this is equivalent to

$$m_0 < c_n(\beta^{-2n} - 1),$$

which is precisely the upper inequality defining I_n. Hence

$$m_{n+1}, \dots, m_{2n} < 0,$$

and $x_{2n} = \eta n/2$.

The momentum after the full cycle is

$$m_{2n} = \beta^{2n} m_0 - c_n(1 - \beta^{2n}).$$

The affine function

$$P_n(m) := \beta^{2n} m - c_n(1 - \beta^{2n})$$

satisfies

$$P_n\big(-c_n(1 + \beta^{-n})\big) = -c_n(1 + \beta^n),$$

and

$$P_n\big(c_n(\beta^{-2n} - 1)\big) = 0.$$

Therefore,

$$P_n(I_n) = \big(-c_n(1 + \beta^n), 0\big) \subset I_n.$$

Since $0 \in I_n$, the same sign pattern repeats on every cycle, proving (60).

It remains to prove (61). The sequence $z_t := |x_t|/\eta$ is n-periodic, with $z_t = |n/2 - t|$ for $0 \le t < n$, and

$$\sum_{t=0}^{n-1} z_t = \begin{cases} n^2/4, & n \text{ is even}, \\ (n^2+1)/4, & n \text{ is odd}. \end{cases}$$

For $S_r := \sum_{t=0}^{r-1} z_t$, $0 \le r < n$, direct summation gives, when $n = 2m$,

$$\begin{aligned} S_r - \frac{rn}{4} &= \begin{cases} \dfrac{r}{2}(m - r + 1), & r \le m+1, \\ \dfrac{1}{2}(r - 2m)(r - m - 1), & r \ge m+1, \end{cases} \\ &\ge -\frac{n^2}{32}, \end{aligned}$$

and, when $n = 2m + 1$,

$$\begin{aligned} S_r - \frac{rn}{4} &= \begin{cases} \dfrac{r}{4}(2m - 2r + 3), & r \le m+1, \\ \dfrac{1}{2}\left(r - \dfrac{6m+5}{4}\right)^2 - \dfrac{4m^2 - 4m - 7}{32}, & r \ge m+1, \end{cases} \\ &\ge -\frac{n^2}{32}. \end{aligned}$$

Writing $T = qn + r$ therefore yields

$$\sum_{t=0}^{T-1} z_t \ge \frac{Tn}{4} - \frac{n^2}{32}.$$

If $T \ge \tau n$, dividing by T proves (61). □

Lemma D.3. *Let $s \in (0,1)$, $\beta = 1 - s^2$, and let $\kappa > 0$ satisfy $\kappa s < 2$. Define*

$$H_\kappa(s) := -\frac{\kappa}{2}\log(1-s^2) - s\,\mathrm{arctanh}\left(\frac{\kappa s}{2}\right).$$

If $H_\kappa(s) > 0$, then

$$\delta_n = \frac{1-\beta^n}{(1-\beta)(1+\beta^n)} - \frac{n}{2} > 0$$

for every integer $1 \le n \le \kappa/s$.

Proof. Let

$$a := -\log\beta$$

and define, for $x \ge 0$,

$$g(x) := \frac{\tanh(ax/2)}{s^2} - \frac{x}{2}.$$

The function g is concave on $[0,\infty)$ and satisfies $g(0) = 0$. Since $\kappa s < 2$, the condition $H_\kappa(s) > 0$ is equivalent to

$$\frac{a\kappa}{2s} > \mathrm{arctanh}\left(\frac{\kappa s}{2}\right),$$

and hence to

$$\tanh\left(\frac{a\kappa}{2s}\right) > \frac{\kappa s}{2}.$$

Thus $g(\kappa/s) > 0$. By concavity,

$$g(x) \ge \frac{sx}{\kappa} g(\kappa/s) > 0, \quad 0 < x \le \frac{\kappa}{s}.$$

Finally,

$$\frac{1-\beta^x}{1+\beta^x} = \tanh\left(\frac{ax}{2}\right),$$

so $g(n) = \delta_n > 0$ for every integer in the stated range. □

Lemma D.4. *For $0 < s \le 3/5$,*

$$H_{7/3}(s) > 0.$$

Proof. Set

$$h(s) := H_{7/3}(s) = -\frac{7}{6}\log(1-s^2) - s\,\mathrm{arctanh}\left(\frac{7s}{6}\right).$$

A direct calculation gives

$$h''(s) = \frac{7s^2(2401s^4 - 2423s^2 + 360)}{3(1-s^2)^2(36-49s^2)^2}.$$

The polynomial

$$p(u) := 2401u^2 - 2423u + 360$$

is strictly decreasing on $[0, 9/25]$, because $p'(u) = 4802u - 2423 < 0$ there, and

$$p(0) = 360, \quad p(9/25) = -\frac{125694}{625}.$$

Thus h'' is first positive and then negative on $(0, 3/5)$. Since $h'(0) = 0$ and

$$h'(3/5) = \frac{665}{816} - \mathrm{arctanh}(7/10) < 0,$$

the function h first increases and then decreases on $[0, 3/5]$.

It remains to check the right endpoint. We have

$$30h(3/5) = \log\left[\left(\frac{25}{16}\right)^{35}\left(\frac{3}{17}\right)^{9}\right] > 0.$$

Hence $h(3/5) > 0$. Together with $h(0) = 0$, this proves $h(s) > 0$ for $0 < s \le 3/5$.

□

Proof of Theorem 5.1. We first derive a lower bound with $\tau = 3$ for general choices of κ and α. Let

$$\kappa > 0, \quad \alpha \in (0, 1).$$

We first derive the bound under the following conditions:

$$\kappa s < 2, \quad H_\kappa(s) > 0, \quad T \geq \frac{3\kappa}{s}, \quad T - 1 \geq \frac{\kappa}{2s}. \tag{62}$$

We verify these conditions below for a specific choice of κ and α.

For a fixed $\eta > 0$, define

$$X_\eta := \min\left\{\frac{2R}{\eta}, \frac{\kappa}{s}\right\}.$$

If $X_\eta \geq 1$, choose $n := \lfloor X_\eta \rfloor$. Then

$$1 \leq n \leq \frac{2R}{\eta}, \quad n \leq \frac{\kappa}{s}, \quad T \geq 3n.$$

Consider the translated quadratic

$$f(\boldsymbol{x}) := \frac{L}{2}\left\|\boldsymbol{x} + \frac{\eta n}{2}\mathbf{1}\right\|_2^2.$$

It belongs to $\mathcal{F}_L(\Delta)$, because

$$\begin{aligned} f(\mathbf{0}) - f^\star &= \frac{Ld\eta^2 n^2}{8} \\ &\leq \frac{Ld\eta^2}{8}\left(\frac{2R}{\eta}\right)^2 \\ &\leq \Delta. \end{aligned}$$

After translating by $\eta n\mathbf{1}/2$, every coordinate follows the scalar recursion in Lemma D.2. The first two conditions in (62) and Lemma D.3 ensure that $\delta_n > 0$. Hence, applying Lemma D.2 with $\tau = 3$ gives

$$\mathcal{P}_T(\eta) \geq \frac{23}{96}Ld\eta n \geq \frac{23}{96}Ld\eta(X_\eta - 1).$$

If $X_\eta < 1$, the expression on the right is negative, and the same inequality follows from $\mathcal{P}_T(\eta) \geq 0$. Therefore, for every $\eta > 0$,

$$\mathcal{P}_T(\eta) \geq \Lambda_2(\eta) := \frac{23}{96}Ld\eta\left(\min\left\{\frac{2R}{\eta}, \frac{\kappa}{s}\right\} - 1\right). \tag{63}$$

