

The Optimal Rate Constant of Heavy-Ball Momentum on Smooth Strongly Convex Functions

Marcos Hernanz

marcos@hernanz.io

Abstract

The heavy-ball method with a constant step size and a constant momentum parameter accelerates gradient descent on quadratics, but no choice of the two parameters does so on the class of smooth strongly convex functions. We determine how much faster than gradient descent it can be made on this class. Writing the best worst-case rate over all tunings as $1 - c(\kappa)/\kappa$, where κ is the condition number, we prove that $c(\kappa)$ tends to 8 in every dimension of two or more. The lower bound on the rate holds for every tuning, because a tuning that is faster than $1 - 8/\kappa$ on quadratics follows a slowly decaying spiral on some function of two variables. The matching upper bound comes from an explicit tuning, with nearly the momentum of Polyak's tuning and a step size about $\sqrt{\kappa}/8$ times smaller, and from a Lyapunov function whose positivity is verified in exact arithmetic. Optimally tuned heavy-ball is therefore asymptotically four times faster than gradient descent on this class.

Contents

1	Introduction	4
2	Related Work	6
3	Setting and Main Results	8
3.1	Worst-case rates	8
3.2	Notation and interpolation conditions	8
3.3	Main results	9
4	Where the Constant 8 Comes From	11
4.1	What quadratics allow	11
4.2	Spirals	11
4.3	The upper bound	12
5	Proof of the Lower Bound	14
5.1	Tunings that beat $1 - 8/\kappa$ on quadratics	14
5.2	Decaying spirals	14
5.3	The window	16
5.4	Roots of unity in the window	17
5.5	Proof of Theorem A	17
5.6	Rest start	18
6	Proof of the Upper Bound	19
6.1	The tuning on quadratics	19
6.2	The Lyapunov function	19
6.3	Structure of the certificate	20
6.4	Exact verification	20
6.5	Exact range	21
6.6	How the certificate was found	21
7	Computer-Assisted Steps	22
7.1	What was checked	22
7.2	Size of the computations and trust base	23
7.3	Floating-point tests	23
8	Finite Condition Numbers	24
8.1	Exact certificates at fixed condition numbers	24
8.2	A numerical bracket	24
8.3	Earlier numerical values	24
8.4	Shape of the tuning and the second-order term	25
9	Discussion and Limitations	26

9.1	What the results settle	26
9.2	Open questions	26
9.3	The constant 8 in the region of Ghadimi et al.	26
9.4	Limitations	27
10	Conclusion	28
	References	29
A	The Certificate of Theorem B	32
B	Polynomials of the Window Lemma	34
C	Exact Certificates at Fixed Condition Numbers	35
D	Results in One Dimension	36
E	Reproducibility	41

1 Introduction

The heavy-ball method of Polyak [24] minimizes a differentiable function f by the iteration

$$x_{t+1} = x_t - \gamma \nabla f(x_t) + \beta(x_t - x_{t-1}), \quad (1)$$

with a step size γ and a momentum parameter β . It is the prototype of the momentum methods used to train neural networks. On quadratics whose Hessian has its eigenvalues in $[\mu, L]$, Polyak's choice of γ and β converges at the rate $1 - 2/\sqrt{\kappa} + O(1/\kappa)$, where $\kappa = L/\mu$ is the condition number. This is the accelerated rate. When the dimension is large compared with the number of iterations, no first-order method has a better one, already on quadratics [19, 21, Thm. 2.1.13]. Gradient descent, the case $\beta = 0$, has the rate $1 - 2/\kappa + O(1/\kappa^2)$ with its best step size.

Beyond quadratics, heavy-ball loses this advantage. On the class $\mathcal{F}_{\mu,L}$ of L -smooth μ -strongly convex functions, Polyak's tuning fails to converge on some functions as soon as $\kappa \geq 9 + 4\sqrt{5}$ [22, Thm. 2.2], and Goujaud et al. [11] proved that no tuning accelerates in dimension two or more. Every choice of (γ, β) either fails to converge on some function of the class or converges no faster than $1 - O(1/\kappa)$, already on quadratics. Their counterexamples are cycles, found numerically by Goujaud et al. [10] and then constructed in closed form. Non-acceleration holds in one dimension as well [7, 16].

This leaves a quantitative question. Let $\rho^*(\kappa)$ be the best worst-case rate of (1) on $\mathcal{F}_{\mu,L}$ over all constant tunings (γ, β) , and write it as $\rho^*(\kappa) = 1 - c(\kappa)/\kappa$. Gradient descent shows that $c(\kappa) \geq 2\kappa/(\kappa + 1)$, and non-acceleration says that $c(\kappa)$ is bounded. How large is it? Goujaud et al. [11, §7] note that the exact rate is open. The thesis of Goujaud [9, §10.3.3] records the proved bounds, which are asymptotically $2 \leq c \leq 100/3$ in our normalization, computes numerical optima at $\kappa = 20, 100$ and 1000 , and asks what the limit is. It poses the related problem of finding analytical expressions for the optimal tuning and rate. Concurrently with this work, Liu et al. [16, App. F] proved a lower bound in one dimension that implies $c \leq 24$ (§2).

This paper determines the limit. In every dimension $d \geq 2$ we prove that $c(\kappa) \rightarrow 8$. The lower bound on the rate (Theorem A) holds for every tuning. If a tuning converges faster than $1 - 8/\kappa$ on quadratics, we construct a function of two variables on which heavy-ball follows the spiral $x_t = (1 - 8/\kappa)^t e^{2\pi i t/K}$ around the minimizer, for some integer K . The upper bound (Theorem B) is an explicit tuning for $\kappa \geq 400$, with momentum $1 - \beta \approx 4/\sqrt{\kappa}$ and step size $\gamma \approx 32/(L\sqrt{\kappa})$, whose worst-case rate is $1 - (8 - 64/\sqrt{\kappa})/\kappa$ in every dimension. Together they give $8 - 64/\sqrt{\kappa} \leq c(\kappa) \leq 8$ (Corollary 1). Optimally tuned heavy-ball is therefore asymptotically four times faster than gradient descent on $\mathcal{F}_{\mu,L}$, and this factor cannot be improved. This answers the question of Goujaud [9] about the limit, and it solves the problem posed there of the optimal tuning and rate to first order in $1/\kappa$.

The constant 8 has a short explanation, which §4 gives in full. Along a spiral the gradient is the iterate scaled and rotated by a fixed nonzero angle, which no quadratic can produce. Convexity allows such a rotation only in exchange for curvature, and the exchange is most favorable at half the largest curvature L . Balancing the two gives a condition $c(2w - 1) \geq 8w^2$ on a free parameter w , which some w satisfies if and only if $c \geq 8$. The tuning of Theorem B has the momentum of Polyak's tuning to first order and a step size about $\sqrt{\kappa}/8$ times smaller. Its proof uses a Lyapunov function that perturbs the mechanical energy of the heavy ball. Three of its six coefficients are forced by the quadratic on which the rate is attained. The other three were fitted numerically and then rounded, and the positivity of the resulting quadratic form is proved for all $\kappa \geq 400$ at once by exact root counting.

Several ingredients come from earlier work. Non-acceleration is due to Goujaud et al. [11] in dimension two and higher, and to Fotopoulos et al. [7] and Liu et al. [16] in one dimension. Our spirals are the decaying version of the root-of-unity cycles of Goujaud et al. [11] and the single-frequency case of the decaying waves of Pasechniuk-Vilensky [22], whose extra coordinate we use for the rest start. The tunings that beat a given rate on quadratics are described in Goujaud et al. [11, Lemma 2.4]. Our certificate belongs to the class with one step of memory and the lag-one multiplier, which is the weighted off-by-one constraint of Lessard et al. [15] (§2). At each fixed $\kappa > 64.82$ the rate of our tuning can be derived from a concurrent criterion of Pasechniuk-Vilensky [22, Thm. 7.2(b)] (Remark 2). What is new is the value 8 of the limit. The lower bound $1 - 8/\kappa$ for every tuning improves on $1 - (100/3 + o(1))/\kappa$ [11] and on the concurrent $1 - (24 + o(1))/\kappa$ [16]. The family that attains it to first order is explicit, and its certificate is valid for all $\kappa \geq 400$ at once.

Parts of both proofs are computer-assisted, and only exact arithmetic is used. Polynomial identities are verified symbolically and the positivity of explicit univariate polynomials by exact Sturm sequences. Section 7 lists every such step with its range, its method and its script. Section 8 compares the asymptotic result with what is known at finite κ , including exact certificates at four values of κ . Appendix D collects results in one dimension. There we prove that the constant $c_1(\kappa)$ of $\mathcal{F}_{\mu,L}(\mathbb{R})$ is at most $50/3 + o(1)$, show that the cycle method used in earlier lower bounds cannot go below 16, and give an exact non-convergent orbit of Polyak's tuning started at rest, for every $\kappa \geq 100$.

2 Related Work

Global analyses of heavy-ball on $\mathcal{F}_{\mu,L}$ give no explicit rate better than that of gradient descent. Ghadimi et al. [8] prove global linear convergence on an explicit region of tunings. Their rate is not explicit in general, and on a sub-family it is the gradient-descent factor $1 - \gamma\mu$ for function values. Goujaud et al. [11, Lemma 2.5] show that the best rate over that region, even on quadratics, is $1 - 8/\kappa + o(1/\kappa)$. This constant 8 is related to ours but describes different tunings (§9.3). Other global analyses of heavy-ball on $\mathcal{F}_{\mu,L}$, or of momentum families that contain it, certify explicit rates for small steps $\gamma = O(\mu/L^2)$ [3, 27]. Danilova et al. [4] study its non-monotone transient behavior on quadratics and prove global convergence under the Polyak–Łojasiewicz condition. Polyak’s own analysis is local [24]. Ugrinovskii et al. [31] prove global asymptotic convergence, without a global rate, on a class of functions with sector-bounded gradients and optimize the local rate. Sanz-Serna and Zygalakis [25, §6] show that heavy-ball has no Lyapunov function of the kind they construct for Nesterov’s method, which would certify an accelerated rate.

Polyak’s tuning can fail on $\mathcal{F}_{\mu,L}$. Lessard et al. [15] found a one-dimensional function at $\kappa = 25$ on which its iterates approach a cycle. Badithela and Seiler [2] solved the linear matrix inequality of the weighted off-by-one constraint analytically for Polyak’s tuning and constructed a function on which the method enters a limit cycle. As Pasechniuk-Vilensky [22, §1.1] describe it, the bound reaches 1 at $\kappa = 9 + 4\sqrt{5}$, where a function of one variable has a three-cycle. Pasechniuk-Vilensky [22, Thm. 2.2] prove that $9 + 4\sqrt{5}$ is the exact threshold, with convergence on all of $\mathcal{F}_{\mu,L}$ below it and a non-convergent trajectory at and above it. Wang et al. [36] claim acceleration for a class of Polyak–Łojasiewicz functions. Their erratum (arXiv v3, App. J) adds a co-diagonalization condition, and Goujaud et al. [12] point to an incorrect inequality in the remaining one-dimensional argument. With time-varying parameters on smooth convex functions, Ma and Zhang [17] prove a lower bound for heavy-ball that holds for every schedule fixed in advance, and He and Zhang [13] use randomized schedules to obtain a faster rate than the best known one for gradient descent. We only consider constant parameters.

Other lower bounds for every tuning come from cycles. Goujaud et al. [11] constructed heavy-ball cycles on the K -th roots of unity in the plane (their Thm. 3.5) and proved that every tuning outside the cycling regions satisfies $\rho \geq (1 - C/\kappa)/(1 + C/\kappa)$ for every $C > 50/3$ (Thm. 3.6), which is $c \leq 100/3$ in our normalization. Goujaud et al. [12] asked whether heavy-ball accelerates on functions of one variable. Fotopoulos et al. [7] answered this negatively by rotating the planar cycles and projecting them onto a line. Building on their construction, Liu et al. [16, App. F] prove that in one dimension every tuning with $\gamma > 0$ and $0 \leq \beta < 1$ has rate at least $(\kappa - 12)/(\kappa + 12)$ when $\kappa \geq 24$. Tunings outside this range have $\rho_Q \geq 1 - 8/\kappa$ by Lemma 1, which is larger. Since ρ^* does not decrease with the dimension, this gives $c \leq 24$ in every dimension.

The concurrent work of Pasechniuk-Vilensky [22] is the closest to ours. It answers the second riddle of Goujaud et al. [12] and develops a rate theory for general tunings, for the rest-start rate taken uniformly over the dimension. Interpolable mixtures of rotating geometric sequences, decaying ones included, give lower bounds (their Lemma 3.4 and Thm. 3.5(c)), and a lag-one Zames–Falb multiplier gives upper bounds (Thm. 7.2(a)). On an explicit region Q of tunings the rate equals the rate on quadratics (Thm. 7.2(b)). They do not optimize over tunings and do not state a limiting constant for the optimal rate. Among the questions they leave open are the rates on the part of their region Ω_1 outside Q and the asymptotics of their gaps as $\kappa \rightarrow \infty$ (their §10). The spiral of Theorem A is the single-frequency case of their waves, and the tuning of Theorem B lies in their region Q (Remark 2). Our contribution relative to theirs is the quantitative bound for every tuning and a family, uniform in κ , that matches it.

Worst-case rates for fixed parameters can be computed or certified by performance estimation [6, 30], by Lyapunov searches [29, 32] and by integral quadratic constraints [15, 18]. Lessard and Seiler [14] synthesize algorithms with bounds on the achievable rate. For heavy-ball with step $1/L$ and numerically optimized momentum, Lessard et al. [15, Fig. 8] report rates between about $1 - 6.4/\kappa$ and $1 - 7.1/\kappa$ for κ between about 850 and 4200 (our reading of their figure, whose resolution we discuss in §8.3). They describe this as comparable to gradient descent, which is true of the order in κ . Van Scoy and Lessard [33, Fig. 4] find $c \approx 5$ at $\kappa = 100$, and Upadhyaya et al. [32, Fig. 2c] find $c \approx 2.2$ at $\kappa = 10$ with step $1/L$, both read off their figures and converted from rates of the squared distance. Goujaud’s thesis [9, Table 10.1] reports optima at $\kappa = 20, 100$ and 1000 (§8.3). The lag-one multiplier behind our certificate is a discrete-time analogue of the Zames–Falb multipliers [37], and Pasechniuk-Vilensky [22, §1.5] note that it coincides with the weighted off-by-one constraint of Lessard et al. [15]. They relate single-harmonic periodic orbits to the periodic counterexamples in absolute stability of Seiler and Carrasco [26], and multi-harmonic obstructions to the duality bounds for Zames–Falb multipliers of Zhang et al. [38].

Spirals have been used for Nesterov’s method as well. Pokutta and Stellato [23] determine the exact worst-case asymptotic rate of Nesterov’s method with its standard parameters, using planar spirals with a common contraction factor and a linear program over measures. That work is concurrent and concerns a different method, but its tools are close to ours.

