An $\Omega(T^{-3/2})$ lower bound for heavy ball

GPT-6 Astra prompted by Wenzhi Gao

September 9, 2026

Abstract

For every predetermined heavy-ball schedule with nonnegative steps and momenta in $[0, 1)$, we prove a worst-case last-iterate error $\Omega(LR^2/T^{3/2})$ on convex L-smooth functions, with zero initial velocity and initial distance at most R. The hard instance is a static Huber chain of dimension at most $T + 1$. Two improvements remove the logarithms from the preceding analysis: a sharp-order local bound $\mathcal{S}_2(N) = O(\sqrt{N})$, and compression of each retention block into one virtual checkpoint. The compression preserves a transfer coefficient small enough for a 2/3-power capacity argument. The proof uses the static-chain construction and momentum-removal lemma of Ma and Zhang [1]; the local estimate, block compression, and their combination are proved here. No matching upper bound is claimed.[1]

1 Main result

Theorem 1. *There is an absolute constant $c > 0$ such that, for every integer $T \geq 1$, $L, R > 0$, and every finite predetermined schedule $\eta_t \geq 0$, $\beta_t \in [0, 1)$, $0 \leq t < T$, there are a dimension $d \leq T + 1$, a differentiable convex L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with a minimizer $x_\star$, and an initialization $\|x_0 - x_\star\| \leq R$ for which*

$$x_{-1} = x_0, \qquad x_{t+1} = x_t - \eta_t \nabla f(x_t) + \beta_t(x_t - x_{t-1})$$

satisfies

$$f(x_T) - f(x_\star) \geq \frac{cLR^2}{T^{3/2}}. \tag{1}$$

The objective can be a scaled and translated static one-sided Huber chain of the form in [1, Section 3].

Schedules may depend on T, L, R, but are fixed before the adversarial objective is chosen. The output is the actual last iterate. The theorem strengthens the bound $\Omega(LR^2/[T^{(1+\sqrt{5})/2} \log T])$ of [1] and removes all logarithmic losses from the preceding three-halves refinement. It does not establish an upper rate or cover schedules chosen adaptively from observed gradients.

The proof normalizes $L = R = 1$. The local estimate below gives linear control in an arbitrary product budget. That control lets us replace all checkpoints within a retention block by a single candidate, with a charge proportional to $d\sqrt{n}$, where d is the block's largest memory and n its length. A weighted selection argument then uses $\sum d^{2/3} n^{1/3} = O(T)$ to prove (1). There is no small-certificate split or global logarithmic product-budget argument in this proof.

[1] A recent result [2] shows an improved upper bound that matches our lower bound using randomized stepsize and momentum.

2 A sharp square-root bound for admissible sequences

For a nonnegative sequence $\eta_1, \dots, \eta_N$, write

$$S_{a,b} = \sum_{i=a+1}^{b} \eta_i, \qquad A_{a,b} = \sum_{i=a+1}^{b} (b-i)\eta_i, \qquad \psi(a,b) = \min\left\{ \frac{S_{a,b}}{4}, \frac{(b-a)S_{a,b}}{2(2+A_{a,b})} \right\}.$$

Here $0 \le a < b \le N$ are integers. For $\Lambda \ge 1$, the sequence is Λ-*admissible* if every chronological family satisfies

$$\prod_{j=1}^{k} \psi(a_j, b_j) \le \Lambda, \qquad 0 \le a_1 < b_1 \le \cdots \le a_k < b_k \le N.$$

The empty product is one. Denote the supremum of the total sequence mass by $\mathcal{S}_\Lambda(N)$.

Theorem 2. *For every integer $N \ge 1$ and every $\Lambda \ge 1$,*

$$\mathcal{S}_2(N) \le 8192\sqrt{N}, \qquad \mathcal{S}_\Lambda(N) \le 8200\Lambda\sqrt{N}. \tag{2}$$

Moreover, $\mathcal{S}_2(N) \ge \sqrt{N}$, so its exponent is sharp.

We first prove a square-root estimate for the fixed small product budget e^{δ_0}, where $\delta_0 = 10^{-4}$. A constant number of prefix-product bands then handles budget two. All logarithms in this section are natural unless a base is displayed.