Let

$$\eta_0 := \frac{R}{T-1}, \quad \eta_c := \frac{2Rs}{\kappa}.$$

The last condition in (62) ensures that $\eta_0 \leq \eta_c$. Define

$$A := -\frac{\alpha}{2}(T-1) + \frac{23}{96}(1-\alpha)\left(\frac{\kappa}{s} - 1\right),$$

$$C := \frac{\alpha}{2}\left(1 - \frac{1}{T}\right) + \frac{23}{96}(1-\alpha)\left(\frac{\kappa}{s} - 1\right), \quad D := \frac{\alpha}{2}\left(1 - \frac{1}{T}\right) - \frac{23}{96}(1-\alpha).$$

By Lemma D.1 and (63),

$$\mathcal{P}_T(\eta) \geq W(\eta) := \alpha\Lambda_1(\eta) + (1-\alpha)\Lambda_2(\eta),$$

where

$$\frac{W(\eta)}{Ld} = \begin{cases} \alpha R + A\eta, & 0 < \eta \leq \eta_0, \\ \dfrac{\alpha R^2}{2T\eta} + C\eta, & \eta_0 \leq \eta \leq \eta_c, \\ \dfrac{\alpha R^2}{2T\eta} + \dfrac{23(1-\alpha)}{48}R + D\eta, & \eta \geq \eta_c. \end{cases}$$

Hence, if

$$C > 0, \quad D > 0, \quad A \le 0, \quad \frac{23(1-\alpha)}{48} + 2\sqrt{\frac{2\alpha D}{T}} - \frac{\alpha\kappa}{sT} \ge 0, \tag{64}$$

then

$$\mathcal{P}_T(\eta) \ge LdR\sqrt{\frac{2\alpha C}{T}} \quad \text{for every } \eta > 0. \tag{65}$$

Indeed, AM–GM gives the claimed bound on the middle branch. The first branch is nonincreasing and is therefore bounded below by its value at η_0. For the last branch, AM–GM gives

$$W(\eta) \ge LdR\left(\frac{23(1-\alpha)}{48} + \sqrt{\frac{2\alpha D}{T}}\right).$$

Since $C - D = 23(1-\alpha)\kappa/(96s)$, the last condition in (64) implies, after squaring, that this is at least the right-hand side of (65).

We now verify these conditions under the assumptions of Theorem 5.1. Choose

$$\kappa = \frac{7}{3}, \quad \alpha = \frac{3}{4}.$$

Since $\beta > 16/25$, we have $0 < s < 3/5$, and therefore $\kappa s < 2$. Moreover, Lemma D.4 gives $H_\kappa(s) > 0$.

The horizon assumption $T \ge 20 + 1/s^2$ implies

$$T - \frac{3\kappa}{s} \ge 20 + \frac{1}{s^2} - \frac{7}{s} = \frac{20s^2 - 7s + 1}{s^2} > 0,$$
$$T - 1 - \frac{\kappa}{2s} \ge 19 + \frac{1}{s^2} - \frac{7}{6s} = \frac{19s^2 - \frac{7}{6}s + 1}{s^2} > 0.$$

Both numerators are positive because their discriminants are negative. Hence (62) holds. Also,

$$D = \frac{121}{384} - \frac{3}{8T} > 0, \tag{66}$$
$$C = D + \frac{161}{1152s} > 0. \tag{67}$$

The condition $A \le 0$ in (64) follows directly from $T - 1 \ge \kappa/(2s)$ and $23/96 \le 3/4$:

$$-\frac{3}{8}(T-1) + \frac{23}{384}\left(\frac{\kappa}{s} - 1\right) \le -\frac{3\kappa}{16s} + \frac{23\kappa}{384s} \le 0.$$

It remains to check the last condition in (64). After dividing that condition by 2 and substituting the chosen parameters, it is equivalent to

$$\frac{23}{384} + \sqrt{\frac{3}{2T}\left(\frac{121}{384} - \frac{3}{8T}\right)} - \frac{7}{8sT} \ge 0. \tag{68}$$

Put $r := 1/T$. The horizon condition $T \ge 20 + 1/s^2$ gives

$$0 < r \le r_s := \frac{s^2}{1 + 20s^2} < \frac{1}{20}.$$

Define

$$E_s(r) := \frac{23}{384} + \sqrt{\frac{3r}{2}\left(\frac{121}{384} - \frac{3r}{8}\right)} - \frac{7r}{8s}.$$

The square-root term is concave in r on $[0, r_s]$. Thus E_s is concave, so it suffices to check the two endpoints. Clearly,

$$E_s(0) = \frac{23}{384} > 0.$$

At the other endpoint,

$$E_s(r_s) = \frac{23(1+20s^2) + 24s\sqrt{121+2276s^2} - 336s}{384(1+20s^2)}$$
$$\geq \frac{23 + 460s^2 - 72s}{384(1+20s^2)} > 0,$$

where we used $\sqrt{121+2276s^2} \geq 11$. This proves (68).

We have therefore established, for every $\eta > 0$,

$$\mathcal{P}_T(\eta) \geq LdR\sqrt{\frac{3C}{2T}}.$$

Using $R^2 = 2\Delta/(Ld)$ and (67), this becomes

$$\mathcal{P}_T(\eta) \geq \sqrt{\frac{3\Delta Ld}{2T}\left(\frac{121}{192} + \frac{161}{576s} - \frac{3}{4T}\right)}. \tag{69}$$

Since $T \geq 20$,

$$\frac{121}{192} - \frac{3}{4T} \geq \frac{7}{12}.$$

Furthermore, $\frac{161}{576} > \frac{5}{18}$. Taking the infimum of (69) over $\eta > 0$ and substituting $s = \sqrt{1-\beta}$ therefore implies the stated lower bound. □

D.3 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 ℓ_1-norm of the gradient as a function of η.

While our analysis establishes a worst-case lower-bound separation that scales as $\Omega((1-\beta)^{-1/4})$, it does not match the $\mathcal{O}((1-\beta)^{-1/2})$ dependence suggested by the optimized upper bound in Proposition 3.4. 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.

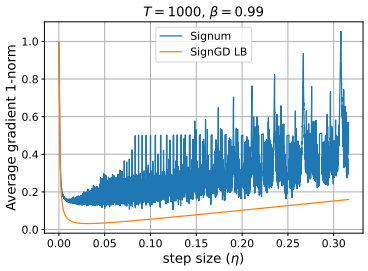

Figure 2: η vs average gradient ℓ_1-norm on $f(x) = \frac{1}{2}x^2$ with $x_0 = 1$, $T = 1000$, $\beta = 0.99$. The SignGD lower bound from (58) is shown for comparison.

This sensitivity also makes stochastic extensions nontrivial. Our proof of Theorem 5.1 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.

E AUXILIARY PROOFS

E.1 AN EQUALITY CASE FOR PROPOSITION 3.1

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 $g(\boldsymbol{x}; \boldsymbol{\xi}) = L\boldsymbol{x} + \boldsymbol{\xi}$, where $\mathbf{1}_d \in \mathbb{R}^d$ denotes the all-ones vector and

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

Then, Assumptions 2.3 and 2.4 are satisfied.

We also have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &= f(\boldsymbol{x}_t - \eta \boldsymbol{g}_t) \\ &= f(\boldsymbol{x}_t) - \eta \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{g}_t + \frac{\eta^2 L}{2}\|\boldsymbol{g}_t\|_2^2. \end{aligned} \tag{70}$$

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 (70) 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,

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

□

E.2 COMPARISON OF SGD AND SHB UPPER BOUNDS

The upper bound in Proposition 3.2 is minimized at

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

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

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

then the minimized upper bound becomes

$$\sigma\sqrt{\frac{5\beta^2+9\beta+8}{\beta+1}\frac{\Delta L}{T}}.$$

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 $T \to \infty$.

E.3 Proof of Proposition 3.3

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

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

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

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

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

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

Dividing by ηT:

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

By minimizing the right-hand side, we have

$$\frac{\Delta}{\eta T} + \frac{Ld}{2}\eta \ge \sqrt{\frac{2\Delta L d}{T}}.$$

□

E.4 Asymptotic Tightness of Proposition 3.3

Assume $T > 2$. Applying Lemma D.1, we have

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

where

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

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

$$\frac{1}{2}\sqrt{2\Delta L d}.$$

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 lower bound and the corresponding minimized upper bound (each optimized with respect to η) is $\geq\sqrt{1-\frac{1}{T}}$.