Methods that evaluate the gradient elsewhere do accelerate on $\mathcal{F}_{\mu,L}$. Measured, as here, by the distance to the minimizer, Nesterov’s method [20] with constant momentum converges at rate at most $(1 - 1/\sqrt{\kappa})^{1/2} \approx 1 - 1/(2\sqrt{\kappa})$ [21, §2.2.1], and the triple momentum method at rate $1 - 1/\sqrt{\kappa}$ [35]. No first-order method has a better rate for the distance to the minimizer when the dimension is large compared with the number of iterations. On $\mathcal{F}_{\mu,L}(\mathbb{R}^{2N+1})$ the lower bound of Drori and Taylor [5] after N steps is of order $(1 - 1/\sqrt{\kappa})^N$, and the method of Taylor and Drori [28] attains it. Nesterov’s method and the triple momentum method evaluate the gradient at a point other than the iterate, which heavy-ball does not. If f is in addition twice continuously differentiable near the minimizer, the asymptotic rate of a globally convergent method is its local rate, which for heavy-ball can be the accelerated rate of quadratics [24]. Ugrinovskii et al. [31] exploit this, and so do Van Scoy and Lessard [34] with a new method. The functions in our lower bound are not twice differentiable at the minimizer, and cannot be. If f were twice differentiable at $x_\star$, the iterates of Theorem A would eventually converge at a rate at most $\rho_Q < 1 - 8/\kappa$.

3 Setting and Main Results

3.1 Worst-case rates

For $0 < \mu < L$, let $\mathcal{F}_{\mu,L}(\mathbb{R}^d)$ be the set of differentiable functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $f - (\mu/2)\|\cdot\|^2$ is convex and ∇f is L -Lipschitz, and let $\mathcal{Q}_{\mu,L}$ be its subset of quadratics. Each $f \in \mathcal{F}_{\mu,L}$ has a unique minimizer $x_\star$. Heavy-ball (1) starts from an arbitrary pair (x_{-1}, x_0) , as in the cycle constructions of Goujaud et al. [11]. For a tuning $(\gamma, \beta) \in \mathbb{R}^2$ and a set of functions $\mathcal{F}$, the rate $\rho_{\gamma,\beta}(\mathcal{F})$ is the smallest $\rho \geq 0$ with the following property. For every $\varepsilon > 0$ there is a T_0 such that

$$\|x_T - x_\star\| \leq (\rho + \varepsilon)^T \max\{\|x_0 - x_\star\|, \|x_{-1} - x_\star\|\} \quad (2)$$

for all $f \in \mathcal{F}$, all initial pairs and all $T \geq T_0$. The rest-start rate $\rho_{\gamma,\beta}^{\text{rest}}$ is defined in the same way with $x_{-1} = x_0$. The optimal rate in dimension d and its constant are

$$\rho_d^\star(\kappa) = \inf_{(\gamma,\beta) \in \mathbb{R}^2} \rho_{\gamma,\beta}(\mathcal{F}_{\mu,L}(\mathbb{R}^d)), \quad c_d(\kappa) = \kappa(1 - \rho_d^\star(\kappa)).$$

By scaling, $\rho_d^\star$ depends on μ and L only through $\kappa = L/\mu$. It does not decrease with d , because adding coordinates on which f is $(\mu/2)\|\cdot\|^2$ embeds $\mathcal{F}_{\mu,L}(\mathbb{R}^d)$ into $\mathcal{F}_{\mu,L}(\mathbb{R}^{d+1})$. Gradient descent with $\gamma = 2/(L + \mu)$ has rate $(\kappa - 1)/(\kappa + 1)$ on $\mathcal{F}_{\mu,L}$ [21, Thm. 2.1.15], so $c_d(\kappa) \geq 2\kappa/(\kappa + 1)$.

Remark 1 (conventions). Goujaud et al. [11] and Goujaud [9] write κ for μ/L and state rates as $(1 - C\mu/L)/(1 + C\mu/L)$. In our notation this is $c = 2C\kappa/(\kappa + C) \rightarrow 2C$. Asymptotically, our constant 8 is $C = 4$ there, and their bound for every $C > 50/3$ is $c \leq 100/3$. They normalize by $\|x_0 - x_\star\|$ and do not specify the second initial point [11, Defs. 1.3–1.5]. Pasechniuk-Vilensky [22] use the rest start (their eq. (1.1)) and take the supremum over all dimensions inside the rate (their eq. (1.2)). Our results hold under each of these conventions. The lower bound holds for arbitrary initial pairs in $d \geq 2$ and from rest in $d \geq 3$, normalized by either initial distance. The upper bound holds for every initial pair in every dimension, with a constant independent of d . Normalized by $\|x_0 - x_\star\|$ alone, it needs x_{-1} to be determined by x_0 , as in the rest start.

3.2 Notation and interpolation conditions

From now on we take $L = 1$ and write

$$q = \mu = 1/\kappa, \quad s = \sqrt{q} = \kappa^{-1/2}, \quad h = \gamma L,$$

with $x_\star = 0$ and $f(x_\star) = 0$. On the quadratic $(\lambda/2)\|x\|^2$, heavy-ball is a linear recursion with characteristic polynomial $z^2 - (1 + \beta - h\lambda)z + \beta$. Let $r(\lambda)$ be the largest modulus of its roots and $\rho_Q = \max_{\lambda \in [q,1]} r(\lambda) = \max\{r(q), r(1)\}$. Running heavy-ball from an eigenvector, or from the real part of a complex one, shows that $\rho_{\gamma,\beta}(\mathcal{F}_{\mu,L}) \geq \rho_Q$, and ρ_Q is the rate on $\mathcal{Q}_{\mu,L}$ [11, Prop. 2.1].

Both proofs rest on the interpolation conditions for $\mathcal{F}_{\mu,L}$. A finite set of triples (x_i, g_i, f_i) is $\mathcal{F}_{\mu,L}$ -interpolable if some $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$ has $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$ for all i . By Taylor et al. [30, Thm. 4], this holds if and only if, for all i and j ,

$$I(i, j) := f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{\|g_i - g_j\|^2 + q\|x_i - x_j\|^2 - 2q\langle g_i - g_j, x_i - x_j \rangle}{2(1 - q)} \geq 0. \quad (3)$$

The lower bound constructs data that satisfy these inequalities, and the upper bound adds two of them with non-negative multipliers.

3.3 Main results

Theorem A (lower bound). *Let $0 < \mu < L$ and $d \geq 2$. For every $(\gamma, \beta) \in \mathbb{R}^2$,*

$$\rho_{\gamma, \beta}(\mathcal{F}_{\mu, L}(\mathbb{R}^d)) \geq 1 - \frac{8}{\kappa}.$$

More precisely, let $\kappa > 8$ and $r = 1 - 8/\kappa$. Either $\rho_Q \geq r$, or there are an integer $K \geq 4$, a function $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^2)$ and an initial pair on which heavy-ball satisfies $x_t - x_\star = r^t R_{2\pi t/K}(x_0 - x_\star)$ for all $t \geq 0$, where R_θ is the rotation of the plane by the angle θ .

Theorem B (achievability). *Let $\kappa \geq 400$, $s = \kappa^{-1/2}$, $C = 8 - 64s$, and*

$$r = 1 - Cs^2, \quad \beta = r(r - 4s), \quad \gamma = \frac{Cs(4 + Cs)}{L}. \quad (4)$$

There is a constant M , depending only on κ , such that for every d , every $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^d)$ and every initial pair (x_{-1}, x_0) ,

$$\|x_t - x_\star\| \leq M r^t (\|x_0 - x_\star\| + \|x_{-1} - x_\star\|) \quad \text{for all } t \geq 0.$$

The rate is attained on $f = (\mu/2)\|x\|^2$, already from rest. Hence

$$\rho_{\gamma, \beta}(\mathcal{F}_{\mu, L}(\mathbb{R}^d)) = \rho_{\gamma, \beta}^{\text{rest}}(\mathcal{F}_{\mu, L}(\mathbb{R}^d)) = 1 - \frac{8 - 64/\sqrt{\kappa}}{\kappa}.$$

The constant M is given in the proof. For example, $M < 78$ at $\kappa = 400$ and $M < 559$ at $\kappa = 10^4$ (Table 3). The same certificate works for every $\kappa \geq 357.2901 \dots$ and for no smaller κ (§6.5).

Corollary 1 (the optimal constant). *For $d \geq 2$ and $\kappa \geq 400$,*

$$1 - \frac{8}{\kappa} \leq \rho_d^\star(\kappa) \leq 1 - \frac{8}{\kappa} + \frac{64}{\kappa^{3/2}}, \quad \text{that is,} \quad 8 - \frac{64}{\sqrt{\kappa}} \leq c_d(\kappa) \leq 8.$$

The upper bound $c_d(\kappa) \leq 8$ holds for all $\kappa > 1$. In particular $c_d(\kappa) \rightarrow 8$, while the constant $2\kappa/(\kappa + 1)$ of gradient descent tends to 2. In one dimension, $c_1(\kappa) \geq 8 - 64/\sqrt{\kappa}$ for $\kappa \geq 400$.

Corollary 2 (rest start). *For $d \geq 3$ and every $(\gamma, \beta) \in \mathbb{R}^2$, $\rho_{\gamma, \beta}^{\text{rest}}(\mathcal{F}_{\mu, L}(\mathbb{R}^d)) \geq 1 - 8/\kappa$. Hence Corollary 1 holds for the rest-start rate as well, for $d \geq 3$.*

In terms of iterations, the spiral of Theorem A has $\|x_N - x_\star\| = r^{N+1} \max\{\|x_0 - x_\star\|, \|x_{-1} - x_\star\|\}$. If $\rho_Q \geq r$, the trajectory $x_t = z^{t+1}$ on $(\lambda/2)\|x\|^2$, with $\mathbb{R}^2$ identified with $\mathbb{C}$ and z a root of modulus ρ_Q for $\lambda = q$ or $\lambda = 1$, decays no faster. To reduce the distance to the minimizer by a factor $\varepsilon < 1$ on every function of $\mathcal{F}_{\mu, L}(\mathbb{R}^2)$, every tuning therefore needs at least $\log(1/\varepsilon)/\log(1/r) - 1 \geq (\kappa/8 - 1)\log(1/\varepsilon) - 1$ iterations. The tuning of Theorem B needs at most $\lceil \kappa(\log(1/\varepsilon) + \log 2M)/(8 - 64/\sqrt{\kappa}) \rceil$. Gradient descent needs about $(\kappa/2)\log(1/\varepsilon)$ iterations, and the triple momentum method about $\sqrt{\kappa}\log(1/\varepsilon)$.

Remark 2 (relation to Pasechniuk-Vilensky). The tuning (4) lies in the region Q of Pasechniuk-Vilensky [22, Thm. 7.2(b)] for every $\kappa > 64.82$. For this family, their membership conditions

(Cor. 7.3(a),(c)) reduce to the positivity of univariate polynomials in s , which we checked by exact root counting (§7). On Q they prove that the rate equals the rate on quadratics, which here is r . At each fixed κ , their theorem therefore gives the rate equality of Theorem B independently. Their statement is for the rest start and loses an arbitrary ε in the rate, although their proof controls arbitrary initial states. Theorem B adds a short, self-contained certificate in the time domain with an explicit constant M for arbitrary initial pairs, and it identifies a family that is asymptotically optimal by Theorem A.

4 Where the Constant 8 Comes From

This section explains the two theorems without full proofs. We keep $L = 1$.

4.1 What quadratics allow

On a quadratic of curvature λ , heavy-ball is a linear recursion. In the flattest direction, $\lambda = q$, it converges at a real rate r if r is a root of $z^2 - (1 + \beta - hq)z + \beta$. Evaluating this polynomial at $z = 1$ gives $hq = (1 - r)(1 - \beta/r)$. A real rate $r = 1 - c/\kappa$ therefore requires

$$\frac{h}{1 - \beta/r} = c.$$

We call the left side the *effective step*. For gradient descent it equals h , and stability in the stiffest direction, $\lambda = 1$, requires $h < 2$, hence $c < 2$. Momentum separates the two requirements. The stiff direction only needs $h \lesssim 2(1 + \beta)$, while the effective step grows without bound as $\beta/r \rightarrow 1$. The momentum cannot exceed r^2 , because the roots for $\lambda = q$ have product β , so one of them has modulus at least $\sqrt{\beta}$. In the coordinates

$$X = 1 - \beta/r \quad \text{and} \quad h = \gamma L,$$

the tunings that beat $r = 1 - 8/\kappa$ on quadratics form a triangle (Lemma 1; 11, Lemma 2.4). Its three sides are $X > \eta := 1 - r$, which bounds the momentum, $h > 8X$, which says that the effective step exceeds 8, and $h < (2 - \eta)(2 - X)$, which keeps the stiff direction stable. On quadratics alone, c can be as large as about $2\sqrt{\kappa}$, which Polyak's tuning attains. Everything that limits c to 8 comes from functions that are not quadratic.

4.2 Spirals

Identify $\mathbb{R}^2$ with $\mathbb{C}$ and look for a heavy-ball trajectory that spirals into the minimizer, $x_t = r^t e^{i\theta t}$. The iteration (1) determines the gradients along it, $g_t = (a + ib)x_t$ with real a and b that depend on the tuning, by (6). The part $a x_t$ is what the quadratic $(a/2)\|x\|^2$ would produce. The part $ib x_t$ is perpendicular to x_t (Figure 1). No quadratic produces it along the whole spiral, because a Hessian H with $Hx_t = (a + ib)x_t$ for all t is the multiplication by $a + ib$, which is symmetric only if $b = 0$. A convex function that is not quadratic can produce some of this rotation, but only a limited amount. The interpolation conditions allow it as long as

$$|b| \leq 2m \tan \frac{\theta}{2}, \quad \text{where} \quad m = \frac{(a - q)(1 - a) - b^2}{2(1 - q)} \approx \frac{a(1 - a)}{2}$$

(Lemma 2). The rotation b has to be paid for by a curvature a strictly between q and 1 and by a large enough turning angle θ .