2.1 From integer intervals to inverse cumulative mass

Let Π_i be the maximum chronological product supported in the first i entries. The empty product and concatenation give

$$1 = \Pi_0 \le \cdots \le \Pi_N \le e^{\delta_0} < 2, \qquad \psi(a,b) \le \frac{\Pi_b}{\Pi_a}. \tag{3}$$

In particular, $\eta_i \le 8$. For an integer interval with $S_{a,b} > 8$, put $r = \Pi_b/\Pi_a \le 2$. The first branch of ψ exceeds r, so $(b-a)S_{a,b} \le 4r + 2rA_{a,b}$. Hence

$$K_{a,b} := (b-a)S_{a,b} - 2A_{a,b} \le 4 + (1-r^{-1})(b-a)S_{a,b} \le 4 + (b-a)S_{a,b}\log r. \tag{4}$$

Set $S = \sum_i \eta_i$, and linearly interpolate the cumulative mass: $F(i) = \sum_{j=1}^{i} \eta_j$. Then $F : [0,N] \to [0,S]$ is nondecreasing and 8-Lipschitz. For real endpoints define

$$K(a,b) = (b-a)(F(a)+F(b)) - 2\int_a^b F(t)\,dt + F(b) - F(a).$$

At integer endpoints this is exactly $K_{a,b}$. On each unit rectangle of endpoint positions it is affine separately in a and b: for example, $\partial_a K = F(a) - F(b) + (b-a-1)F'(a)$ has derivative zero inside a cell. Thus $K(a,b)$ is the convex combination of its four integer-corner values.

Place the logarithmic prefix increments on the mass axis:

$$\nu = \sum_{i=1}^{N} (\log \Pi_i - \log \Pi_{i-1})\delta_{F(i)}, \qquad \nu(\mathbb{R}) \le \delta_0.$$

For $0 \le s < S$, choose the right-continuous inverse $Q(s) = \sup\{t \in [0,N] : F(t) \le s\}$, and set $Q(S) = N$. Then $F(Q(s)) = s$. Extend Q by zero to the left of 0 and by N to the right of S. Its Stieltjes measure $\mu = dQ$ satisfies

$$\mu \ge \tfrac{1}{8}\, ds, \qquad \mu([0,S]) \le N. \tag{5}$$

Indeed, $s_2 - s_1 \le 8(Q(s_2) - Q(s_1))$. Flat pieces of F produce jumps of Q; no positive-step assumption is imposed. We use Stieltjes and distributional identities when necessary. Choices of inverse representatives and moving-endpoint atom conventions affect only null sets in the integrals below.

Fix

$$M = 32, \qquad c_0 = \frac{7}{8}, \qquad C_0 = \frac{9}{4}, \qquad \ell = S - M.$$

If $S \le 2M$, then $S \le 64\sqrt{N}$ already. Assume henceforth that $\ell \ge M$. For $0 \le s \le \ell$, put

$$n(s) = Q(s+M) - Q(s), \qquad d(s) = \nu([s-8, s+M+8]).$$

At the real endpoints $a = Q(s)$ and $b = Q(s+M)$, all four integer corner intervals have mass between $M - 16 = 16$ and $M + 16 = 48$. Their lengths are at most $n(s) + 2$, and all relevant prefix increments have mass locations in $[s-8, s+M+8]$. Consequently, (4) and bilinear interpolation imply

$$K(a,b) \le 4 + (n(s)+2)(M+16)d(s) \le 4 + \frac{9}{4} M n(s) d(s),$$

where $n(s) \ge M/8 = 4$ was used in the last step.