E.5 Proof of Proposition 3.4

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

$$\begin{aligned}
f(\boldsymbol{x}_{t+1})&\leq f(\boldsymbol{x}_t)-\eta\nabla f(\boldsymbol{x}_t)^\top\mathrm{sign}(\boldsymbol{m}_{t+1})+\frac{L}{2}\eta^2\|\mathrm{sign}(\boldsymbol{m}_{t+1})\|_2^2\\
&\leq f(\boldsymbol{x}_t)-\eta\nabla f(\boldsymbol{x}_t)^\top\mathrm{sign}(\nabla f(\boldsymbol{x}_t))\\
&\quad+\eta\nabla f(\boldsymbol{x}_t)^\top(\mathrm{sign}(\nabla f(\boldsymbol{x}_t))-\mathrm{sign}(\boldsymbol{m}_{t+1}))+\frac{\eta^2Ld}{2}\\
&=f(\boldsymbol{x}_t)-\eta\|\nabla f(\boldsymbol{x}_t)\|_1+\eta\nabla f(\boldsymbol{x}_t)^\top(\mathrm{sign}(\nabla f(\boldsymbol{x}_t))-\mathrm{sign}(\boldsymbol{m}_{t+1}))+\frac{\eta^2Ld}{2}. \qquad (71)
\end{aligned}$$

Now, we show that $x(\mathrm{sign}(x)-\mathrm{sign}(y))\leq 2|x-y|$ for any $x,y\in\mathbb{R}$. If $\mathrm{sign}(x)=\mathrm{sign}(y)$, then $\mathrm{sign}(x)-\mathrm{sign}(y)=0$ and the inequality holds. Otherwise, we have $xy\leq 0$, which implies $|x-y|\geq|x|$. Since $1-\beta>0$, we have $\mathrm{sign}(\boldsymbol{m}_{t+1})=\mathrm{sign}((1-\beta)\boldsymbol{m}_{t+1})$, and thus

$$\begin{aligned}
\nabla f(\boldsymbol{x}_t)^\top(\mathrm{sign}(\nabla f(\boldsymbol{x}_t))-\mathrm{sign}(\boldsymbol{m}_{t+1}))&\leq 2\|\nabla f(\boldsymbol{x}_t)-(1-\beta)\boldsymbol{m}_{t+1}\|_1\\
&\leq 2\sqrt{d}\|\nabla f(\boldsymbol{x}_t)-(1-\beta)\boldsymbol{m}_{t+1}\|_2.
\end{aligned}$$

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

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

Dividing by ηT, we obtain

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

By the L-smoothness, we have

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

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

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

Minimizing the right-hand side gives

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

□

E.6 Proof of Muon Upper Bound

We now extend the argument of Proposition 3.4 to Muon. Let $r := \min\{m, n\}$.

Proposition E.1. *Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L-smooth. Let $\{\boldsymbol{X}_t\}_{t=0}^{T-1}$ denote the iterates generated by Muon, with $\sigma = 0$. Then, with $\Delta = f(\boldsymbol{X}_0) - f^\star$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{X}_t)\|_* \le \frac{\Delta}{\eta T} + \frac{Lr}{2}\left(\frac{1+3\beta}{1-\beta}\right)\eta + \frac{2\sqrt{2\Delta Lr}}{T}\frac{\beta}{1-\beta}. \tag{72}$$

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

Proof. Proposition E.1 is a zero initial momentum variant of the exact-SVD Muon bound in Shen et al. (2026, Theorem 4.1). While Shen et al. (2026) initialize the momentum with the initial gradient, we use $\boldsymbol{M}_0 = \boldsymbol{0}_{m\times n}$. To account for this difference, we use the corresponding zero-initialized exact-SVD analysis in Kim & Oh (2026, Appendix C.1), whose momentum recursion matches ours up to indexing. We specialize an intermediate bound from their proof to our setting.

Let $\|\cdot\|_{\mathrm{op}}$ denote the spectral norm. By the norm inequalities $\|\boldsymbol{A}\|_* \le \sqrt{r}\|\boldsymbol{A}\|_{\mathrm{F}}$ and $\|\boldsymbol{A}\|_{\mathrm{F}} \le \sqrt{r}\|\boldsymbol{A}\|_{\mathrm{op}}$, our L-smoothness assumption implies

$$\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\|_* \le \sqrt{r}\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\|_{\mathrm{F}} \le \sqrt{r}L\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}} \le rL\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{op}}.$$

Thus, substituting $\sigma = 0$ into the bound obtained in the proof of Kim & Oh (2026, Theorem 4) and using the smoothness constant rL gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{X}_t)\|_* \le \frac{\Delta}{\eta T} + \frac{rL\eta}{2} + \frac{2\beta rL\eta}{1-\beta} + \frac{2}{T}\sum_{t=1}^{T}\beta^t\|\nabla f(\boldsymbol{X}_0)\|_*.$$

Moreover,

$$\|\nabla f(\boldsymbol{X}_0)\|_* \le \sqrt{r}\|\nabla f(\boldsymbol{X}_0)\|_{\mathrm{F}} \le \sqrt{2\Delta Lr},$$

where the second inequality follows from L-smoothness. Since $\sum_{t=1}^{T}\beta^t \le \frac{\beta}{1-\beta}$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{X}_t)\|_* \le \frac{\Delta}{\eta T} + \frac{Lr}{2}\left(\frac{1+3\beta}{1-\beta}\right)\eta + \frac{2\sqrt{2\Delta Lr}}{T}\frac{\beta}{1-\beta}.$$

Minimizing the right-hand side over $\eta > 0$ yields the stated result. □

E.7 Proof of Proposition 6.1

Proof. We prove that both $\boldsymbol{X}_t$ and $\boldsymbol{M}_t$ are diagonal by induction on t.

Base case. Both $\boldsymbol{X}_0$ and $\boldsymbol{M}_0$ are diagonal by initialization.

Inductive step. Since $\boldsymbol{X}_t$ and $\boldsymbol{M}_t$ are diagonal by the inductive hypothesis, and $\boldsymbol{X}' = \operatorname{diag}(\boldsymbol{x}')$ is diagonal, so is

$$\boldsymbol{G}_t = \nabla f(\boldsymbol{X}_t) = L(\boldsymbol{X}_t + \boldsymbol{X}').$$

Thus, $\boldsymbol{M}_{t+1}$ is diagonal. Let $\boldsymbol{M}_{t+1} = \operatorname{diag}(m_1^{(t+1)}, \dots, m_d^{(t+1)})$ and $I_{t+1} := \{i : m_i^{(t+1)} \neq 0\}$. Then, a compact SVD of $\boldsymbol{M}_{t+1}$ is given by

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= [\boldsymbol{e}_i]_{i \in I_{t+1}}, \\
\boldsymbol{S}_{t+1} &= \operatorname{diag}\left(|m_i^{(t+1)}|\right)_{i \in I_{t+1}}, \\
\boldsymbol{V}_{t+1} &= [\operatorname{sign}(m_i^{(t+1)})\boldsymbol{e}_i]_{i \in I_{t+1}}.
\end{aligned}$$

Consequently,

$$\boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^\top = \operatorname{diag}\left(\operatorname{sign}(m_1^{(t+1)}), \dots, \operatorname{sign}(m_d^{(t+1)})\right).$$

Thus, $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta \boldsymbol{U}_{t+1}\boldsymbol{V}_{t+1}^\top$ is also diagonal.