Now place the tuning on the edge $h = cX$, where its rate on the quadratic of curvature q is about $1 - c/\kappa$, and write $1 - \cos \theta = 2Xw$ for some $w > 0$. Suppose that X is small and that η is much smaller than X , which is the case for the tunings that matter as $\kappa \rightarrow \infty$. Then (6) and (7), with $a = \operatorname{Re} A/h$ and $b = \operatorname{Im} A/h$, give $a \approx 4w/c$, $|b| \approx 2\sqrt{Xw}/c$ and $\tan(\theta/2) \approx \sqrt{Xw}$. After division by $\sqrt{Xw}/c$, the condition above becomes $2 \leq 4w(1 - 4w/c)$, that is,

$$c(2w - 1) \geq 8w^2.$$

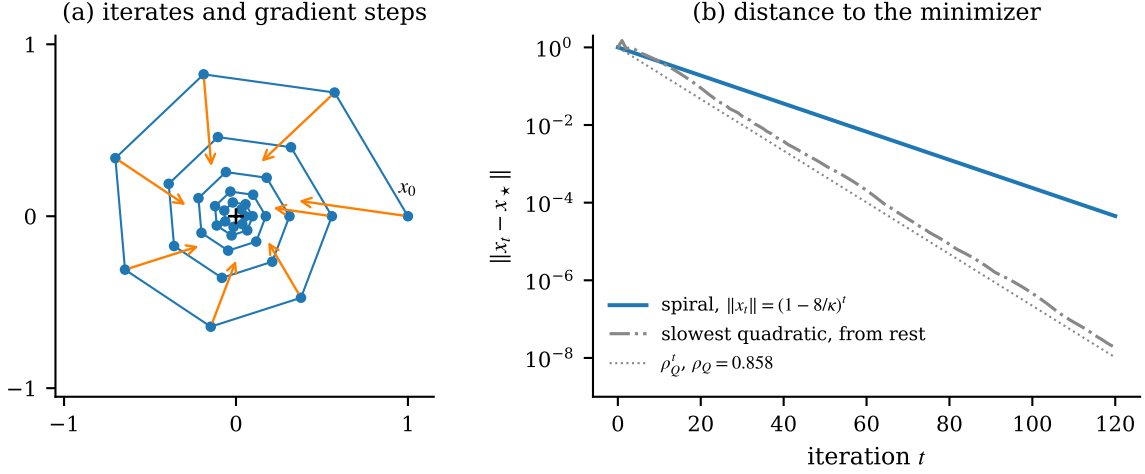


Figure 1: A spiral of Lemma 2 at $\kappa = 100$, for the tuning $\gamma L = 2.5$, $\beta = 0.736$. (a) The iterates turn by $2\pi/7$ per step and contract by $r = 1 - 8/\kappa = 0.92$. The gradients are the iterates scaled and rotated by a fixed nonzero angle, which no quadratic produces. The orange arrows are the gradient steps $-\gamma \nabla f(x_t)$. (b) On this function heavy-ball converges at the rate 0.92, although on every quadratic of the class it converges at the rate $\rho_Q = 0.858$ or faster.

Some w satisfies this if and only if

$$c \geq \min_{w > 1/2} \frac{8w^2}{2w - 1} = 8,$$

and the minimum is attained at $w = 1$. At $c = 8$ the best spiral has $a = 1/2$, the middle of the range of curvatures, $m = 1/8$ and $|b| = \sqrt{X}/4$, and it turns by $\theta \approx 2\sqrt{X}$ per step. More precisely, with $\vartheta = \eta/X$, the polynomial Φ of Lemma 2 satisfies

$$\Phi(cX, 2Xw) = 2X^2[c(2w - 1 + \vartheta) - 8w^2] + O(X^3).$$

A tuning that beats $1 - 8/\kappa$ on quadratics has $h > 8X$, so it lies beyond the edge $c = 8$, and decaying spirals at the rate $1 - 8/\kappa$ exist for it. The proof of Theorem A makes this precise. It finds a window of values of (h, u) around $u = 2X$ on which $\Phi > 0$ (Lemma 4), and a root of unity inside it (Lemma 5).

4.3 The upper bound

For the converse we need a tuning that lies on the edge $h = CX$ for its own rate, with C just below 8, and a proof that no function in $\mathcal{F}_{\mu,L}$ slows it down. The tuning (4) satisfies $X = s(4 + Cs)$ and $h = CX$, so its effective step is C . Its momentum agrees with Polyak's to first order, $1 - \beta \approx 4/\sqrt{\kappa}$, while its step $\gamma L \approx 32/\sqrt{\kappa}$ is about $\sqrt{\kappa}/8$ times smaller than Polyak's.

The proof uses a Lyapunov function. Let $\varphi = f - (\mu/2)\|\cdot\|^2$, which is convex and $(1 - q)$ -smooth. To leading order in s , the Lyapunov function is

$$V_t \approx f(x_{t-1}) + \frac{\|x_t - x_{t-1}\|^2}{2\gamma} + \frac{1}{2}\langle x_t - x_{t-1}, \nabla \varphi(x_{t-1}) \rangle + 2s\|\nabla \varphi(x_{t-1})\|^2.$$

The first two terms are the potential and the kinetic energy of the heavy ball. The third couples the velocity with the gradient, and the fourth is a small multiple of the squared gradient. We show that $V_{t+1} \leq r^2 V_t$ on every $f \in \mathcal{F}_{\mu,L}$. The proof adds two inequalities satisfied by φ , one between consecutive iterates and one between the iterate and the minimizer, with multipliers r^2 and $1 - r^2$. These are the only multipliers that cancel the function values. What remains is a quadratic form in $x_t, x_{t-1}, \nabla\varphi(x_{t-1})$ and $\nabla\varphi(x_t)$ that must be positive semidefinite. Because the quadratic $(\mu/2)\|x\|^2$ attains the rate r , the form must vanish in the corresponding direction, and this fixes three of the six entries of the matrix that defines V . We found the other three numerically, as power series in s , rounded them, and proved positivity for all $\kappa \geq 400$ by exact root counting. Section 4.2 predicts this outcome. With effective step $C < 8$ no spiral can slow the tuning down, and the certificate shows that no other function can either.

5 Proof of the Lower Bound

This section proves Theorem A. Fix $q < 1/8$, so that $r = 1 - 8q \in (0, 1)$. For $q \geq 1/8$ there is nothing to prove. Write $\eta = 1 - r = 8q$. The proof has four steps. A tuning that beats r on quadratics lies in a triangle (Lemma 1). A decaying spiral at rate r is a heavy-ball trajectory on some function of $\mathcal{F}_{\mu,L}$ as soon as a polynomial Φ is non-negative (Lemma 2). This polynomial is positive on an explicit window (Lemma 4), and some root of unity falls in that window (Lemma 5).

5.1 Tunings that beat $1 - 8/\kappa$ on quadratics

For a tuning (h, β) define

$$W = \frac{r^2 - \beta}{r}, \quad X = W + \eta = 1 - \frac{\beta}{r}, \quad \text{so that} \quad \beta = r(1 - X). \quad (5)$$

Lemma 1 (quadratic necessary conditions). *If $\rho_Q < r$, then*

$$\eta < X < \frac{2(2 - \eta)}{10 - \eta} \leq \frac{2}{5} \quad \text{and} \quad 8X < h < (2 - \eta)(2 - X).$$

In particular $0 < \beta < 1$, $W > 0$ and $h > 0$.

Proof. By the Schur–Cohn criterion, both roots of $z^2 - az + b$, with a and b real, lie in the open disc of radius r if and only if $|b| < r^2$ and $|a|r < r^2 + b$. We apply it to $\lambda = q$, where $a = 1 + \beta - qh$, and to $\lambda = 1$, where $a = 1 + \beta - h$. First, $\beta < r^2$ gives $W > 0$, that is, $X > \eta$. With $\beta = r(1 - X)$ one checks that

$$(r^2 + \beta) - (1 + \beta - qh)r = r(qh - \eta X) \quad \text{and} \quad (r^2 + \beta) + (1 + \beta - h)r = r((2 - \eta)(2 - X) - h).$$

So $\lambda = q$ gives $qh > \eta X$, that is, $h > 8X$, and $\lambda = 1$ gives $h < (2 - \eta)(2 - X)$. The two are compatible only if $8X < (2 - \eta)(2 - X)$, that is, $X < 2(2 - \eta)/(10 - \eta)$, which is at most $2/5$. Then $\beta = r(1 - X) \in (0, 1)$. $\square$

These conditions describe the interior of the sublevel set $\text{SLS}(r)$ of Goujaud et al. [11, Lemma 2.4] in the coordinates (X, h) . The edge $h = 8X$ is their edge $\gamma = (1 - \rho)(1 - \beta/\rho)/\mu$, the edge $X = \eta$ is $\beta = \rho^2$, and the third edge is the stability edge of the stiff direction. The compatibility condition forces $\eta < 6 - 4\sqrt{2}$, so for $q \geq 0.043$ every tuning has $\rho_Q \geq r$.

5.2 Decaying spirals

Identify $\mathbb{R}^2$ with $\mathbb{C}$, with the inner product $\langle v, v' \rangle = \text{Re}(v\bar{v}')$. For an integer $K \geq 3$ let $\theta = 2\pi/K$, $z = re^{i\theta}$ and $u = 1 - \cos \theta = 2\sin^2(\pi/K)$. If heavy-ball is to follow $x_t = z^t$, then (1) forces the gradients

$$h g_t = (1 + \beta)x_t - \beta x_{t-1} - x_{t+1} = A x_t, \quad \text{where} \quad A = 1 + \beta - \frac{\beta}{r}e^{-i\theta} - re^{i\theta}. \quad (6)$$

With $\beta = r^2 - rW$ a short computation gives

$$\text{Re } A = \alpha := u(2 - \eta - X) + \eta X \quad \text{and} \quad \text{Im } A = -W \sin \theta. \quad (7)$$

Lemma 2 (decaying spirals). *Let $r \in (0, 1]$, $\eta = 1 - r$ and $K \geq 3$, with θ, z and u as above. Let (h, β) be a tuning with $h > 0$ and $W \geq 0$, where W and X are defined by (5) with this r , let α be as in (7), and let*

$$\Phi(h, u) = -qh^2 + h[(1 + q)\alpha - (1 - q)(2 - u)W] - \alpha^2 - W^2u(2 - u). \quad (8)$$

If $\Phi(h, u) \geq 0$, there is an $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^2)$ on which heavy-ball started at $(x_{-1}, x_0) = (z^{-1}, 1)$ gives $x_t = z^t$ for all $t \geq 0$. In particular $\rho_{\gamma, \beta}(\mathcal{F}_{\mu, L}(\mathbb{R}^2)) \geq r$.

Proof. Take the data $x_t = z^t$, $g_t = (a + ib)x_t$ and $f_t = (a/2)|x_t|^2$ for $t \geq 0$, with $a = \operatorname{Re} A/h$ and $b = \operatorname{Im} A/h$, together with the minimizer $(0, 0, 0)$. By (6), heavy-ball maps (x_{t-1}, x_t) to x_{t+1} on any function that interpolates these data. Substituting into (3) gives, for all $i \neq j$,

$$I(i, j) = m |x_i - x_j|^2 + b \operatorname{Im}(x_j \bar{x}_i), \quad I(i, \star) = I(\star, i) = m |x_i|^2, \quad (9)$$

where $m = ((a - q)(1 - a) - b^2)/(2(1 - q))$ is the quantity of §4.2. For $\{i, j\} = \{n, n + k\}$ with $k \geq 1$, this is $r^{2n}[m|1 - z^k|^2 \pm b r^k \sin k\theta]$, and both signs occur. So the data are interpolable if and only if $m \geq 0$ and

$$m |1 - z^k|^2 \geq |b| r^k |\sin k\theta| \quad \text{for all } k \geq 1. \quad (10)$$

Since

$$|1 - z^k|^2 = (1 - r^k)^2 + 2r^k(1 - \cos k\theta) \geq 4r^k \sin^2 \frac{\pi k}{K} \quad \text{and} \quad |\sin k\theta| = 2 \left| \sin \frac{\pi k}{K} \cos \frac{\pi k}{K} \right|,$$

the right side of (10) vanishes when $\sin(\pi k/K) = 0$. For the other k , inequality (10) follows from $2m |\sin(\pi k/K)| \geq |b| |\cos(\pi k/K)|$. Since $|\tan(\pi k/K)| \geq \tan(\pi/K)$ for these k , this holds as soon as $|b| \leq 2m \tan(\pi/K)$, a condition that implies $m \geq 0$. Now $h|b| = W \sin \theta$ by (7), and $\sin \theta / \tan(\pi/K) = 1 + \cos \theta = 2 - u$. Multiplying by $(1 - q)h^2 / \tan(\pi/K) > 0$ turns the condition into

$$2(1 - q)h^2 m - (1 - q)hW(2 - u) \geq 0.$$

Expanding $2(1 - q)h^2 m = (\alpha - qh)(h - \alpha) - (\operatorname{Im} A)^2$ with $(\operatorname{Im} A)^2 = W^2u(2 - u)$ shows that this is the condition $\Phi(h, u) \geq 0$, and this condition implies $m \geq 0$. So every finite subset of the data is interpolable, and Lemma 3 below provides a single $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^2)$. $\square$

At $r = 1$ the bound $|1 - z^k|^2 \geq 4r^k \sin^2(\pi k/K)$ is an equality, the condition $|b| \leq 2m \tan(\pi/K)$ is necessary as well as sufficient, and among the tunings that converge on quadratics, the set $\{\Phi \geq 0\}$ is the roots-of-unity cycling region of Goujaud et al. [11, Thm. 3.5] (checked symbolically). Lemma 2 is therefore the decaying version of their cycles, and the single-frequency case of the decaying waves of Pasechniuk-Vilensky [22, Lemma 3.4, Thm. 3.5(c)]. What it adds is a closed-form condition, $\Phi \geq 0$, that can be checked for all tunings at once.

Lemma 3 (infinite interpolation). *Let $(x_t, g_t, f_t)_{t \geq 0}$ in $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be such that every finite subfamily, together with $(x_\star, 0, f_\star)$, is $\mathcal{F}_{\mu, L}$ -interpolable. Then a single $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^d)$ has $f(x_t) = f_t$ and $\nabla f(x_t) = g_t$ for all t , and $f(x_\star) = f_\star$ and $\nabla f(x_\star) = 0$.*

Proof. Let $f_N \in \mathcal{F}_{\mu, L}(\mathbb{R}^d)$ interpolate the first N points and $(x_\star, 0, f_\star)$. Then $f_\star + (\mu/2)\|x - x_\star\|^2 \leq f_N(x) \leq f_\star + (L/2)\|x - x_\star\|^2$, and the gradients ∇f_N are L -Lipschitz with $\nabla f_N(x_\star) = 0$. By the Arzelà–Ascoli theorem and a diagonal argument, a subsequence has $f_N \rightarrow f$ and $\nabla f_N \rightarrow G$

locally uniformly. Then f is C^1 with $\nabla f = G$. Moreover, $f - (\mu/2)\|\cdot\|^2$ is convex as a pointwise limit of convex functions, and ∇f is L -Lipschitz as a limit of L -Lipschitz maps. So $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$. Since f_N interpolates the point t for every $N > t$, and $(x_\star, 0, f_\star)$ for every N , so does f . $\square$

Pasechniuk-Vilensky [22, Thm. 3.2] prove the interpolation of infinite data sets directly, and it follows from the extension theorem of Azagra and Mudarra [1] as well.