The inverse-area identity is

$$\int_a^b t\, dF(t) = \int_s^{s+M} Q(u)\, du.$$

Integration by parts therefore gives

$$K(a,b) = M\left[1 - Q(s) - Q(s+M) + \frac{2}{M}\int_s^{s+M} Q(u)\, du\right].$$

Thus

$$Q(s) + Q(s+M) - \frac{2}{M}\int_s^{s+M} Q(u)\, du \ge c_0 - C_0 n(s) d(s). \tag{6}$$

In addition, Tonelli's theorem gives

$$0 \le d(s) \le \delta_0, \qquad \int_0^\ell d(s)\, ds \le (M+16)\delta_0 = \frac{3}{2} M \delta_0. \tag{7}$$

2.2 Parabolic smoothing and a one-sided comparison

Define

$$W(s) = \frac{1}{M}\int_0^M u(M-u) Q(s+u)\, du, \qquad V(s) = W'(s) = \frac{1}{M}\int_{[s,s+M]} (t-s)(s+M-t)\, d\mu(t).$$

The compact parabolic kernel is globally Lipschitz, so V is Lipschitz and in particular absolutely continuous. Differentiating in distributions, and hence almost everywhere, yields both

$$V'(s) = Q(s) + Q(s+M) - \frac{2}{M}\int_s^{s+M} Q(u)\,du \geq c_0 - C_0 n(s)d(s), \tag{8}$$

$$V'(s) = \int_{[s,s+M]} \left(\frac{2(t-s)}{M} - 1\right) d\mu(t). \tag{9}$$

The nonnegative kernel and (5) also give

$$V(s) \geq 0, \qquad \int_0^\ell V(s)\,ds = W(\ell) - W(0) \leq \frac{NM^2}{6}. \tag{10}$$

Split the window measure into its first quarter and the rest:

$$m_-(s) = \mu([s, s+M/4]), \qquad m_+(s) = \mu((s+M/4, s+M]).$$

Almost everywhere $n = m_- + m_+$. The signed kernel in (9) is at most $-1/2$ on the first quarter and at most one elsewhere. Combining this with (8) and $d \leq \delta_0$ gives

$$(1/2 - C_0\delta_0)m_- \leq (1 + C_0\delta_0)m_+ - c_0.$$

Since $C_0\delta_0 \leq 1/4$, we obtain $m_- \leq 5m_+$ and $n \leq 6m_+$. For $0 \leq s \leq \ell - M$, another application of Tonelli gives

$$\begin{aligned}
\int_s^{s+M} V(r)\,dr &\geq \int_{(s+M/4,s+M]} \left(\int_s^t \frac{(t-r)(M-t+r)}{M}\,dr\right) d\mu(t) \\
&\geq \left(\int_0^{M/4} \frac{u(M-u)}{M}\,du\right) m_+(s) = \frac{5M^2}{192} m_+(s).
\end{aligned}$$

Consequently,

$$n(s) \leq \frac{1152}{5M^2}\int_s^{s+M} V(r)\,dr \qquad (0 \leq s \leq \ell - M). \tag{11}$$

The direction of this comparison matters: it uses only future values of V. The small product budget prevents most of the window's time span from concentrating at its departing edge.

2.3 A positive backward test function

Put $b_0 = 2592/(5M^2)$ and define positive integral operators

$$\begin{aligned}
(\mathcal{B}h)(r) &= b_0 \int_{\max(0,r-M)}^{\min(r,\ell-M)} h(s)d(s)\,ds, \\
(\mathcal{K}h)(r) &= \int_r^\ell (\mathcal{B}h)(u)\,du.
\end{aligned}$$

Empty integrals are zero. Let $g(r) = \ell - r$ and use the weighted norm

$$\|h\|_g = \sup_{0 \leq r < \ell} \frac{|h(r)|}{\ell - r}, \qquad h(\ell) = 0.$$

We claim

$$\|\mathcal{K}\|_g \le \varkappa := \frac{7776}{5}\delta_0 = 0.15552 < \frac{1}{3}. \tag{12}$$

Indeed, Tonelli gives

$$(\mathcal{K}g)(r) = b_0 \int_0^{\ell-M} d(s)(\ell-s)\,|[r,\ell]\cap[s,s+M]|\,ds.$$

A nonempty intersection requires $s \ge r-M$, so $\ell-s \le \ell-r+M$. If $\ell-r \ge M$, the product of this quantity and the intersection length is at most $2M(\ell-r)$. If $\ell-r<M$, the intersection length is at most $\ell-r$ and $\ell-s\le 2M$, giving the same bound. Thus

$$(\mathcal{K}g)(r) \le 2b_0M(\ell-r)\int_0^\ell d(s)\,ds \le 3b_0M^2\delta_0 g(r) = \varkappa g(r).$$

Positivity extends the estimate to signed h and proves (12).