Now, let

$$\begin{aligned}
\boldsymbol{M}_t &= \operatorname{diag}(m_1^{(t)}, \dots, m_d^{(t)}), \\
\boldsymbol{X}_t &= \operatorname{diag}(x_1^{(t)}, \dots, x_d^{(t)}), \\
\boldsymbol{X}' &= \operatorname{diag}(x'_1, \dots, x'_d).
\end{aligned}$$

Then, for any $i \in \{1, \dots, d\}$ and $t \in \{0, \dots, T-1\}$, we have

$$\begin{aligned}
m_i^{(t+1)} &= \beta m_i^{(t)} + L\big(x_i^{(t)} + x'_i\big), \\
x_i^{(t+1)} &= x_i^{(t)} - \eta \operatorname{sign}(m_i^{(t+1)}).
\end{aligned}$$

This is exactly the Signum update in Algorithm 1 on the function $\frac{L}{2}\|\boldsymbol{x} + \boldsymbol{x}'\|_2^2$, with initial point $(x_1^{(0)}, \dots, x_d^{(0)}) = \boldsymbol{x}_0$ and zero initial momentum. Therefore, the diagonal entries of $\boldsymbol{X}_t$ and $\boldsymbol{M}_t$ coincide with the Signum iterates $\boldsymbol{x}_t$ and $\boldsymbol{m}_t$, respectively, for every t, which proves $\boldsymbol{X}_t = \operatorname{diag}(\boldsymbol{x}_t)$ and $\boldsymbol{M}_t = \operatorname{diag}(\boldsymbol{m}_t)$. □

F 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 (Drori & Teboulle, 2014), 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 & Shamir (2020).

All PEP instances are implemented using the PEPit toolbox (Goujaud et al., 2024), and the resulting semidefinite programs are solved using the MOSEK solver (ApS, 2025).

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 $(T + 2)$ (Taylor et al., 2017). We analyze specific worst-case instances for $T = 50$ and $\beta = 0.95$. 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, $\boldsymbol{X} = [\boldsymbol{x}_0 \quad \cdots \quad \boldsymbol{x}_{T-1}]^\top$ and $\boldsymbol{G} = [\nabla f(\boldsymbol{x}_0) \quad \cdots \quad \nabla f(\boldsymbol{x}_{T-1})]^\top$ 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 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 4.4.

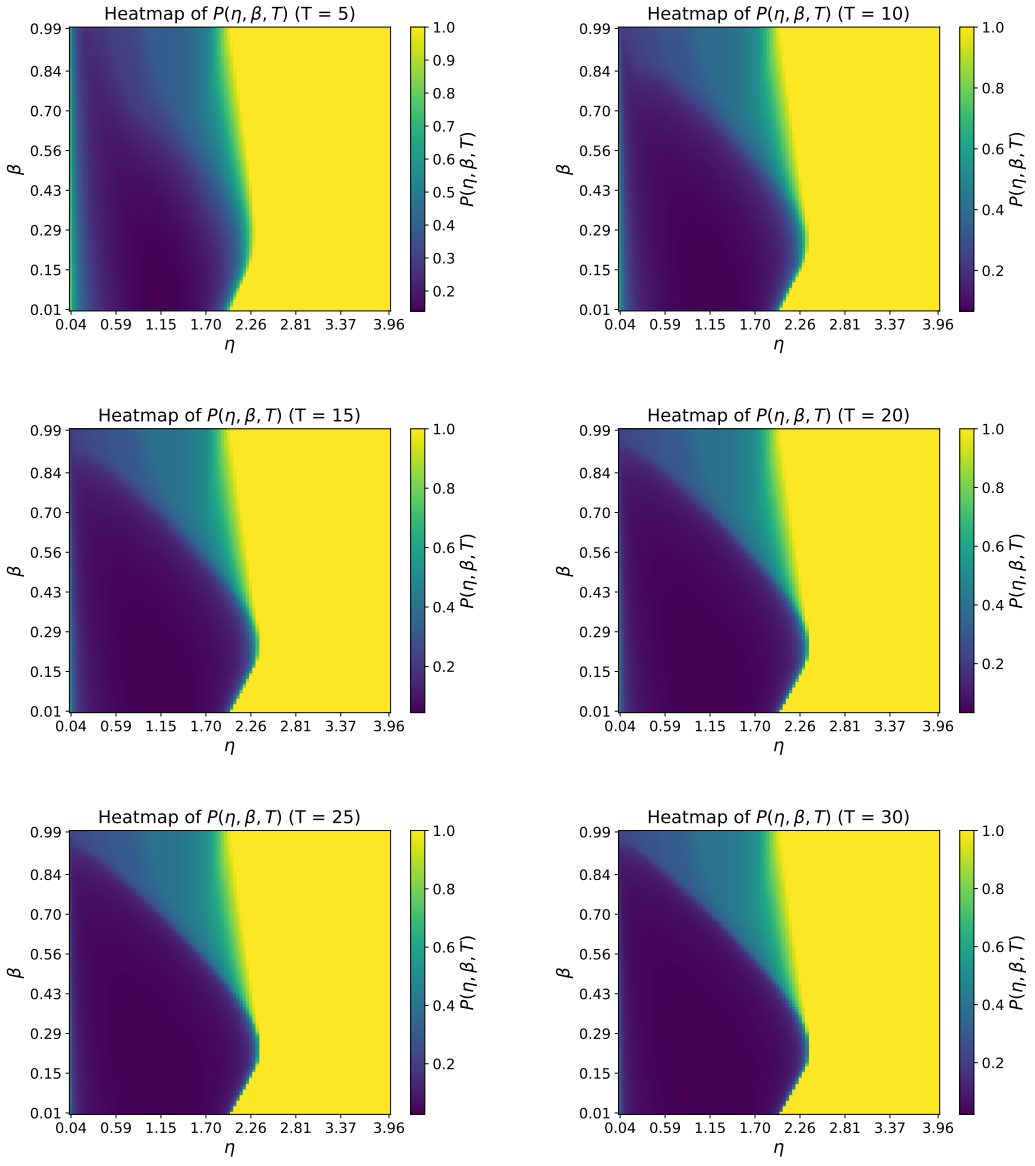

Figure 3: PEP results in terms of best-iterate squared gradient ℓ_2-norm ($T \in \{5, 10, 15, 20, 25, 30\}$).