5.3 The window

From now on we set $W = X - \eta$ and $q = \eta/8$ in (8), so that Φ is a polynomial in h, u, η and X .

Lemma 4 (window). *Let $0 < \eta < X \leq 2/5$. Then $\Phi(h, u) > 0$ for all*

$$h \in [8X, (2 - \eta)(2 - X)] \quad \text{and} \quad u \in [2X(1 - \sqrt{X}/2), 2X(1 + \sqrt{X}/2)].$$

Proof. The polynomial Φ is concave in h , since the coefficient of h^2 is $-q$, and concave in u , since the coefficient of u^2 is $W^2 - (2 - \eta - X)^2 = -4(1 - \eta)(1 - X) < 0$, independently of h . A function on a rectangle that is concave in each variable separately attains its minimum at a corner, so it suffices to check the four corners. Write $u = 2X(1 + \nu)$ with $\nu = \pm\sqrt{X}/2$, so that $4\nu^2 = X$, and write $X = y^2$ and $\eta = \vartheta X$ with $(\vartheta, y) \in (0, 1) \times (0, \sqrt{2/5}]$. The following polynomial identities are checked symbolically. At the lower corner $h = 8X$,

$$\Phi(8X, u) = X^2[4X(X + 2 + 7\nu) + 2\vartheta E_1 + 2\vartheta^2 X E_2] = y^4 F_1(\vartheta, y), \quad (11)$$

where $E_1 = 8 - 5X + 2X\nu + 8X\nu^2 - 8X^2(1 + \nu)^2$ and $E_2 = 2X(1 + \nu) - 2\nu - 3$. At the upper corner $h = (2 - \eta)(2 - X)$,

$$\Phi((2 - \eta)(2 - X), u) = (1 - X(1 + \nu)) y^2 F_2(\vartheta, y), \quad (12)$$

where

$$F_2 = 4(2 + 4\nu - X(3 + 4\nu)) + \vartheta(6 - 9X + 11X^2 - 16X\nu(1 - X)) - \frac{\vartheta^2 X}{2}(6 + 3X - X^2).$$

Here $1 - X(1 + \nu) \geq 1 - 0.4 \cdot 1.32 > 0$. Both F_1 and F_2 are quadratics in ϑ with negative leading coefficients, namely $2y^2 E_2$, where $E_2 \leq 2X(1 + |\nu|) + 2|\nu| - 3 < 0$, and $-(y^2/2)(6 + 3y^2 - y^4)$. So they are concave in ϑ , and it suffices to check $\vartheta \in \{0, 1\}$. With $\nu = \pm y/2$ this leaves the univariate polynomials

$$\begin{aligned} F_1(0, y) &= 4y^2(y^2 \pm 7y/2 + 2), & F_1(1, y) &= 16 - 8y^2 \pm 14y^3 - 4y^4 \mp 14y^5 - 4y^6, \\ F_2(0, y) &= 8 \pm 8y - 12y^2 \mp 8y^3, & F_2(1, y) &= 14 \pm 8y - 24y^2 \mp 16y^3 + 19y^4/2 \pm 8y^5 + y^6/2. \end{aligned}$$

Each of $F_1(0, y)/y^2$, $F_1(1, y)$, $F_2(0, y)$ and $F_2(1, y)$ is positive at $y = 7/11$ and has no root in $[0, 7/11] \supset [0, \sqrt{2/5}]$, by exact Sturm counts. For instance, $y^2 - 7y/2 + 2 > 0$ for $y < 0.71$. Hence $F_1 > 0$ and $F_2 > 0$ on the box. $\square$

Appendix B gives an independent check of the same lemma by interval arithmetic.

5.4 Roots of unity in the window

Lemma 5 (roots of unity). *For every $X \in (0, 2/5]$ there is an integer $K \geq 4$ with $2X(1 - \sqrt{X}/2) \leq u_K \leq 2X(1 + \sqrt{X}/2)$, where $u_K = 2 \sin^2(\pi/K)$.*

Proof. Let $y = \sqrt{X}$, $u^+ = 2X(1 + y/2)$ and $u^- = 2X(1 - y/2)$, and let K be the least integer $K \geq 3$ with $u_K \leq u^+$. Since $u_3 = 3/2 > 1.06 > u^+$, we have $K \geq 4$ and $u_{K-1} > u^+$. As $\sin x/x$ decreases on $(0, \pi]$, $\sin(\pi/K)/\sin(\pi/(K-1)) \geq (K-1)/K$, so $u_K \geq (1 - 1/K)^2 u_{K-1} > (1 - 1/K)^2 u^+$. From $u_K \leq u^+$ and $\pi/K \leq \pi/4$ we get $1/K \leq \arcsin(\zeta)/\pi$ with $\zeta = \sqrt{u^+/2} = y\sqrt{1 + y/2}$. So $u_K \geq u^-$ follows from

$$\left(1 - \frac{\arcsin \zeta}{\pi}\right)^2 \geq \frac{u^-}{u^+} = \frac{1 - y/2}{1 + y/2}, \quad (13)$$

which we now prove for all $y \in (0, \sqrt{2/5}]$.

We first bound $\arcsin$ by a linear function. Since $y \leq \sqrt{2/5}$,

$$\zeta = y\sqrt{1 + y/2} \leq \sqrt{2/5}\sqrt{1 + \sqrt{2/5}/2} < 0.7256 < 3/4.$$

Since $\arcsin$ is convex on $[0, 1]$ and vanishes at 0, $\arcsin \zeta \leq (4\zeta/3) \arcsin(3/4)$ for $\zeta \in [0, 3/4]$. The alternating Taylor series gives

$$\sin 0.8481 \geq 0.8481 - \frac{0.8481^3}{6} + \frac{0.8481^5}{120} - \frac{0.8481^7}{5040} > 0.75002,$$

so $\arcsin(3/4) < 0.8481$. With $\pi > 3.1415$ this gives $\arcsin(\zeta)/\pi \leq 0.36\zeta$.

Next we reduce (13) to a polynomial inequality. Put $w = \sqrt{1 + y/2} = 1 + \delta$, where

$$0 < \delta \leq \sqrt{1 + \sqrt{2/5}/2} - 1 < 0.1473.$$

Then $y = 2(w^2 - 1)$ and $0.36\zeta = 0.72w(w^2 - 1)$. Since $u^-/u^+ = (2 - w^2)/w^2$, (13) follows from

$$w(1 - 0.72w(w^2 - 1)) \geq \sqrt{2 - w^2},$$

both sides being non-negative.

Finally, $w(w^2 - 1) = \delta(1 + \delta)(2 + \delta) \leq 2.4637\delta$ for $\delta \leq 0.1473$, and $0.72 \cdot 2.4637 \leq 1.78$. So the left side is at least $(1 + \delta)(1 - 1.78\delta) = 1 - 0.78\delta - 1.78\delta^2 > 0$. The right side is $\sqrt{1 - 2\delta - \delta^2} \leq 1 - \delta - \delta^2/2$. The difference of the two bounds is $\delta(0.22 - 1.28\delta) \geq 0$ for $\delta \leq 0.1718$. $\square$

Each constant in this proof is checked in exact arithmetic by `code/verify_rou_analytic.py`. Independently, `code/verify_lowerbound.py` checks (13) on $[0.1, \sqrt{2/5}]$ by interval arithmetic on 4000 subintervals.

5.5 Proof of Theorem A

If $\rho_Q \geq r$ there is nothing to prove, since $\rho_{\gamma, \beta}(\mathcal{F}_{\mu, L}) \geq \rho_Q$. Otherwise Lemma 1 gives $\eta < X \leq 2/5$ and $h \in (8X, (2 - \eta)(2 - X))$, with $W > 0$. Lemma 5 gives a $K \geq 4$ with u_K in the window of

Lemma 4, so $\Phi(h, u_K) > 0$. Lemma 2 then gives an $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^2)$ with $\|x_t\| = r^t$. Extending f by $(\mu/2)\|\cdot\|^2$ in the other coordinates gives the result in $\mathbb{R}^d$. Since $\|x_T\| = r^{T+1} \max\{\|x_0\|, \|x_{-1}\|\}$, the bound (2) fails for every $\rho < r$. $\square$

5.6 Rest start

Proof of Corollary 2. We first treat quadratics. On $(\lambda/2)x^2$ started at rest, $x_1 = (1 - h\lambda)x_0$, and x_t is a linear combination of z_1^t and z_2^t when the roots are distinct. If the coefficient of the dominant root z_1 vanishes, then $1 - h\lambda = z_2$, so $z_1 = \beta$, and either $\beta = 0$, which is gradient descent with $x_t = (1 - h\lambda)^t x_0$, or $z_2 = 1$, and the method does not converge. The case of a double root is similar. So heavy-ball started at rest still converges at the rate $r(\lambda)$ or does not converge, and $\rho_{\gamma, \beta}^{\text{rest}} \geq \min\{\rho_Q, 1\}$.

Now suppose $\rho_Q < r$, and take K and the planar spiral data of the proof of Theorem A, in the plane $\mathbb{C}$. Replace the first gradient by $g_0 = (1 - z)/h$, so that heavy-ball started at rest at $x_0 = 1$ lands on $x_1 = z$. Add a third coordinate, on which f is $(\lambda/2)y^2$ for some $\lambda \in (q, 1)$, started at $y_{-1} = y_0 = Y$, so that the space is $\mathbb{C} \times \mathbb{R} \cong \mathbb{R}^3$. Then $y_t = Y a_t$ with $a_{-1} = a_0 = 1$ and $a_{t+1} = (1 + \beta - h\lambda)a_t - \beta a_{t-1}$. The inequalities (3) split over orthogonal blocks,

$$I(i, j) = I^\circ(i, j) + c_\lambda(y_i - y_j)^2, \quad c_\lambda = \frac{(\lambda - q)(1 - \lambda)}{2(1 - q)} > 0,$$

with $y_\star = 0$, where I° is computed from the planar data. For $i, j \geq 1$ and the minimizer, $I^\circ \geq 0$ by Lemma 2. The pairs that involve the index 0 have $|I^\circ| \leq M_0$ for a constant M_0 , because all data are bounded ($r < 1$). Each $a_j(\lambda) - 1$ with $j \geq 1$ is a polynomial of degree j in λ with leading coefficient $(-h)^j \neq 0$. So for all but countably many λ , $a_j \neq 1$ for every $j \geq 1$, and since $a_j \rightarrow 0$, this gives $\delta_\lambda = \inf_{j \geq 1} (1 - a_j)^2 > 0$. Taking $Y^2 \geq M_0/(c_\lambda \min\{\delta_\lambda, 1\})$ makes every inequality hold. Lemma 3 gives an $f \in \mathcal{F}_{\mu, L}(\mathbb{R}^3)$, and $\|x_t\| \geq |z^t| = r^t$ with $x_{-1} = x_0 = (1, Y)$. $\square$

This is the construction of Pasechniuk-Vilensky [22, §5.1 and Lemma A.5], specialized to one decaying wave.

6 Proof of the Upper Bound

This section proves Theorem B. Throughout, $L = 1$, $\ell = 1 - q$, $x_\star = 0$ and $f(x_\star) = 0$.

6.1 The tuning on quadratics

With the tuning (4), $1 + \beta - hq = r + \beta/r$ is an identity in s , so the polynomial $z^2 - (1 + \beta - hq)z + \beta$ of the direction of curvature q has the roots r and $\beta/r = r - 4s$. Equivalently, $X = 1 - \beta/r = s(4 + Cs)$ and $h = CX$, so the tuning lies on the edge $h = CX$ of §4.1 for its own rate. Running heavy-ball on $(\mu/2)\|x\|^2$ from $x_{-1} = x_0/r$ gives $x_t = r^t x_0$. Started at rest, both roots are excited, and by the argument in the proof of Corollary 2 the coefficient of r^t is non-zero. In the direction of curvature 1 the roots are complex with modulus $\sqrt{\beta} < r$, because the discriminant $(1 + \beta - h)^2 - 4\beta$ is negative for $0 < s \leq 1/20$. So $\rho_Q = r$, and it remains to show that no function in $\mathcal{F}_{\mu,L}$ does worse than the quadratics.

6.2 The Lyapunov function

Let $\varphi = f - (q/2)\|\cdot\|^2$. It is convex and ℓ -smooth, with $\varphi(0) = 0$ and $\nabla\varphi(0) = 0$. Let $p_t = \nabla\varphi(x_t)$, so that $\nabla f(x_t) = p_t + qx_t$ and heavy-ball reads

$$x_{t+1} = (1 + \beta - hq)x_t - \beta x_{t-1} - h p_t. \quad (14)$$

Convexity and smoothness of φ give, for consecutive iterates and for the minimizer,

$$\begin{aligned} I_t^1 &= \varphi(x_{t-1}) - \varphi(x_t) - \langle p_t, x_{t-1} - x_t \rangle - \frac{1}{2\ell} \|p_{t-1} - p_t\|^2 \geq 0, \\ I_t^2 &= -\varphi(x_t) + \langle p_t, x_t \rangle - \frac{1}{2\ell} \|p_t\|^2 \geq 0. \end{aligned}$$

With $\xi_t = (x_t, x_{t-1}, p_{t-1})$ and a symmetric 3×3 matrix $P = (P_{ij})$, define

$$V_t = \xi_t^\top (P \otimes I_d) \xi_t + \varphi(x_{t-1}). \quad (15)$$

Proposition 1 (dissipation identity). *Let $P = P(s)$ be the matrix of Appendix A. Along (14),*

$$r^2 V_t - V_{t+1} - r^2 I_t^1 - (1 - r^2) I_t^2 = v_t^\top (F \otimes I_d) v_t, \quad \text{where} \quad v_t = (x_t, x_{t-1}, p_{t-1}, p_t), \quad (16)$$

for a symmetric 4×4 matrix $F = F(s)$ of rational functions of s . For $0 < s \leq 1/20$, $F \geq 0$ and $G = P + e_3 e_3^\top / (2\ell) \succ 0$.

The values of φ cancel in (16). The multiplier r^2 of I_t^1 removes the term $r^2 \varphi(x_{t-1})$ of $r^2 V_t$, and since the two multipliers add up to 1 they remove the term $-\varphi(x_t)$ of $-V_{t+1}$. The identity is between quadratic forms with scalar coefficients, so it holds in every dimension once it holds for $d = 1$.