The convergent Neumann series defines

$$\tau = (I+\mathcal{K})^{-1}g = \sum_{j=0}^{\infty}(-\mathcal{K})^j g.$$

Although the series alternates, its proximity to g ensures positivity:

$$\|\tau-g\|_g \le \frac{\varkappa}{1-\varkappa} < \frac{1}{2}, \qquad \frac{1}{2}(\ell-r)\le\tau(r)\le 2(\ell-r).$$

Since $\mathcal{B}$ consists of sliding integrals of an integrable function, it maps into continuous functions. Hence $\mathcal{K}$ maps into continuously differentiable functions. Differentiating $\tau = g-\mathcal{K}\tau$ yields

$$\tau(\ell)=0, \qquad \tau'(r) = (\mathcal{B}\tau)(r)-1. \tag{13}$$

Multiply (8) by $\tau\ge 0$ and integrate by parts:

$$\begin{aligned} c_0\int_0^\ell \tau &\le \int_0^\ell \tau V' + C_0\int_0^\ell \tau n d \\ &= -\tau(0)V(0) + \int_0^\ell V - \int_0^\ell(\mathcal{B}\tau)V + C_0\int_0^\ell \tau n d \\ &\le \int_0^\ell V - \int_0^\ell(\mathcal{B}\tau)V + C_0\int_0^\ell \tau n d. \end{aligned} \tag{14}$$

By (11) and the definition of $\mathcal{B}$,

$$C_0\int_0^{\ell-M}\tau(s)n(s)d(s)\,ds \le \int_0^\ell(\mathcal{B}\tau)(r)V(r)\,dr.$$

Thus all interior damping in (14) cancels. On the last strip, $\tau\le 2M$ and $d\le\delta_0$. Moreover,

$$\int_{\ell-M}^{\ell} n(s)\,ds \le M\mu([0,S]) \le MN.$$

Its remaining charge is therefore at most $2C_0M^2\delta_0N$. Since $\int_0^\ell\tau\ge\ell^2/4$, we conclude that

$$\frac{c_0}{4}\ell^2 \le M^2N\left(\frac{1}{6}+2C_0\delta_0\right).$$

The square-root coefficient on the right is less than 28. Therefore $S=M+\ell\le 32+28\sqrt{N}\le 64\sqrt{N}$. The previously separated case $S\le 2M$ satisfies the same bound. We have proved

$$\mathcal{S}_{e^{10^{-4}}}(N) \le 64\sqrt{N}. \tag{15}$$

2.4 Product bands and arbitrary budgets

For a 2-admissible sequence, remove entry i whenever $\lfloor \log \Pi_i/\delta_0 \rfloor$ increases. At most

$$J = \left\lfloor \frac{\log 2}{\delta_0} \right\rfloor \leq 6931$$

entries are removed. Between two removals, the endpoint prefix ratio is less than e^{δ_0}. Appending any chronological family in that run to a maximizing preceding prefix shows that the entire run is e^{δ_0}-admissible. The removed entries have mass at most eight each. Hence (15) and Cauchy–Schwarz give

$$\sum_i \eta_i \leq 64\sqrt{(J+1)N} + 8\min(J, N) \leq \left(64\sqrt{J+1} + 8\sqrt{J}\right)\sqrt{N} < 8192\sqrt{N}.$$

For general $\Lambda \geq 1$, use dyadic prefix bands instead. Write $J = \lfloor \log_2 \Lambda \rfloor$. Removing the at most J band crossings leaves at most $J+1$ 2-admissible runs. For each removed entry let $r_i = \Pi_i/\Pi_{i-1} \geq 1$. Then

$$\eta_i \leq 4r_i, \qquad \prod_{i \text{ removed}} r_i \leq \Lambda, \qquad \sum_{i \text{ removed}} (r_i - 1) \leq \Lambda - 1.$$

The last inequality follows by expanding the product of the nonnegative increments. Therefore the removed mass is at most $4(J + \Lambda - 1) \leq 8\Lambda$. Since $\sqrt{J+1} \leq \Lambda$,

$$\sum_i \eta_i \leq 8192\sqrt{(J+1)N} + 8\Lambda \leq 8200\Lambda\sqrt{N}.$$

This establishes both upper bounds in (2).