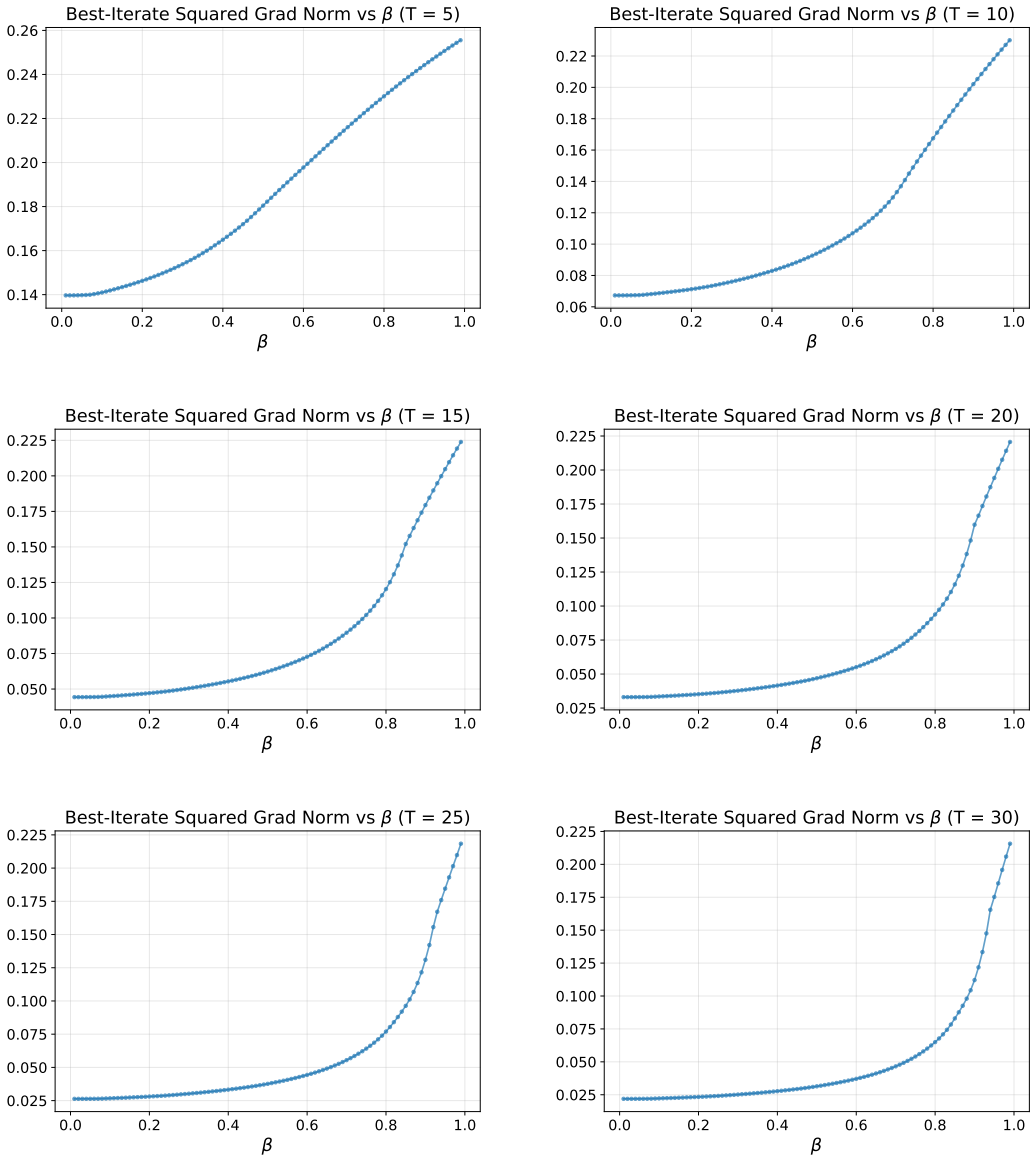

Figure 4: Minimized worst-case value $\inf_\eta P(\eta, \beta, T)$ as a function of the momentum parameter β for different values of T.

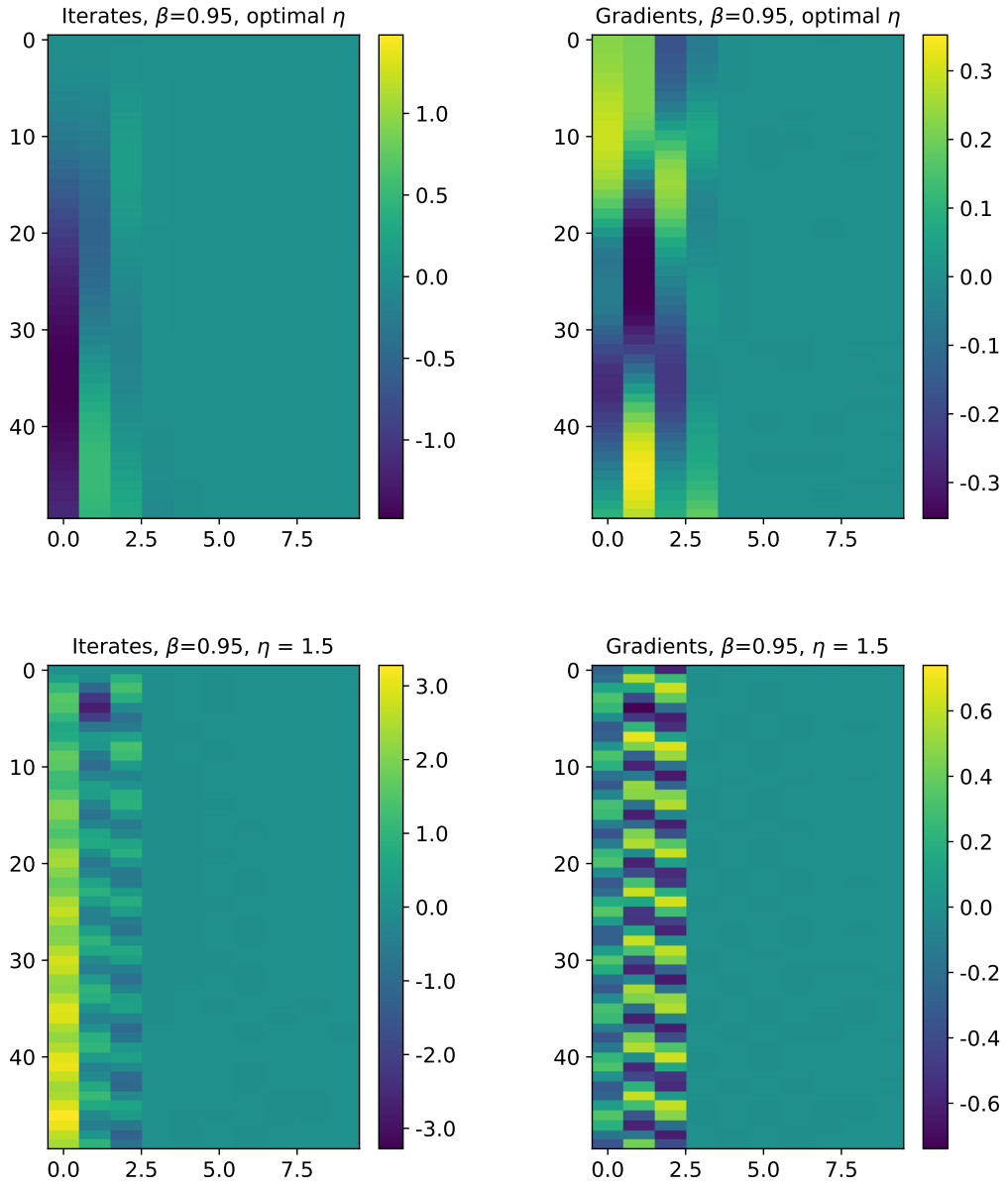

Figure 5: Heatmaps of representative PEP worst-case instances for $T = 50$ and $\beta = 0.95$. We compare the instance obtained at the optimized step size $\eta^\star \approx 0.02348$ with the instance obtained at the larger step size $\eta = 1.5$. The plots show the iterate matrix $\boldsymbol{X}$, 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.