Proof of Theorem B from Proposition 1. By (16), $V_{t+1} \leq r^2 V_t - r^2 I_t^1 - (1 - r^2) I_t^2 \leq r^2 V_t$, so $V_t \leq r^{2t} V_0$. Since $\varphi(x) \geq \|\nabla\varphi(x)\|^2 / (2\ell)$, we have $V_t \geq \xi_t^\top (G \otimes I_d) \xi_t \geq \lambda_{\min}(G) \|x_t\|^2$. Moreover $\|p_{-1}\| \leq \ell \|x_{-1}\|$ and $\varphi(x_{-1}) \leq (\ell/2) \|x_{-1}\|^2$, so

$$V_0 \leq \|P\| (\|x_0\|^2 + (1 + \ell^2) \|x_{-1}\|^2) + \frac{\ell}{2} \|x_{-1}\|^2,$$

where $\|P\|$ is the spectral norm. Hence $\|x_t\| \leq Mr^t(\|x_0\| + \|x_{-1}\|)$ with

$$M^2 = \frac{\|P\|(1 + \ell^2) + \ell/2}{\lambda_{\min}(G)}.$$

Together with the quadratic lower bound above, $\rho_{\gamma,\beta} = \rho_{\gamma,\beta}^{\text{rest}} = r$. $\square$

6.3 Structure of the certificate

On $(\mu/2)\|x\|^2$ we have $p \equiv 0$, and the slow solution has $x_{t-1} = x_t/r$. All the I 's vanish there, and V decays at the rate r^2 . So the right side of (16) vanishes at $v_0 = (1, 1/r, 0, 0)$, and $F \geq 0$ forces $Fv_0 = 0$. This determines P_{11} , P_{12} and P_{13} from the other entries (Appendix A). It makes $F \geq 0$ equivalent to $F_3 \geq 0$, where F_3 is the lower-right 3×3 block of F . Indeed, writing $v = v^{(1)}v_0 + w$ with $w^{(1)} = 0$ gives $v^\top Fv = w^\top Fw$. The three free entries are

$$\begin{aligned} P_{22} &= \frac{1}{s} \left(\frac{1}{64} + \frac{5}{32}s + \frac{5}{4}s^2 + 26s^3 \right), \\ P_{23} &= - \left(\frac{1}{4} + \frac{9}{8}s + 11s^2 + 248s^3 \right), \\ P_{33} &= s(2 + 14s + 441s^2 + 3210s^3). \end{aligned}$$

To leading order in s , using $\gamma \approx 32s$ and $\varphi + (q/2)\|\cdot\|^2 = f$,

$$V_t = f(x_{t-1}) + \frac{\|x_t - x_{t-1}\|^2}{2\gamma} + \frac{1}{2}\langle x_t - x_{t-1}, \nabla\varphi(x_{t-1}) \rangle + 2s\|\nabla\varphi(x_{t-1})\|^2 + \dots, \quad (17)$$

where the dots stand for the cross terms $(s/2)\langle x_t - x_{t-1}, x_{t-1} \rangle$ and $4s^2\langle x_{t-1}, \nabla\varphi(x_{t-1}) \rangle$ to leading order, and for corrections of higher order in s to the coefficients shown. The certificate is therefore a small perturbation of the mechanical energy of the heavy ball, potential plus kinetic.

6.4 Exact verification

In the basis

$$U = \begin{pmatrix} 1 & 4 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix}, \quad \det U = 2,$$

the leading principal minors of $U^\top F_3 U$ and of G are rational functions of s with the expansions

$$\begin{aligned} U^\top F_3 U : & \quad \frac{1}{8} + \frac{3}{4}s + O(s^2), & \quad \frac{s}{4} + O(s^2), & \quad 2s^3 + O(s^4), \\ G : & \quad \frac{1}{64s} + O(1), & \quad \frac{s}{128} + O(s^2), & \quad \frac{s}{256} + O(s^2). \end{aligned}$$

Write each minor as $n(s)/d(s)$ in lowest terms. We verify by an exact Sturm count that the polynomial nd has no root in $(0, 1/20]$, and that it is positive at $s = 1/20$. So all six minors are positive on $(0, 1/20]$, and Sylvester's criterion gives $U^\top F_3 U > 0$ and $G > 0$. The identity (16), the null vector, the root r and the sign of the discriminant in the direction of curvature 1 are all checked symbolically or by exact root counting. This proves Proposition 1, and with it Theorem B.

6.5 Exact range

The minors of G and the first minor of $U^\top F_3 U$ have no sign change on $(0, 1/5)$. The second minor of $U^\top F_3 U$ first changes sign at $s \approx 0.060896$ ($\kappa \approx 269.67$) and the third at $s_1 \approx 0.0529041$, that is, $\kappa = 1/s_1^2 \approx 357.2901$. The discriminant for curvature 1 stays negative up to $s \approx 0.122$. At s_1 the third minor vanishes and the other five are positive, so $F_3 \geq 0$ by continuity and the conclusions of Proposition 1 still hold. For every $s \in (s_1, 1)$, either P is undefined or some minor is negative, so that $F \not\geq 0$ or $G \not> 0$. So this certificate proves Theorem B for every $\kappa \geq 1/s_1^2$ and for no smaller κ . We state $\kappa \geq 400$ for simplicity. For $64.82 < \kappa < 1/s_1^2$ the rate of the same tuning still follows from Pasechniuk-Vilensky [22] (Remark 2), but there $8 - 64/\sqrt{\kappa}$ is below 4.62, while the certified tuning at $\kappa = 10^2$ in Table 2 already gives 5.032.

6.6 How the certificate was found

For fixed s , the matrices P with $F \geq 0$ and $G > 0$ form a spectrahedron, which we explored numerically with a semidefinite programming solver. Imposing the null vector leaves three unknowns. At $s = 1/50, 1/100$ and $1/400$ its extent in P_{22} and in P_{23} is at most about $12s$ times the size of these entries, so the set is thin, while P_{33} can vary by about three times its size and change sign. We computed its analytic center along the family $C = 8 - \tau s$, whose rate is $1 - 8/\kappa + \tau\kappa^{-3/2}$, fitted the expansions of the three entries in powers of s , rounded the coefficients to simple rationals, and verified the result exactly. None of this search is part of the proof. A formal expansion of the minors suggests that, within this certificate class and the family $\beta = r(r - \sigma s)$, feasibility as $s \rightarrow 0$ requires $\tau > 6\sigma + 96/\sigma$, which is smallest at $\sigma = 4$, where it reads $\tau > 48$. So the value $\tau = 64$ is not far from what lag-one certificates of this form can give. We did not try to prove this threshold.

7 Computer-Assisted Steps

This section states what was checked by computer, how, and on what range. No step of either proof relies on floating-point arithmetic.

7.1 What was checked

Table 1 lists every computer-assisted step of the proofs, together with the auxiliary exact computations quoted in the text. “Symbolic” means an identity between polynomials or rational functions with rational coefficients, verified by expanding the difference to zero. “Sturm” means an exact count of the real roots of a univariate polynomial with rational coefficients on a closed interval, together with an exact evaluation of its sign at one point. “Interval” means outward-rounded interval arithmetic.

Table 1: Computer-assisted steps. “Range” is the exact parameter range on which the check holds. The scripts are in code/.

Claim	What is checked	Range	Method	Script
Lemma 1	the two Schur–Cohn identities	all parameters	symbolic	verify_lowerbound
Lemma 2	reduction (9), formulas (7), identity for Φ	all parameters	symbolic	verify_lowerbound
	at $r = 1$, Φ gives the region of Goujaud et al. [11, Thm. 3.5]	all parameters	symbolic	verify_lowerbound, verify_symbolic
Lemma 4	corner identities (11) and (12), u^2 -coefficient	all parameters	symbolic	verify_lowerbound
	8 end polynomials and 4 concavity coefficients have no root	$y \in [0, 7/11]$	Sturm	verify_lowerbound
	(independent) $F_1, F_2 > 0$ on the box	$[0, 1] \times [0, \sqrt{2/5}]$	interval, 76 boxes	verify_lowerbound
Lemma 5	the constants of the proof	as stated	exact rationals	verify_rou_analytic
	(independent) inequality (13)	$y \in [0.1, \sqrt{2/5}]$	interval, 4000 pieces	verify_lowerbound
Proposition 1	identity (16), null vector v_0 , r is a root for curvature q	all s	symbolic	verify_achievability
	6 leading minors positive	$s \in (0, 1/20]$	Sturm	verify_achievability
	discriminant for curvature 1 negative	$s \in (0, 1/20]$	Sturm	verify_achievability
	first sign change of each minor, validity at s_1 and failure for every larger s	$s \in (0, 1)$	real-root isolation	verify_achievability --range
Remark 2	membership in the region Q of Pasechniuk-Vilensky [22] (sufficient conditions)	$s \in (0, 0.12420979]$	Sturm	verify_pv_membership
Table 2	fixed- κ certificates (Appendix C)	$\kappa \in \{10^2, 400, 10^3, 10^4\}$	exact rationals	fixed_kappa_certificates
§9.3	spiral bound for the corner tuning	$\kappa = 10^2, \dots, 10^5$	interval	ghadimi_corner

7.2 Size of the computations and trust base

In Lemma 4 the polynomials have degree at most 6 and small rational coefficients. They are printed in the proof and in Appendix B, and can be checked by hand. In Proposition 1 the six products nd have degrees 12, 26 and 34 (minors of $U^\top F_3 U$) and 14, 18 and 26 (minors of G), with integer coefficients of up to 21 digits. The discriminant for curvature 1 has degree 12. The entries of P are printed in Appendix A. The tuning, the matrices P, F, U and G , the six minors with their products nd , and the discriminant are written in exact form to `results/optimal/achievability.json`, so the check can be repeated in any computer algebra system.

All exact computations use SymPy 1.14 under Python 3.12, namely rational arithmetic, polynomial expansion, `Poly.count_roots` (Sturm sequences) and `Poly.real_roots` (real-root isolation). Interval arithmetic uses `mpmath` 1.3 (`mpmath.iv`, 30 to 40 digits, outward rounding). Each check runs in under a minute on a laptop.

The proofs are not formalized in a proof assistant, so the trust base includes the correctness of the polynomial routines of SymPy and of `mpmath.iv`. We reduce this dependence in three ways. The lower bound does not need interval arithmetic, because Lemma 5 has a proof by hand and the univariate polynomials of Lemma 4 are short, and both lemmas are checked a second time by an independent route. The upper bound depends on exact root counts for seven polynomials of degree at most 34, which any computer algebra system can repeat from the published data. Finally, the certificate of Theorem B was found numerically, but only its exact form is used. Floating point appears only in the numerical columns of Table 2 and in the tests below, and none of these is used in a proof.

7.3 Floating-point tests

As a sanity check, separate from the proofs, we built the objects of the proofs in floating point. For 400 random tunings with $q \in [10^{-7}, 0.042]$ that satisfy the conditions of Lemma 1, we built the spiral with the K of Lemma 5 and checked (10) directly for all $k \leq 3000$. The smallest relative slack was 0.15. For 20 random tunings with $\rho_Q < r$ and $q \in [10^{-4}, 0.04]$, we built the rest-start construction of Corollary 2 in $\mathbb{R}^3$, verified that heavy-ball started at rest reproduces it, and checked all pairwise inequalities (3) among up to 600 points and the minimizer. The smallest slack was $4.9 \cdot 10^{-3}$.

8 Finite Condition Numbers

Corollary 1 is asymptotic. This section compares it with what is known at finite κ .

8.1 Exact certificates at fixed condition numbers

The certificate class of §6 applies to a single rational tuning at a fixed κ , where every entry is a rational number. For $\kappa = 10^2, 400, 10^3$ and 10^4 we searched numerically for the fastest tuning in the region Q of Pasechniuk-Vilensky [22], rounded it to a rational tuning whose rate r is a root of the polynomial for curvature q , solved a small semidefinite program for the free entries of P , rounded them, and verified the certificate exactly (Appendix C). This proves $c(\kappa) \geq 5.032, 6.173, 6.734$ and 7.548 at these values of κ , in every dimension. These constants exceed $8 - 64/\sqrt{\kappa}$ at the same κ , so the tuning of Theorem B is not optimal at finite κ .

8.2 A numerical bracket

We computed two further quantities in floating point, which are not certified. On the side of tunings, we minimized the exact rate ρ_Q over a grid of tunings in the region Q of Pasechniuk-Vilensky [22, Thm. 7.2(b)], where their theorem gives $\rho = \rho_Q$ (code/pv_region_search.py). On the side of functions, for each tuning we took the larger of ρ_Q and the rate of the slowest decaying spiral, and minimized over a grid of tunings refined near the optimum. The spirals are either root-of-unity spirals as in Lemma 2 or general log-spirals $x_t = r^t e^{it\theta}$ with any θ , and for the latter the function values $f_t = Fr^{2t}$ may have any scale F . Table 2 collects these numbers with the proved bounds.

Table 2: Bounds on $c(\kappa) = \kappa(1 - \rho^*(\kappa))$ for $d \geq 2$. Lower bounds on c come from tunings, and upper bounds from functions that slow every tuning down. Lower bounds are rounded down and upper bounds up. The columns marked “num.” are floating point.

κ	lower bounds on c (tunings)			upper bounds on c (functions)		
	Theorem B	exact (Appendix C)	best on Q (num.)	log-spirals (num.)	root-of-unity spirals (num.)	Theorem A
10^2	–	5.032	5.034	5.37	5.44	8
400	4.800	6.173	6.175	–	6.50	8
10^3	5.976	6.734	6.736	6.98	6.99	8
10^4	7.360	7.548	7.549	–	7.66	8
10^5	7.797	–	7.851	–	–	8
10^6	7.936	–	–	–	–	8
∞	8					8

8.3 Earlier numerical values

The thesis of Goujaud [9, Table 10.1] reports numerical optima at $\kappa = 20, 100$ and 1000 . With Remark 1 they convert to $c \approx 3.42, 5.03$ and 7.49 , if they refer to the rate of the distance. The value at $\kappa = 100$ agrees with Table 2, and the tuning listed with it, $\beta = 0.65$ and $\gamma L = 1.57$, is close to our certified tuning $\beta = 0.650$ and $\gamma L = 1.587$ in Table 4. We could not reconcile the value at $\kappa = 1000$. The tuning listed with it, $\beta = 0.839$ and $\gamma = 1.01$ read as γL , already has a rate on quadratics that

corresponds to $c \approx 6.5$, and 7.49 exceeds our numerical spiral bound 6.98. The thesis may use a different notion of rate. Van Scoy and Lessard [33] find $c \approx 5$ at $\kappa = 100$, in agreement with Table 2. The values of Lessard et al. [15, Fig. 8], for the restricted family with step $1/L$, lie below our numerical spiral bounds for $\kappa \lesssim 4000$, as valid bounds must. Their values of $1 - \rho$ appear to lie on a grid of spacing about 2^{-12} , which limits the resolution in c to about $\kappa/4096$.