Finally, take $\eta_i = \sqrt{i} - \sqrt{i-1}$. Its total mass is $\sqrt{N}$. For an integer interval, put $m = \sqrt{b} - \sqrt{a}$. Summation by parts and monotonicity of the square root give

$$\begin{aligned} K_{a,b} &= (b-a)(\sqrt{a} + \sqrt{b}) - 2\sum_{j=a}^{b-1} \sqrt{j} \\ &\leq (b-a)(\sqrt{a} + \sqrt{b}) - 2\int_a^b \sqrt{t}\,dt + 2m = 2m - \frac{m^3}{3} \leq \frac{4\sqrt{2}}{3} < 4. \end{aligned}$$

Thus the second branch of every score is at most one. The sequence is 1-admissible and hence 2-admissible, proving the matching exponent claimed in Theorem 2.

3 Static responses and tight retention blocks

For update indices $0 \leq t < T$, define

$$m_t = \sum_{u=t}^{T-1} \prod_{v=t+1}^{u} \beta_v, \qquad h_t = m_t \eta_t, \qquad H = \sum_{t=0}^{T-1} h_t.$$

Then $m_{T-1} = 1$, $m_t = 1 + \beta_{t+1} m_{t+1}$, and $1 \leq m_t \leq T - t$. For $s \leq t$, let

$$\rho_{s,t} = \frac{m_t}{m_s} \prod_{v=s+1}^{t} \beta_v.$$

Retention is multiplicative and lies in $[0,1]$. We use the exact responses from [1, Appendix A]:

$$F_t = \sum_{s\le t} h_s\rho_{s,t}, \qquad \Delta_t = \sum_{s<t} \frac{F_s}{m_s} \le \sum_{s<t} h_s,$$
$$S_{u,t} = m_u(1-\rho_{u,t}), \qquad \mathcal{A}_{u,t} = \sum_{u<s<t} h_s(1-\rho_{s,t}),$$
$$W_{u,t} = \sum_{u<s\le t} h_s\rho_{s,t}.$$

The starting response and warm-transfer gain are

$$G_{\text{start}}(t) = \frac{F_t}{2(2+\Delta_t)}, \qquad G(u,t) = \frac{W_{u,t}}{4m_u}\min\left\{1, \frac{2S_{u,t}}{2+\mathcal{A}_{u,t}}\right\}. \tag{16}$$

Here $u<t$ for a warm transfer. In particular $S_{u,t} \ge 1$. The momentum-removal lemma [1, Lemma A.2] states

$$G(u,t) \ge \rho_{u,t}\psi(u,t) \quad \text{if } \rho_{u,t} > 0, \tag{17}$$

where the score uses the original steps on $(u,t]$.

A starting response, chronological warm transfers, and intervening waits are realized together by a single static chain [1, Lemma A.1]. A wait from s to t preserves at least $\rho_{s,t}$ times the saved remaining displacement. We need the following explicit extension concerning the final wait.

Lemma 1 (Deferred terminal penalty)**.** *Suppose a static path has last actual checkpoint v and saved remaining displacement at least A_v. For any $b \ge v$, a terminal penalty can be installed at query b, giving a static witness with error at least*

$$\frac{(\rho_{v,b}A_v)^2}{4(1+\sum_{t>b} h_t)}.$$

This waiting step adds no coordinate.