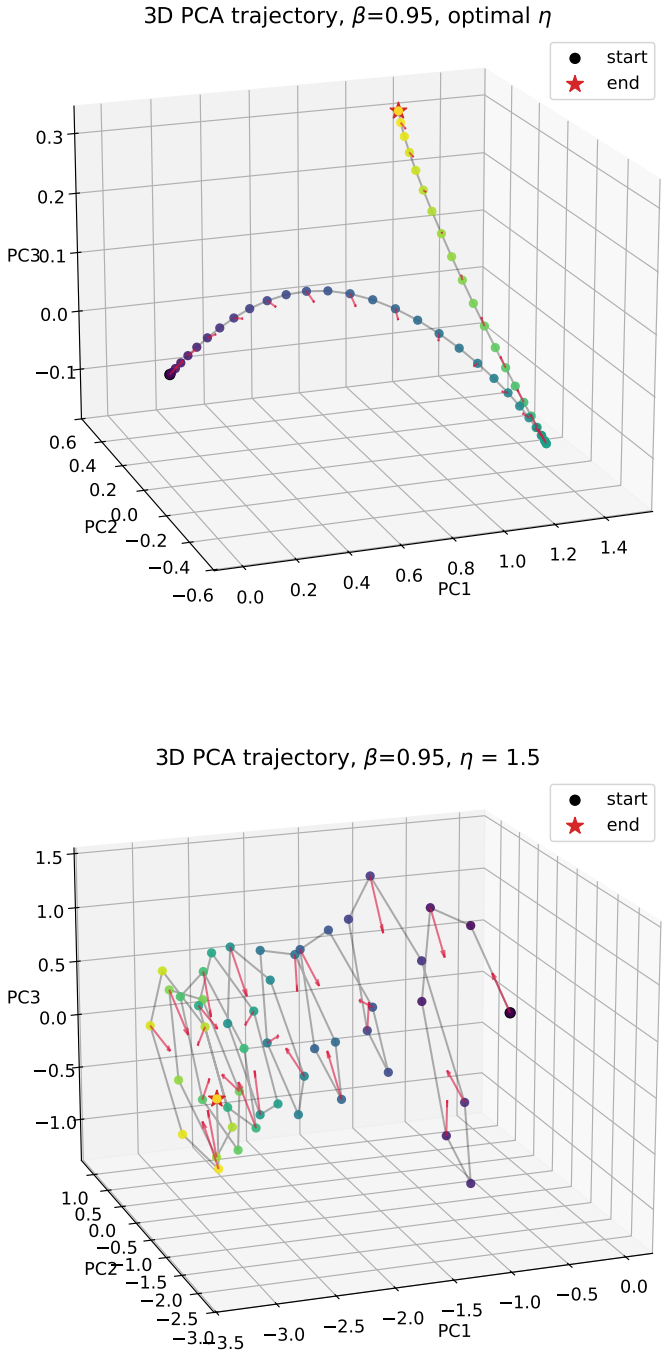

Figure 6: Trajectories of the worst-case iterates for $T = 50$ and $\beta = 0.95$, projected onto the first three principal components for $\eta = \eta^\star \approx 0.02348$ and $\eta = 1.5$, with arrows indicating the *negative* gradient directions at each iterate.

G Details and Results of Deep Learning Experiments

We compare methods with and without Polyak-style momentum in two settings: ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky et al., 2009), and a nanoGPT-style Transformer (Karpathy, 2022) on character-level Tiny Shakespeare (Karpathy, 2015). We report training cross-entropy and averaged gradient norms, selecting the learning rate separately for each momentum coefficient and metric.

G.1 Setup

CIFAR-10. We train ResNet-18 on CIFAR-10 without data augmentation. For minibatch training, we compare SGD and SHB with batch sizes 64, 128, and 256 for $30{,}000$ updates using three seeds. Loss and gradient norms are evaluated on a fixed class-balanced subset of $5{,}000$ training images every 100 updates. Minibatch updates use the full CIFAR-10 training set, while the fixed $5{,}000$-image subset is used for evaluation. We also run full-batch GD and HB for $1{,}024$ updates on the same fixed $5{,}000$-image objective using five seeds.

Tiny Shakespeare. We train a nanoGPT-style character-level Transformer with context length 256. We compare Adam (Kingma & Ba, 2014), Signum (Bernstein et al., 2018), and Muon (Jordan et al., 2024) using minibatch sizes 128 and 256 for $3{,}000$ updates, and also use a full-batch objective consisting of $3{,}921$ fixed non-overlapping contexts for 500 updates. Minibatch gradient norms are evaluated on $1{,}024$ fixed training contexts every 10 updates. All Tiny Shakespeare experiments use three seeds.

Optimizers. All methods use a constant learning rate and $\beta \in \{0, 0.5, 0.9\}$. For Adam, β denotes the first-moment coefficient, with $\beta_2 = 0.999$. Signum uses the standard heavy-ball momentum buffer followed by a sign update. Muon applies the momentum update followed by Newton–Schulz orthogonalization to parameters with at least two dimensions; vector and scalar parameters are updated by SGD with momentum. We use no weight decay, and dropout is set to zero. Experiments were run on NVIDIA RTX 3090 GPUs.

Learning rates. We search learning rates of the form $\eta = 0.1 \times 2^k$, with the exponent ranges in Table 1. The learning rate is selected separately for each setting, optimizer, β, and metric. For training loss, we select the learning rate with the lowest mean loss over the final 10% of updates; for the gradient-norm criterion, we select the learning rate with the lowest average gradient norm over the recorded states.

Table 1: Learning-rate grids for each experiment.

Setting	Optimizer	Exponents k
CIFAR-10, minibatch size 64	SGD/SHB	$-6, \dots, 8$
CIFAR-10, minibatch size 128 or 256	SGD/SHB	$-4, \dots, 8$
CIFAR-10, full-batch size 5,000	GD/HB	$-6, \dots, 8$
Tiny Shakespeare, minibatch size 128 or 256	Adam, Signum	$-12, \dots, 1$
Tiny Shakespeare, minibatch size 128 or 256	Muon	$-9, \dots, 0$
Tiny Shakespeare, full-batch size 3,921	Adam, Signum, Muon	$-13, \dots, 8$

Metrics. We report training cross-entropy and cumulative averages of gradient norms. For SGD/SHB, we use the squared gradient ℓ_2-norm; for Adam and Signum, the gradient ℓ_1-norm; and for Muon, the Frobenius norm of the gradients of matrix parameters, which serves as a computationally efficient surrogate for the nuclear norm used in our theory. Learning rates are selected separately for each reported metric. Curves show the mean across seeds, with shading indicating the seed-wise minimum–maximum range.

G.2 Results

Figures 7–10 show the resulting comparisons. Across most settings, the momentum variants attain lower loss and gradient norms, although the best momentum coefficient depends on the optimizer, batch size, and metric. In particular, for Signum with minibatch size 256, the gradient comparison

Table 2: Selected learning-rate exponents. Each entry gives the exponents selected using training loss and the gradient norm, respectively, for $\eta = 0.1 \times 2^k$.

Setting	Optimizer	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$
CIFAR-10, minibatch 64	SGD/SHB	$(0, 6)$	$(-1, 7)$	$(-4, 4)$
CIFAR-10, minibatch 128	SGD/SHB	$(4, 7)$	$(1, 7)$	$(-1, 6)$
CIFAR-10, minibatch 256	SGD/SHB	$(5, 7)$	$(1, 6)$	$(0, 5)$
CIFAR-10, full-batch 5,000	GD/HB	$(3, 1)$	$(1, -1)$	$(0, -1)$
Tiny Shakespeare, minibatch 128	Adam	$(-9, -2)$	$(-7, -1)$	$(-7, -1)$
	Signum	$(-8, -2)$	$(-8, -2)$	$(-10, -4)$
	Muon	$(-5, -2)$	$(-6, -3)$	$(-7, -4)$
Tiny Shakespeare, minibatch 256	Adam	$(-9, -2)$	$(-7, -2)$	$(-7, -2)$
	Signum	$(-8, -2)$	$(-9, -2)$	$(-10, -4)$
	Muon	$(-5, -2)$	$(-5, -3)$	$(-7, -4)$
Tiny Shakespeare, full-batch 3,921	Adam	$(-8, -2)$	$(-8, -1)$	$(-10, 0)$
	Signum	$(-8, -1)$	$(-8, -1)$	$(-9, -2)$
	Muon	$(-2, 2)$	$(-3, -3)$	$(-6, -4)$

favors $\beta = 0$, while the loss comparison favors momentum. The full-batch experiments also show favorable comparisons for momentum, indicating that the empirical benefit is not limited to minibatch stochasticity.