8.4 Shape of the tuning and the second-order term

Expanding (4) gives

$$\gamma L = 32s - 192s^2 + O(s^3), \quad 1 - \beta = 4s + 16s^2 + O(s^3), \quad \frac{\gamma}{1 - \beta} \rightarrow \frac{8}{L}.$$

Polyak's tuning has $\gamma L = 4/(1 + s)^2$ and $1 - \beta = 4s/(1 + s)^2$. The momenta agree to first order, and the step of Theorem B is smaller by a factor of about $\sqrt{\kappa}/8$. At $\kappa = 10^4$ the tuning is $\gamma L = 0.2998$ and $\beta = 0.9586$, with rate $r = 1 - 7.36/\kappa$. The certified tunings of Table 4 have the same shape, $\beta \approx r(r - \sigma s)$, with $\sigma = 2.65, 3.20, 3.45$ and 3.85 at $\kappa = 10^2, 400, 10^3$ and 10^4 against $\sigma = 4$ in Theorem B, so they use slightly more momentum.

Corollary 1 gives $\rho^* = 1 - 8/\kappa + c_2\kappa^{-3/2} + \dots$ with $c_2 \in [0, 64]$, if such an expansion exists. At $\kappa = 10^3$ and 10^4 the numerical bracket gives values of $\sqrt{\kappa}(8 - c(\kappa))$ between about 32 and 45. This suggests that $c_2 > 0$, so that $\rho^*(\kappa)$ is strictly above $1 - 8/\kappa$ at large finite κ . Only the leading constant is proved.

9 Discussion and Limitations

9.1 What the results settle

For $d \geq 2$, the optimal worst-case rate of heavy-ball with constant parameters on $\mathcal{F}_{\mu,L}$ is $1 - 8/\kappa$ to first order, and an explicit family of tunings attains it. On this class, momentum divides the number of iterations of gradient descent by four asymptotically, whereas on quadratics it divides it by a factor of order $\sqrt{\kappa}$. In our family the momentum is Polyak's to first order and the step is about $\sqrt{\kappa}/8$ times smaller, and the best tunings we found numerically at fixed κ have the same shape (§8). We have not proved that every asymptotically optimal family must look like this.

9.2 Open questions

Theorem B is optimal only as $\kappa \rightarrow \infty$. At fixed κ better tunings exist (§8), and we have no closed form for them, so the problem posed by Goujaud [9, §10.3.3] is solved to first order only. The second-order term is open as well. We prove $c_2 \in [0, 64]$, while the numerics point to values between about 32 and 45. A second-order analysis of decaying spirals, possibly with several frequencies as in Pasechniuk-Vilensky [22], could raise the lower end. Lag-one certificates of our form seem unable to lower the upper end below 48 (§6.6), so this may need certificates with longer memory. For κ below 357.2901 ... we have exact certificates at isolated values only.

The spirals need two dimensions. In one dimension Theorem B still applies, so $8 - 64/\sqrt{\kappa} \leq c_1(\kappa) \leq 50/3 + o(1)$ (Appendix D), and whether $c_1(\kappa) \rightarrow 8$ is open. A proof of $c_1 \leq 8 + o(1)$ would require functions of one variable that slow every tuning down as much as the decaying spirals do, and by Proposition 2 single-jump cycles do not suffice. Similarly, Corollary 2 uses a third coordinate, and we do not know whether the rest-start rate on $\mathcal{F}_{\mu,L}(\mathbb{R}^2)$ obeys the same lower bound. Our functions are not twice differentiable at the minimizer (§2), whereas the one-dimensional lower bound of Liu et al. [16] holds for C^∞ functions. Whether $1 - 8/\kappa$ persists on smoother subclasses of $\mathcal{F}_{\mu,L}$ is open.

The results concern constant (γ, β) . Variable parameters and restarts are not covered, and neither are methods that evaluate the gradient elsewhere, such as Nesterov's method and the triple momentum method, which accelerate. The results are worst-case statements for exact gradients on smooth strongly convex functions, and they do not directly describe the stochastic training of neural networks.

9.3 The constant 8 in the region of Ghadimi et al.

Goujaud et al. [11, Lemma 2.5] show that the best rate on quadratics over the region where Ghadimi et al. [8, Thm. 4] prove global convergence on $\mathcal{F}_{\mu,L}$ is $1 - 8/\kappa + o(1/\kappa)$. The two constants are related but describe different tunings. For small steps, the region of Ghadimi et al. is $\gamma L \lesssim 4(1 - \beta)$. In the triangle of §4.1 for the rate $1 - c/\kappa$, the corner where the momentum edge $X = \eta$ meets the edge $h = cX$ of the effective step is $\beta = r^2$, $\gamma L = c^2/\kappa$, with $1 - \beta \approx 2c/\kappa$. It lies in the region of Ghadimi et al. if and only if $c \lesssim 8$. In floating point the minimum of their Lemma 2.5 is attained at this corner, with $\gamma L = 52.3/\kappa$, $62.6/\kappa$ and $63.9/\kappa$ at $\kappa = 10^2$, 10^3 and 10^4 . The step printed in the lemma, $2(1 + \beta^*)/(L + \mu)$, exceeds $2/L$ and lies outside the region. Our tuning has $\gamma L \approx 32/\sqrt{\kappa}$ instead. On $\mathcal{F}_{\mu,L}$ the corner $\beta = (1 - 8/\kappa)^2$, $\gamma L = 64/\kappa$ is much slower than on quadratics. Decaying spirals (Lemma 2, which holds for every r) prove $c \leq 3.681, 3.972, 3.998$ and 4.000 for it at $\kappa = 10^2, 10^3, 10^4$ and 10^5 (interval arithmetic, code/ghadimi_corner.py). The

lag-one criterion of Pasechniuk-Vilensky [22] gives $c \geq 2.92, 3.14, 3.16$ and 3.16 there in floating point. Their Lemma 2.5 therefore does not give 8 on $\mathcal{F}_{\mu,L}$, and we know of no deeper connection between the two constants.

9.4 Limitations

The proofs of Lemma 4 and Proposition 1 use exact computer algebra (§7). The numerical results at finite κ are in floating point and partly rely on a recent result that has not yet been refereed [22], but none of them is used in a proof. Our literature search, up to October 15, 2026, found no earlier statement of the limit 8. Given the concurrent activity on this problem, independent work on the same question is possible.

10 Conclusion

With constant parameters and the best tuning, heavy-ball converges on L -smooth μ -strongly convex functions at the rate $1 - 8/\kappa + O(\kappa^{-3/2})$ in every dimension $d \geq 2$, four times faster than gradient descent. The lower bound is a decaying root-of-unity spiral, which exists for every tuning that is faster than $1 - 8/\kappa$ on quadratics. The upper bound is an explicit tuning, with the momentum of Polyak's tuning to first order and a step about $\sqrt{\kappa}/8$ times smaller, certified by a Lyapunov function close to the mechanical energy and verified exactly for all $\kappa \geq 400$. The constant 8 comes from a trade-off between rotation and curvature. Along a spiral the gradient must rotate, and convexity allows rotation only in exchange for curvature. This exchange is cheapest at half the largest curvature. The optimal tuning at finite κ , the second-order term and the constant in one dimension remain open. The same pair of tools, decaying spirals and lag-one certificates, may give the optimal constants of other methods with a fixed structure.

References

- [1] D. Azagra and C. Mudarra. An extension theorem for convex functions of class $C^{1,1}$ on Hilbert spaces. *Journal of Mathematical Analysis and Applications*, 446(2):1167–1182, 2017. [doi:10.1016/j.jmaa.2016.09.015](https://doi.org/10.1016/j.jmaa.2016.09.015); [arXiv:1603.00241](https://arxiv.org/abs/1603.00241).
- [2] A. Badithela and P. Seiler. Analysis of the heavy-ball algorithm using integral quadratic constraints. In *American Control Conference (ACC)*, pages 4081–4085, 2019. [doi:10.23919/ACC.2019.8814459](https://doi.org/10.23919/ACC.2019.8814459).
- [3] B. Can and M. Gürbüzbalaban. Entropic risk-averse generalized momentum methods. *Optimization Methods and Software*, 40(6):1535–1583, 2025. [doi:10.1080/10556788.2025.2549356](https://doi.org/10.1080/10556788.2025.2549356); [arXiv:2204.11292](https://arxiv.org/abs/2204.11292).
- [4] M. Danilova, A. Kulakova, and B. Polyak. Non-monotone behavior of the heavy ball method. In *Difference Equations and Discrete Dynamical Systems with Applications*, volume 312 of *Springer Proceedings in Mathematics & Statistics*, pages 213–230. Springer, 2020. [doi:10.1007/978-3-030-35502-9_9](https://doi.org/10.1007/978-3-030-35502-9_9); [arXiv:1811.00658](https://arxiv.org/abs/1811.00658).
- [5] Y. Drori and A. Taylor. On the oracle complexity of smooth strongly convex minimization. *Journal of Complexity*, 68:101590, 2022. [doi:10.1016/j.jco.2021.101590](https://doi.org/10.1016/j.jco.2021.101590); [arXiv:2101.09740](https://arxiv.org/abs/2101.09740).
- [6] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1–2):451–482, 2014. [doi:10.1007/s10107-013-0653-0](https://doi.org/10.1007/s10107-013-0653-0); [arXiv:1206.3209](https://arxiv.org/abs/1206.3209).
- [7] K. Fotopoulos, M. Helcig, K. Deck, and J. Alsbach. Cycling and non-acceleration of the heavy-ball method in one dimension. ODI Blog, ETH Zürich, AI for Math and Optimization Seminar, No. 01, July 7, 2026. <https://odi.inf.ethz.ch/blog/heavy-ball-one-dimension/>, accessed October 14, 2026.
- [8] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *European Control Conference (ECC)*, pages 310–315, 2015. [doi:10.1109/ECC.2015.7330562](https://doi.org/10.1109/ECC.2015.7330562). We cite theorem numbers from [arXiv:1412.7457v1](https://arxiv.org/abs/1412.7457v1).
- [9] B. Goujaud. *Constructive Approaches to Worst-Case Complexity Analyses of Gradient Methods for Convex Optimization: Contributions, New Insights, and Novel Results*. PhD thesis, Institut Polytechnique de Paris, 2024. NNT 2024IPPAX039, HAL [tel-04746799](https://tel.archives-ouvertes.fr/tel-04746799).
- [10] B. Goujaud, A. Dieuleveut, and A. Taylor. Counter-examples in first-order optimization: A constructive approach. *IEEE Control Systems Letters*, 7:2485–2490, 2023. [doi:10.1109/LCSYS.2023.3286277](https://doi.org/10.1109/LCSYS.2023.3286277); [arXiv:2303.10503](https://arxiv.org/abs/2303.10503).
- [11] B. Goujaud, A. Taylor, and A. Dieuleveut. Provable non-accelerations of the heavy-ball method. *Mathematical Programming*, 2025. Online first, October 27, 2025, [doi:10.1007/s10107-025-02269-2](https://doi.org/10.1007/s10107-025-02269-2). We cite theorem numbers from [arXiv:2307.11291v2](https://arxiv.org/abs/2307.11291v2).
- [12] B. Goujaud, A. Taylor, and A. Dieuleveut. Open problem: Two riddles in heavy-ball dynamics, 2025. [arXiv:2502.19916](https://arxiv.org/abs/2502.19916).
- [13] C. He and S. Zhang. Heavy-ball method under randomized schedules, 2026. [arXiv:2609.09743](https://arxiv.org/abs/2609.09743).
- [14] L. Lessard and P. Seiler. Direct synthesis of iterative algorithms with bounds on achievable worst-case convergence rate. In *American Control Conference (ACC)*, pages 119–125, 2020. [doi:10.23919/ACC45564.2020.9147401](https://doi.org/10.23919/ACC45564.2020.9147401); [arXiv:1904.09046](https://arxiv.org/abs/1904.09046).
- [15] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. [doi:10.1137/15M1009597](https://doi.org/10.1137/15M1009597); [arXiv:1408.3595](https://arxiv.org/abs/1408.3595).
- [16] K. Liu, Q. Wu, Z. Sun, M. Chen, S. Li, and B. Qu. Sharp initialization guarantees for the heavy-ball method, 2026. [arXiv:2610.10369](https://arxiv.org/abs/2610.10369).
- [17] J. Ma and J. Zhang. A lower bound for the heavy-ball method on smooth convex functions, 2026. [arXiv:2609.08656](https://arxiv.org/abs/2609.08656).
- [18] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997. [doi:10.1109/9.587335](https://doi.org/10.1109/9.587335).

- [19] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- [20] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.
- [21] Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, second edition, 2018. doi:10.1007/978-3-319-91578-4.
- [22] D. Pasechniuk-Vilensky. Two answers on heavy-ball dynamics, 2026. arXiv:2610.09912; listed on arXiv as Dmitry Pasechnyuk-Vilensky.
- [23] S. Pokutta and B. Stellato. The exact worst-case asymptotic rate of Nesterov acceleration for strongly convex optimization, 2026. arXiv:2610.15125.
- [24] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi:10.1016/0041-5553(64)90137-5.
- [25] J. M. Sanz-Serna and K. C. Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3):1542–1565, 2021. doi:10.1137/20M1364138; arXiv:2009.00673.
- [26] P. Seiler and J. Carrasco. Construction of periodic counterexamples to the discrete-time Kalman conjecture. *IEEE Control Systems Letters*, 5(4):1291–1296, 2021. doi:10.1109/LCSYS.2020.3033443; arXiv:2009.02468.
- [27] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195(1–2):79–148, 2022. doi:10.1007/s10107-021-01681-8; arXiv:1810.08907.
- [28] A. Taylor and Y. Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1–2):557–594, 2023. doi:10.1007/s10107-022-01839-y; arXiv:2101.09741.
- [29] A. Taylor, B. Van Scoy, and L. Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4897–4906, 2018. arXiv:1803.06073.
- [30] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1–2):307–345, 2017. doi:10.1007/s10107-016-1009-3; arXiv:1502.05666.
- [31] V. Ugrinovskii, I. R. Petersen, and I. Shames. Global convergence and asymptotic optimality of the heavy ball method for a class of nonconvex optimization problems. *IEEE Control Systems Letters*, 6: 2449–2454, 2022. doi:10.1109/LCSYS.2022.3163408; arXiv:2202.02914.
- [32] M. Upadhyaya, S. Banert, A. B. Taylor, and P. Giselsson. Automated tight Lyapunov analysis for first-order methods. *Mathematical Programming*, 209(1–2):133–170, 2025. doi:10.1007/s10107-024-02061-8; arXiv:2302.06713.
- [33] B. Van Scoy and L. Lessard. The speed–robustness trade-off for first-order methods with additive gradient noise, 2021. arXiv:2109.05059 (v3, November 2025).
- [34] B. Van Scoy and L. Lessard. The fastest known first-order method for minimizing twice continuously differentiable smooth strongly convex functions. *IEEE Control Systems Letters*, 9:655–660, 2025. doi:10.1109/LCSYS.2025.3576681; arXiv:2506.01168.
- [35] B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018. doi:10.1109/LCSYS.2017.2722406.
- [36] J.-K. Wang, C.-H. Lin, A. Wibisono, and B. Hu. Provable acceleration of heavy ball beyond quadratics for a class of Polyak–Łojasiewicz functions when the non-convexity is averaged-out. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22839–22864, 2022. arXiv:2206.11872; v3 adds an erratum (App. J).
- [37] G. Zames and P. L. Falb. Stability conditions for systems with monotone and slope-restricted nonlinearities. *SIAM Journal on Control*, 6(1):89–108, 1968. doi:10.1137/0306007.