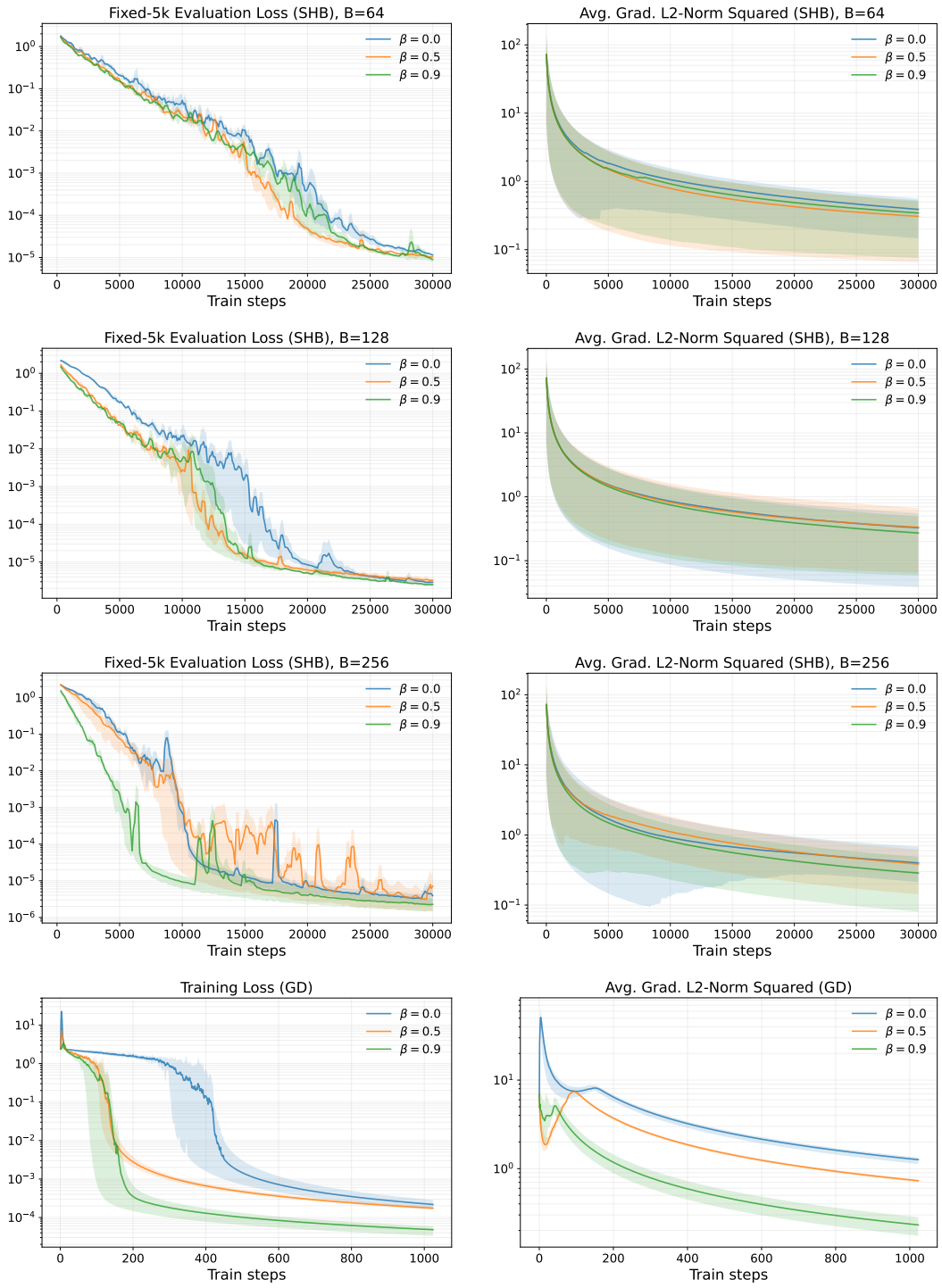

Figure 7: CIFAR-10/ResNet-18 results for SGD/SHB. The first three rows use minibatch sizes 64, 128, and 256, respectively, and the bottom row uses the fixed 5,000-image full-batch objective. Loss is shown on the left and the cumulative average squared ℓ_2 gradient norm on the right. Each β uses the learning rate selected for the corresponding metric. Solid curves and shading show the seed mean and minimum–maximum range. Minibatch losses are smoothed with a moving average of window size 300.

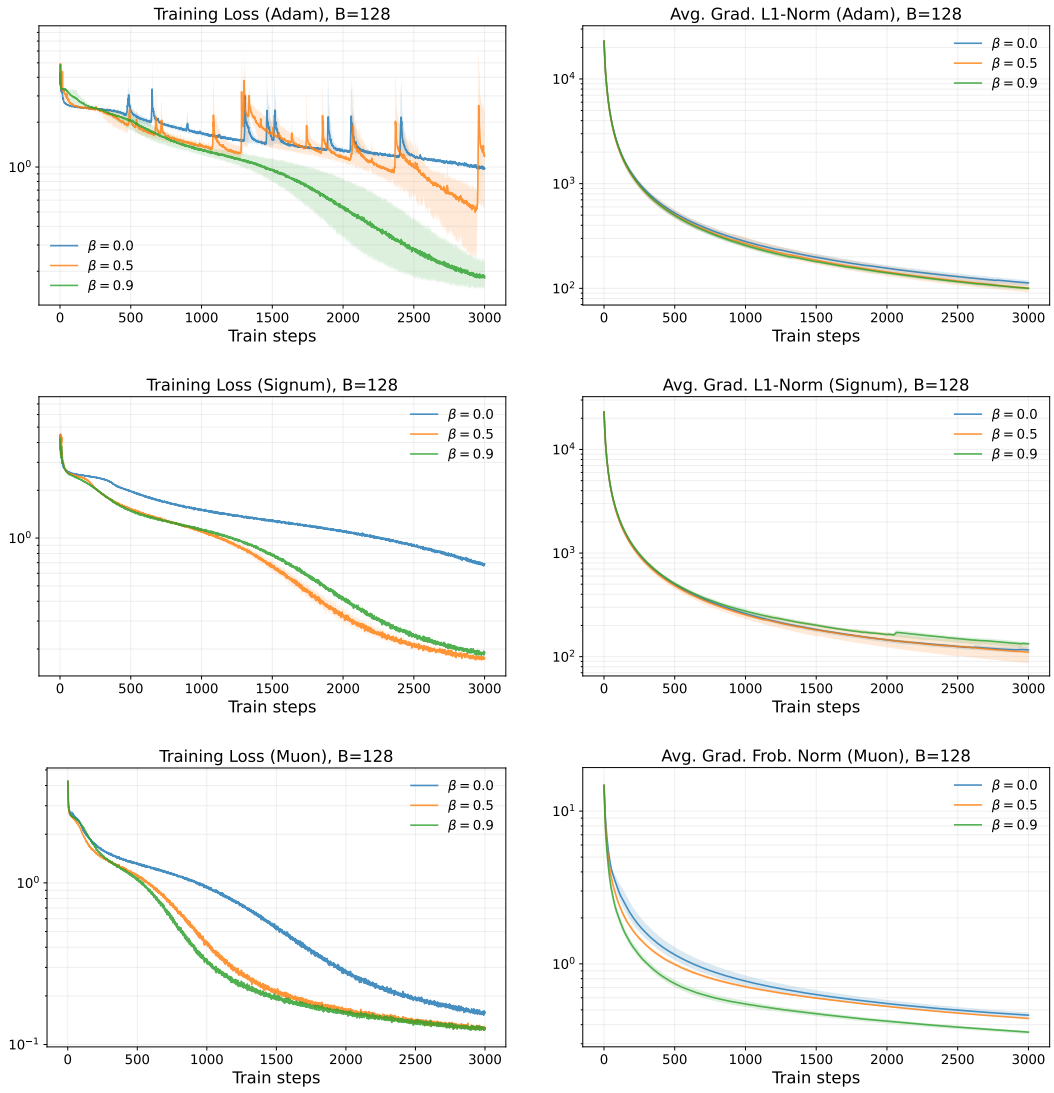

Figure 8: Tiny Shakespeare/GPT results with minibatch size 128. Rows correspond to Adam, Signum, and Muon; loss is shown on the left and the cumulative average gradient norm on the right. For Adam and Signum this is the ℓ_1 gradient norm, and for Muon it is the Frobenius gradient norm over Muon-updated parameters. Each β uses the learning rate selected for the corresponding metric. Solid curves and shading show the mean and minimum–maximum range across three seeds.