- [38] J. Zhang, J. Carrasco, and W. P. Heath. Duality bounds for discrete-time Zames–Falb multipliers. *IEEE Transactions on Automatic Control*, 67(7):3521–3528, 2022. [doi:10.1109/TAC.2021.3095418](https://doi.org/10.1109/TAC.2021.3095418); [arXiv:2008.11975](https://arxiv.org/abs/2008.11975).

Appendix

A The Certificate of Theorem B

As polynomials in $s = \kappa^{-1/2}$, the tuning of Theorem B is

$$\begin{aligned} r &= 1 - 8s^2 + 64s^3, \\ \gamma L &= 32s - 192s^2 - 1024s^3 + 4096s^4, \\ \beta &= 1 - 4s - 16s^2 + 160s^3 - 192s^4 - 1024s^5 + 4096s^6. \end{aligned}$$

Table 3 gives the tuning and the constant M of Theorem B at a few values of κ .

Table 3: The tuning and the constant of Theorem B at sample values of κ , rounded, with M rounded up.

κ	r	β	γL	$\ P\ $	$\lambda_{\min}(G)$	M
400	0.988	0.778544	1.0176	1.44	$5.6 \cdot 10^{-4}$	78
10^3	0.994024	0.862348	0.791643	1.57	$2.3 \cdot 10^{-4}$	127
10^4	0.999264	0.958558	0.299817	3.51	$2.4 \cdot 10^{-5}$	559
10^6	0.999992	0.995984	0.031807	31.6	$2.5 \cdot 10^{-7}$	$1.6 \cdot 10^4$

The Lyapunov matrix P in (15) has the entries

$$\begin{aligned} P_{11} &= \frac{25600s^5 - 1920s^4 - 1632s^3 - 100s^2 - 12s - 1}{64s(16s^2 - 2s - 1)r^2}, \\ P_{12} &= -\frac{26112s^5 - 1984s^4 - 1632s^3 - 92s^2 - 12s - 1}{64s(16s^2 - 2s - 1)r}, \\ P_{13} &= \frac{1984s^3 + 88s^2 + 9s + 2}{8r}, \\ P_{22} &= \frac{1664s^3 + 80s^2 + 10s + 1}{64s}, \\ P_{23} &= -\frac{1984s^3 + 88s^2 + 9s + 2}{8}, \\ P_{33} &= s(3210s^3 + 441s^2 + 14s + 2). \end{aligned}$$

The first three entries solve $Fv_0 = 0$ with $v_0 = (1, 1/r, 0, 0)$. In terms of the free entries, with $C = 8 - 64s$ and $\sigma = 4$ the coefficient of s in $\beta = r(r - \sigma s)$, they are

$$\begin{aligned} P_{11} &= \frac{2CP_{22}s - 2Cs^3 + 2\sigma P_{22} - \sigma s^2 + 2s}{2(Cs + \sigma)(Cs^2 - 1)^2}, \\ P_{12} &= \frac{2CP_{22}s - Cs^3 + 2\sigma P_{22} - \sigma s^2 + s}{2(Cs + \sigma)(Cs^2 - 1)}, \\ P_{13} &= \frac{P_{23}}{Cs^2 - 1}. \end{aligned}$$

To leading order, $F = w_0 w_0^\top / 8 + O(s)$ with $w_0 = (1, -1, 2, -2)$. The leading part of F has rank one, and the positivity of F is decided at orders s and s^3 . The basis U of §6.4 makes this visible

in the minors. In the variables $e = x_t - x_{t-1}$, x_{t-1} and $p = p_{t-1}$, the quadratic part of V_t has the coefficients

$$\begin{aligned} \|e\|^2 &: \frac{1}{64s} + \frac{5}{32} + O(s), & 2\langle e, x_{t-1} \rangle &: \frac{s}{4} + O(s^3), & \|x_{t-1}\|^2 &: \frac{s^2}{2} + O(s^3), \\ 2\langle e, p \rangle &: \frac{1}{4} + O(s), & 2\langle x_{t-1}, p \rangle &: 2s^2 + O(s^3), & \|p\|^2 &: 2s + O(s^2). \end{aligned}$$

Together with $\varphi(x_{t-1})$ this gives (17).

B Polynomials of the Window Lemma

With $\nu = \pm y/2$, $X = y^2$ and $\eta = \vartheta y^2$, the corner polynomials of Lemma 4 are, sorted by powers of ϑ ,

$$F_1^\pm = 4y^2(y^2 \pm 7y/2 + 2) + 2\vartheta(8 - 5y^2 \pm y^3 - 6y^4 \mp 8y^5 - 2y^6) + 2\vartheta^2 y^2(-3 \mp y + 2y^2 \pm y^3),$$

$$F_2^\pm = (8 \pm 8y - 12y^2 \mp 8y^3) + \vartheta(6 - 9y^2 \mp 8y^3 + 11y^4 \pm 8y^5) - \frac{\vartheta^2 y^2}{2}(6 + 3y^2 - y^4).$$

The second route proves $F_1^\pm > 0$ and $F_2^\pm > 0$ on $[0, 1] \times [0, \sqrt{2/5}]$ by interval evaluation of the monomial expansion and bisection. Since F_1 vanishes at $\vartheta = y = 0$, the script writes $F_1 = F_1(0, y) + \vartheta F_{1a}$ and shows that $F_1(0, y)/y^2$ and F_{1a} are positive. This takes 9 boxes for $\nu > 0$ and 67 for $\nu < 0$.

C Exact Certificates at Fixed Condition Numbers

Each row of Table 4 is a rational tuning with $r = 1 - c/\kappa$ a root of its polynomial for curvature q , together with the free entries (P_{22}, P_{23}, P_{33}) of a certificate of the form of §6. The entries P_{11} , P_{12} and P_{13} follow from $Fv_0 = 0$. The script code/fixed_kappa_certificates.py checks in exact rational arithmetic that r is the dominant root on quadratics and that the leading principal minors of F_3 and G are positive, and hence that the rate of the tuning on $\mathcal{F}_{\mu,L}(\mathbb{R}^d)$ is r in every dimension.

Table 4: Exact certificates at fixed κ (results/optimal/fixed_kappa.json).

κ	c	β	γL	P_{22}	P_{23}	P_{33}
100	$\frac{629}{125}$	$\frac{203195907}{312500000}$	$\frac{4958407}{3125000}$	$\frac{368289}{10^6}$	$-\frac{82409}{250000}$	$\frac{40071}{250000}$
400	$\frac{6173}{1000}$	$\frac{101480295257}{125 \cdot 10^9}$	$\frac{133279034566039}{123070937500000}$	$\frac{521519}{10^6}$	$-\frac{146987}{500000}$	$\frac{4403}{50000}$
1000	$\frac{3367}{500}$	$\frac{878213437313}{10^{12}}$	$\frac{387381978567129}{496633000000000}$	$\frac{351169}{500000}$	$-\frac{139493}{500000}$	$\frac{14499}{250000}$
10^4	$\frac{1887}{250}$	$\frac{960020029523}{10^{12}}$	$\frac{74017896690099}{249811300000000}$	$\frac{349997}{200000}$	$-\frac{51973}{200000}$	$\frac{393}{20000}$

D Results in One Dimension

This appendix collects results on $\mathcal{F}_{\mu,L}(\mathbb{R})$. Whether heavy-ball accelerates on functions of one variable, the first of the two riddles of Goujaud et al. [12] (their Open Problem 1), was settled negatively before this work. It was first answered by Fotopoulos et al. [7], in a blog post of July 7, 2026, with a Lean 4 formalization of the core construction that rests on two stated external inputs. They rotate the roots-of-unity cycles of Goujaud et al. [11] by a suitable phase and project them onto a line. The blog reports, by personal communication, that Dieuleveut and collaborators had also solved the problem. We found no such preprint as of October 15, 2026. Combined with Goujaud et al. [11, Thm. 3.6], this gives $\rho \geq 1 - (100/3 + o(1))/\kappa$ on $\mathcal{F}_{\mu,L}(\mathbb{R})$ for every tuning. Building on that construction, Liu et al. [16, App. F, Thm. 8] gave a complete proof with a fixed phase. It shows that every tuning with $\gamma > 0$ and $0 \leq \beta < 1$ has rate at least $(\kappa - 12)/(\kappa + 12)$ for $\kappa \geq 24$, on C^∞ functions with any Hessian-Hölder bound and with locally attracting cycles. Pasechniuk-Vilensky [22] credit the blog with settling the riddle. We include only what we believe adds to these works, namely a better constant (Theorem C), the limit of the cycle method (Proposition 2), and an exact rest-start orbit of Polyak's tuning for all $\kappa \geq 100$ (Theorem D).

D.1 Projected rotations

In one dimension, interpolation is a condition on consecutive secants. For $x_1 < \dots < x_n$, some $f \in \mathcal{F}_{\mu,L}(\mathbb{R})$ has $f'(x_i) = g_i$ for all i if and only if

$$\mu \leq \frac{g_{i+1} - g_i}{x_{i+1} - x_i} \leq L \quad \text{for } i < n.$$

Necessity is clear, and for sufficiency f' can be taken piecewise linear through the points (x_i, g_i) , with slope μ outside $[x_1, x_n]$. No function values need to be prescribed, since heavy-ball uses only gradients, and coinciding points need equal gradients. We use the following statement, due in this form to Fotopoulos et al. [7] for the interval of phases and to Liu et al. [16, App. F] for the fixed phase. We include a short proof for completeness.

Fix $K \geq 3$, $\beta < 1$ and $\gamma > 0$, take $L = 1$ so that $h = \gamma$ and $\gamma\mu = qh$, and let $\omega = 2\pi/K$ and $u = 1 - \cos \omega$. Define

$$\alpha = (1 + \beta)u, \quad E = (1 - \beta)\sin \omega > 0, \quad \Phi_K = (\alpha - qh)(h - \alpha) - E^2 - E(1 - q)h \cot \frac{\pi}{K}. \quad (18)$$

Substituting $E \cot(\pi/K) = (1 - \beta)(2 - u)$ and $E^2 = (1 - \beta)^2 u(2 - u)$ shows that Φ_K is the polynomial $\Phi(h, u)$ of (8) at $r = 1$, and α is the α of (7) at $r = 1$. For a phase ψ , let $x_t = \cos(\omega t + \psi)$, with indices mod K , and $\gamma g_t = (1 + \beta)x_t - \beta x_{t-1} - x_{t+1}$. These are the only gradients for which heavy-ball runs through the points x_t periodically.

Lemma 6 (projected rotations). *For $\psi \in (0, \pi/K)$ the points x_t are distinct, and their sorted consecutive secant slopes, multiplied by γ , take the two values $\sigma_- = \alpha - E \cot \psi$ and $\sigma_+ = \alpha + E \cot(\pi/K - \psi)$. Hence the x_t form a heavy-ball K -cycle on some $f \in \mathcal{F}_{\mu,L}(\mathbb{R})$ if and only if $qh \leq \sigma_-$ and $\sigma_+ \leq h$, and such a phase exists if and only if $\Phi_K \geq 0$.*

Proof. By the addition formulas, $\gamma g_t = \alpha x_t + E y_t$ with $y_t = \sin(\omega t + \psi)$. Sorting by $x = \cos \theta$ means sorting by the folded angle $|\theta| \in (0, \pi)$. The folded angles of the points alternate between the two sets $\psi + \omega\mathbb{Z}$ and $-\psi + \omega\mathbb{Z}$, with gaps alternating between 2ψ and $\omega - 2\psi$. For an upper point $(\cos A, \sin A)$ and a lower point $(\cos B, -\sin B)$ the scaled secant is $\alpha - E \cot((A - B)/2)$, which gives σ_- for $A - B = 2\psi$ and σ_+ for $A - B = -(\omega - 2\psi)$. With $a_1 = (\alpha - qh)/E$ and

$a_2 = (h - \alpha)/E$, the two conditions read $\psi \geq \operatorname{arccot} a_1$ and $\pi/K - \psi \geq \operatorname{arccot} a_2$. They are compatible if and only if $\operatorname{arccot} a_1 + \operatorname{arccot} a_2 \leq \pi/K$, with arccot taking values in $(0, \pi)$. Since $a_1 + a_2 = (1 - q)h/E > 0$, if $a_1 \leq 0$ or $a_2 \leq 0$ then $a_1 a_2 \leq 0$, so both this condition and $\Phi_K \geq 0$ fail. For $a_1, a_2 > 0$ the addition formula for the cotangent turns the condition into $E^2(a_1 a_2 - 1) - E^2(a_1 + a_2) \cot(\pi/K) = \Phi_K \geq 0$. $\square$

The proof applies to every phase. Shifting the time shifts the phase by ω , and the two slopes depend on the phase only through twice the phase modulo ω . So a phase $\psi' \in (-\pi/K, 0)$ gives the same slopes as the phase $\psi' + \pi/K \in (0, \pi/K)$, and a multiple of π/K makes two points with different gradients coincide. Replacing ψ by $\psi - \pi/K$ and reflecting $x \mapsto -x$, which adds π to the phase, therefore keeps both slopes. The cycle then visits its points in the order

$$(x_0, x_1, x_2, \dots, x_{K-2}, x_{K-1}) = (x^{(0)}, x^{(1)}, x^{(3)}, x^{(5)}, \dots, x^{(4)}, x^{(2)}),$$

where $x^{(0)} < x^{(1)} < \dots < x^{(K-1)}$ are the sorted points. This is the ordering conjectured by Goujaud et al. [12], and we checked it for $K = 3, \dots, 39$. If $\Phi_K > 0$ and $r(\mu) < 1$, flattening f' to slope μ near each point of the cycle makes the cycle attracting, and mollifying makes f smooth. This is the one-dimensional version of Goujaud et al. [11, §4–5] and Liu et al. [16, Thm. 8].

D.2 The constant 50/3

Lemma 7 (explicit cycle region). *Let $0 < q < 6/25$, $0 \leq \beta < 1$, $\delta = 1 - \beta$, $0 < h \leq 4$ and $h(6 - 25q) \geq 50\delta$. Then $\Phi_K \geq 0$ for some $K \geq 4$.*

Proof. Write $u_K = 1 - \cos(2\pi/K)$. With $\beta = 1 - \delta$, (18) becomes

$$\Phi_K = -\chi(u_K) + 2qh(\delta + u_K(1 - \delta)) + 2u_K\delta(2u_K - \delta), \quad (19)$$

where $\chi(u) = 4u^2 - 2hu + 2h\delta + qh^2$. The middle term is non-negative, and so is the last one when $u_K \geq \delta/2$. It therefore suffices to find a $K \geq 4$ with $\chi(u_K) \leq 0$ and $u_K \geq \delta/2$.

The quadratic χ is non-positive on $[u_-, u_+]$, where $u_- + u_+ = h/2$ and $u_- u_+ = (2h\delta + qh^2)/4$. Since $(u_- + u_+)^2/(u_- u_+) = h/(2\delta + qh)$ and $R + 2 + 1/R = 25/6$ at $R = 3/2$, we get $u_+/u_- \geq 3/2$ if and only if $h(6 - 25q) \geq 50\delta$. The window lies at the right place, because $u_- \leq h/4 \leq 1 = u_4$ and $u_- = u_- u_+/u_+ > (2h\delta + qh^2)/(2h) > \delta$. The values u_K decrease to 0 with $u_K/u_{K+1} \leq 3/2$ for $K \geq 4$. Indeed, $u_4/u_5 = 1.447\dots$, and for $K \geq 5$, $u_K/u_{K+1} \leq ((K+1)/K)^2 \leq 1.44$ because $\sin x/x$ decreases. Let $K^* = \max\{K \geq 4 : u_K \geq u_-\}$. Then $u_- \leq u_{K^*} \leq 1.5 u_{K^*+1} < 1.5 u_- \leq u_+$ and $u_{K^*} > \delta/2$. $\square$

Lemma 8 (quadratic rate). *For $\beta \in [0, 1)$ and $\gamma > 0$, with $\bar{s} = 1 - \sqrt{\beta}$, we have $1 - r(\mu) \leq \min\{\bar{s}, \gamma\mu/\bar{s}\}$.*

Proof. Let $p(z) = z^2 - (1 + \beta - \gamma\mu)z + \beta$, so that $p(1) = \gamma\mu$. If the roots are complex, then $r = \sqrt{\beta}$, so $1 - r = \bar{s}$, and $|1 + \beta - \gamma\mu| < 2\sqrt{\beta}$ gives $\gamma\mu > \bar{s}^2$. If they are real, $r_1 \geq r_2$, then $r_1 r_2 = \beta \geq 0$. If $r(\mu) \geq 1$ there is nothing to prove. Otherwise $r \geq \sqrt{\beta}$, so $1 - r \leq \bar{s}$. If $0 \leq r_2 \leq r_1$, then $r_2 \leq \sqrt{\beta}$ and $1 - r_1 = \gamma\mu/(1 - r_2) \leq \gamma\mu/\bar{s}$. If $r_2 \leq r_1 \leq 0$, then $\gamma\mu = (1 - r_1)(1 - r_2) \geq 1 \geq 1 - r$. $\square$

Theorem C (non-acceleration on $\mathcal{F}_{\mu,L}(\mathbb{R})$ with constant $50/3$). *For every $0 < \mu < L$ and every $(\gamma, \beta) \in \mathbb{R}^2$,*

$$\rho_{\gamma,\beta}(\mathcal{F}_{\mu,L}(\mathbb{R})) \geq \frac{1-10q}{1+10q}, \quad \text{and, if } q \leq 1/25, \quad \rho_{\gamma,\beta}(\mathcal{F}_{\mu,L}(\mathbb{R})) \geq 1 - \frac{100q}{6+25q}.$$

So $c_1(\kappa) \leq 100/(6+25/\kappa) \rightarrow 50/3$. Moreover $\rho_{\gamma,\beta} \geq 1$ if $\beta \in [0, 1)$, $q < 6/25$ and $h(6-25q) \geq 50(1-\beta)$.

Proof. We show that $1-\rho \leq 20q/(1+10q)$, and $1-\rho \leq 100q/(6+25q)$ when $q \leq 1/25$. The two bounds agree at $q = 1/25$, and the first is trivial for $q \geq 1/10$. A non-constant periodic orbit forces $\rho \geq 1$, and $\rho \geq \rho_Q$ always. We go through the tunings case by case.

If the tuning is unstable on quadratics, that is, $|\beta| \geq 1$, $h \leq 0$ or $h \geq 2(1+\beta)$, then $\rho \geq \rho_Q \geq 1$. If $-1 < \beta < 0$ and $0 < h < 2(1+\beta)$, the roots of p satisfy $r_1 > 0 > r_2$, so $1-r_1 = \gamma\mu/(1-r_2) < qh < 2q$, and both target bounds exceed $2q$.

Now let $0 \leq \beta < 1$, $\delta = 1-\beta$ and $0 < h < 2(1+\beta)$, so that in particular $h < 4$, and first assume $q \leq 1/25$. If $h(6-25q) \geq 50\delta$, then Lemmas 6 and 7 give a cycle, so $\rho \geq 1$. Otherwise $qh < 50q\delta/(6-25q)$ with $\delta = \bar{s}(2-\bar{s})$ and $\bar{s} = 1-\sqrt{\beta}$, and Lemma 8 gives $1-\rho \leq \min\{\bar{s}, 50q(2-\bar{s})/(6-25q)\}$. This is at most the crossing value $\bar{s}^* = 100q/(6+25q)$, which is at most $20q/(1+10q)$ if and only if $q \leq 1/25$. If instead $1/25 < q < 1/10$, Lemma 8 with $h < 4$ gives $1-\rho \leq \min\{\bar{s}, 4q/\bar{s}\} \leq 2\sqrt{q}$, and $2\sqrt{q} \leq 20q/(1+10q)$ if and only if $10q - 10\sqrt{q} + 1 \leq 0$, which holds for $(4-\sqrt{15})/10 \leq q \leq (4+\sqrt{15})/10$, an interval that contains $(1/25, 1/10)$.

The last claim of the theorem follows from Lemmas 6 and 7 when $h \leq 4$. When $h > 4$, we have $h > 2(1+\beta)$, and the tuning is unstable on quadratics. $\square$

The bound $c_1 \leq 50/3 + o(1)$ improves on 24 [16] and $100/3$ [7, 11]. Together with Theorem B, $8 - 64/\sqrt{\kappa} \leq c_1(\kappa) \leq 50/3 + o(1)$. In the notation of Remark 1, $50/3$ corresponds to $C = 25/3$, and should not be confused with the threshold $C > 50/3$ of Goujaud et al. [11].

D.3 The limit of the cycle method

We call a cycle on the K -th roots of unity *single-jump* if it advances by one root per step, as the planar cycles of Goujaud et al. [11] and their projections in Lemma 6 do. Fotopoulos et al. [7] describe cycles that jump by several roots per step. Single-jump cycles combined with the quadratic bound cannot prove anything below 16.

Proposition 2 (the cycle method stops at 16).

- (a) If $0 < \delta < 1$, $h > 0$ and $h(1+q)^2 < 8\delta(1-\delta)(1-q)$, then $\Phi_K < 0$ for every $K \geq 3$.
- (b) At $\delta = 32q$ and $h = 8\delta(1-\delta)(1-q)^2/(1+q)^2$, which satisfies (a), we have $(1-\rho_Q)/q \rightarrow 16$ as $q \rightarrow 0$.

So no combination of single-jump roots-of-unity cycles, with any K and any phase, with the quadratic bound proves $\rho \geq 1 - c/\kappa$ with $c < 16$ for all small q . This applies to the bounds of Goujaud et al. [11], Fotopoulos et al. [7], Liu et al. [16] and Theorem C.

Proof. (a) Dropping the terms $-qh^2$, $-2qh\delta u_K$ and $-2u_K\delta^2$ from (19) leaves $\Phi_K \leq -4(1-\delta)u_K^2 + 2h(1+q)u_K - 2h\delta(1-q)$, whose maximum over u_K is

$$\frac{h}{4(1-\delta)}[h(1+q)^2 - 8\delta(1-\delta)(1-q)] < 0.$$

(b) Write $h = 8\delta c_q$ with $c_q = (1-\delta)(1-q)^2/(1+q)^2 \rightarrow 1$. At curvature μ , the substitution $z = 1 - v$ turns the characteristic polynomial into $v^2 - (\delta + qh)v + qh$, so

$$1 - r(\mu) = \frac{16q c_q}{1 + 8qc_q + \sqrt{(1 + 8qc_q)^2 - c_q}} \rightarrow 16q.$$

At curvature L , for $q < 0.028$ the roots are complex, so $1 - r(L) = 1 - \sqrt{1 - 32q} \rightarrow 16q$. $\square$

Part (a) is consistent with the description of the union of the cycle regions over real K in Goujaud [9, eqs. (10.22)–(10.23)], and Pasechniuk-Vilensky [22, §7.4] show that the union of these regions is the complement of their region Ω_1 , on which they prove convergence (their Thm. 7.1(i)). The proof of (a) bounds Φ_K for every real $u_K \geq 0$, and hence for every real $K > 2$. In the description of Pasechniuk-Vilensky [22, §7.4 and Thm. 7.1(ii)], every tuning with a cycle that jumps by several roots of unity has $\Phi_K \geq 0$ for some real $K > 2$, so these cycles do not reach the tunings of (a) either. As far as we know the value 16 is new, though simple. In dimension two and higher, Theorem A shows that the true constant, 8, is below this limit, which is why its proof uses decaying spirals instead of cycles.

D.4 Polyak's tuning from rest

Several works start heavy-ball at rest, $x_{-1} = x_0$: Pasechniuk-Vilensky [22] throughout, and Lessard et al. [15, App. B] and Liu et al. [16, App. H] in their non-convergence examples. Heavy-ball is invertible for $\beta \neq 0$, so an orbit started at rest can land exactly on a cycle only if the cycle contains a rest state, that is, $x_{K-1} = x_0$.

Theorem D (Polyak's tuning fails from rest in one dimension). *Let $\kappa \geq 100$, $L = 1$, $s = \kappa^{-1/2}$, $\gamma = 4/(1+s)^2$, $\beta = ((1-s)/(1+s))^2$ and $\delta = 1 - \beta$. Let f' be the piecewise-linear function through the points*

$$\left(-1, \frac{-(1+\beta) + \delta/\beta}{\gamma}\right), \quad \left(-\frac{\delta}{\beta}, \frac{-\delta(1+2\beta)}{\beta\gamma}\right), \quad \left(0, \frac{\delta}{\gamma}\right), \quad \left(1, \frac{1}{\gamma}\right),$$

extended with slope μ . Then $f \in \mathcal{F}_{\mu,L}(\mathbb{R})$, and heavy-ball started at $x_{-1} = x_0 = 1$ produces the sequence $1, 0, -1, -\delta/\beta, 1, 1, 0, -1, \dots$ of period five. In particular it never converges.

Proof. The points $(1, 0, -1, -\delta/\beta, 1)$ form a five-cycle with a repeated point, and the two gradients at the repeated point agree because $\beta(1 + \delta/\beta) = 1 = x_0 - x_1$. The order $-1 < -\delta/\beta < 0 < 1$ holds because $1 - 6s + s^2 > 0$. The six slope conditions $\gamma\mu \leq \gamma \cdot \text{secant} \leq \gamma L$ reduce to the positivity on $(0, 1/10]$ of the rational functions of s in Table 5. The binding one is $1 - 10s + 2s^3 - s^4$, which is decreasing and equals 0.0019 at $s = 1/10$. One-dimensional interpolation completes the proof. $\square$

Earlier results of this kind hold at a single κ or in higher dimension. Lessard et al. [15, §4.6 and App. B] show at $\kappa = 25$, by a finite verification, that the rest start $x_0 = 3.3$ enters the basin of

Table 5: The slope conditions of Theorem D, with σ the scaled secant γ -secant and $s = \kappa^{-1/2}$. All entries are positive on $(0, 1/10]$.

interval and σ	$\sigma - \gamma\mu$	$\gamma L - \sigma$
$[0, 1], \sigma = \beta$	$\frac{1 - 3s}{1 + s}$	$\frac{3 - s}{1 + s}$
$[-\delta/\beta, 0], \sigma = 1 + 3\beta$	$\frac{4(1 - s)}{(1 + s)^2}$	$\frac{4s(1 - s)}{(1 + s)^2}$
$[-1, -\delta/\beta], \sigma = \frac{(1 + \beta)(\beta - 2\delta)}{\beta - \delta}$	$\frac{2(1 - 10s + 2s^3 - s^4)}{(1 + s)^2(1 - 6s + s^2)}$	$\frac{2(1 - 2s + 10s^3 - s^4)}{(1 + s)^2(1 - 6s + s^2)}$
order of the points	$1 - \delta/\beta = \frac{1 - 6s + s^2}{(1 - s)^2}$	$\delta/\beta = \frac{4s}{(1 - s)^2}$

an attracting three-cycle. Liu et al. [16, App. H, Prop. 10] show at $\kappa = 100$ that every rest start in $[37/50, 19/25]$ converges to an attracting three-cycle. Pasechniuk-Vilensky [22] prove failure from rest for every $\kappa \geq 9 + 4\sqrt{5}$, in $\mathbb{R}^3$ (their Thm. 2.2). Theorem D holds for all $\kappa \geq 100$ in one dimension, with exact landing on the orbit. Whether Polyak's tuning fails from rest in one dimension for every $\kappa \geq 9 + 4\sqrt{5}$ remains open. With a free middle point in place of -1 , the same shape appears to be feasible for κ above about 45.76. This threshold is numerical, and exact rational instances at $\kappa = 49$ and 64 are verified.

E Reproducibility

All scripts are at <https://github.com/marcoshernanz/heavy-ball-optimal-rate>. They run with SymPy 1.14, mpmath 1.3, NumPy and Matplotlib, and CVXPY 1.9.3 with Clarabel is needed only for the optional numerical searches. Table 6 lists what each script checks or produces.

Table 6: Scripts in code/ and what they check or produce.

Script	Checks or produces
verify_lowerbound.py	all checks for Theorem A and Corollary 2 in Table 1 and the floating-point tests, writes results/optimal/lowerbound.json and ends with ALL OK
verify_rou_analytic.py	the constants of the proof of Lemma 5
verify_achievability.py 1/20	Proposition 1 for $\kappa \geq 400$, writes results/optimal/achievability.json; with --range it checks the range of §6.5, and with --table it computes Table 3
verify_pv_membership.py	the tuning (4) lies in the region Q of Pasechniuk-Vilensky [22] (Remark 2)
fixed_kappa_certificates.py	Appendix C, writes results/optimal/fixed_kappa.json
ghadimi_corner.py	§9.3
verify_symbolic.py, stallcycles.py	the identities of Appendix D, and the rest-start orbit of Theorem D in exact rationals
pv_region_search.py, decaying_spirals.py, scan_spiral_constant.py	the numerical columns of Table 2 (best on Q, root-of-unity spirals, log-spirals)
make_figures_optimal.py paper	Figure 1