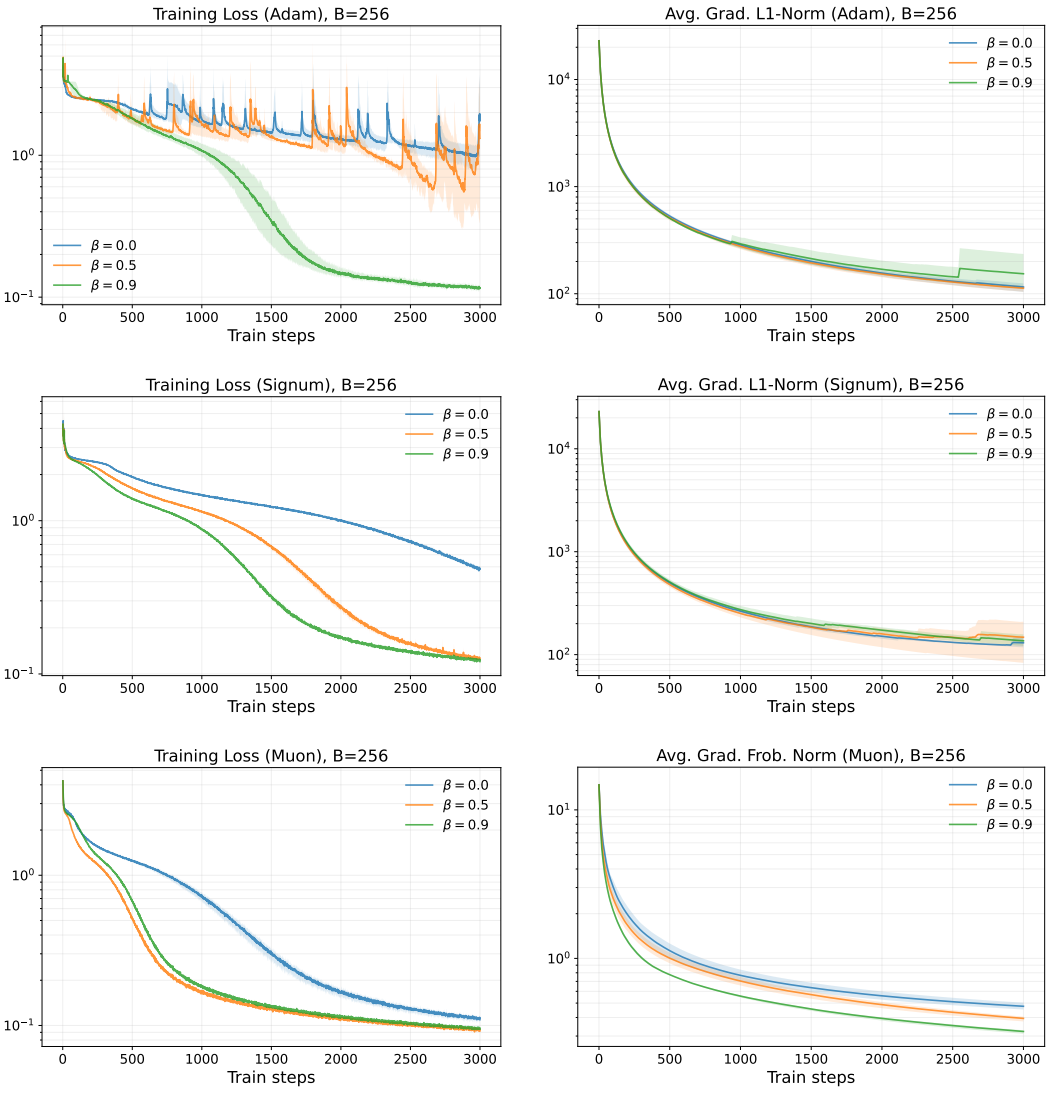

Figure 9: Tiny Shakespeare/GPT results with minibatch size 256. Rows correspond to Adam, Signum, and Muon; loss is shown on the left and the cumulative average gradient norm on the right. Each β uses the learning rate selected for the corresponding metric. Solid curves and shading show the mean and minimum–maximum range across three seeds.

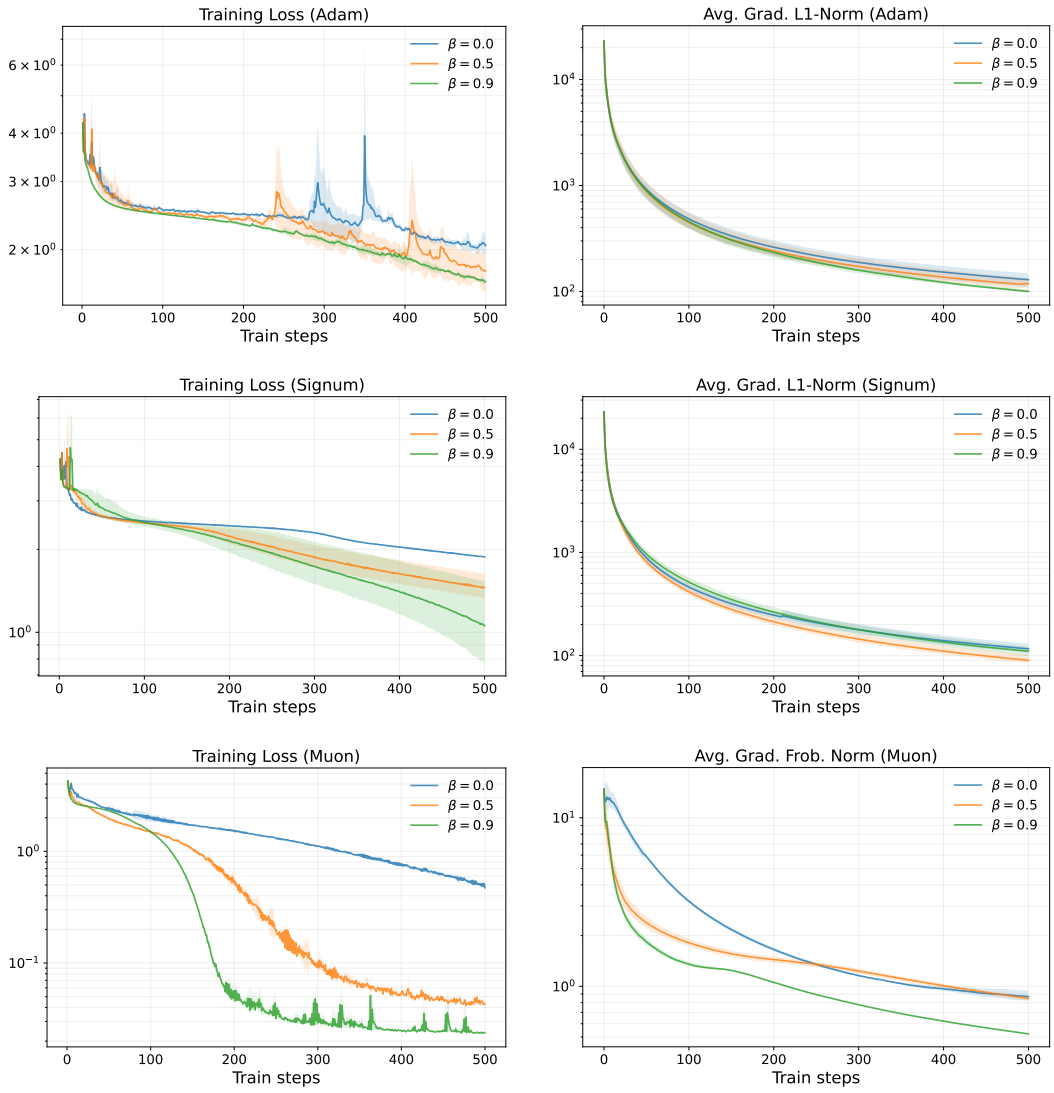

Figure 10: Tiny Shakespeare/GPT results for 500 updates on the *full-batch* objective of 3,921 fixed non-overlapping contexts. Rows correspond to Adam, Signum, and Muon; loss is shown on the left and the cumulative average gradient norm on the right. Each β uses the learning rate selected for the corresponding metric. Solid curves and shading show the mean and minimum–maximum range across three seeds.