Proof. Before the terminal penalty is added, the frontier receives only nonnegative velocity increments from its left link. Its position is nondecreasing, and its remaining displacement after update b is at least $\rho_{v,b}A_v$. Set the terminal offset to its position at query b. The penalty then has zero gradient through query b, preserving this velocity. The supporting-line argument in [1, Appendix A.4, step 4], now at time b, gives precisely the displayed bound. The change is only in the terminal offset and threshold. $\square$

Thus endpoints used below are virtual checkpoints: an actual transfer may have ended earlier and its frontier may subsequently wait. We work directly with these static witnesses, rather than assuming the paper's displayed certificate family already includes the extension.

Lemma 2 (Tight retention partition)**.** *For any $\gamma \in (0,1)$ there is a partition into consecutive nonempty blocks $I_j = [a_j, b_j]$ such that every within-block retention is at least γ. With $n_j = |I_j|$ and $d_j = \max_{t\in I_j} m_t$,*

$$\sum_j n_j = T, \qquad \sum_j d_j \le \left(1 + \frac{1}{1-\gamma}\right)T.$$

Proof. Proceed backwards: from endpoint b, choose the smallest $a \le b$ with $\rho_{a,b} \ge \gamma$, then repeat on the remaining prefix. Multiplicativity shows that every internal retention is at least $\rho_{a,b}$. If $b_1 < \cdots < b_J = T-1$ are the block endpoints, maximality gives $\rho_{b_j,b_{j+1}} < \gamma$ for $j < J$. The exact identity

$$m_s(1-\rho_{s,t}) = \sum_{u=s}^{t-1} \prod_{v=s+1}^{u} \beta_v \le t-s$$

therefore implies

$$\sum_{j=1}^{J} m_{b_j} \le \frac{T-1}{1-\gamma} + 1.$$

Since $m_t \le 1 + m_{t+1}$, within each block $d_j \le m_{b_j} + n_j - 1$. Sum these inequalities to conclude. Zero momenta cause no difficulty. □

4 Compressing a block into one checkpoint

Fix for the remainder

$$\gamma = \frac{99}{100}, \qquad \varepsilon = \frac{1}{10}, \qquad C_0 = 8200, \qquad \kappa = \frac{2(1+\varepsilon)}{\gamma} = \frac{20}{9}. \tag{18}$$

For the blocks in Lemma 2, set

$$w_j = \sum_{t \in I_j} h_t, \qquad B_j = \frac{C_0}{\varepsilon} d_j \sqrt{n_j} = 82000 d_j \sqrt{n_j}.$$

Lemma 3 (Virtual-block transfer)**.** *Suppose an incoming frontier has a saved remaining displacement at the endpoint u of an earlier block. Put*

$$c = 2m_u + \sum_{u<t<a_j} h_t.$$

There is a chronological transfer construction through I_j, followed by waiting to b_j, whose gain is at least

$$\frac{w_j}{\kappa(c+B_j)}. \tag{19}$$

A starting construction through I_j reaches b_j with remaining displacement at least

$$\frac{w_j}{\kappa(2 + \sum_{t<a_j} h_t + B_j)}. \tag{20}$$

Proof. Write $I_j = [a,b]$, $X_t = \sum_{s=a}^{t} h_s$, and $X_{a-1} = 0$. Let Q_t be the largest chronological product of ψ over intervals inside $(t,b]$, including the empty product. These are finite maxima and $Q_t \ge 1$.

We first retain the full numerator of the warm response. Multiplicativity implies

$$\mathcal{A}_{u,t} \le \frac{S_{u,t}}{m_u} \sum_{u<s<t} h_s.$$

Using $S_{u,t} \ge 1$ in (16) yields

$$G(u,t) \ge \frac{W_{u,t}}{2(2m_u + \sum_{u<s<t} h_s)}. \tag{21}$$

Append a suffix attaining Q_t, with waits between its transfers and a final wait to b. By (17), the suffix gain, including waits, is at least $\rho_{t,b}Q_t$. Furthermore,

$$W_{u,t}\rho_{t,b} \geq \sum_{s=a}^{t} h_s\rho_{s,b} \geq \gamma X_t.$$

Thus the completed gain is at least

$$\frac{\gamma X_t Q_t}{2(c+X_{t-1})}. \tag{22}$$

Retention is paid once, because the prefix and suffix factors multiply before they are bounded. The starting formula gives the same estimate with $c = 2 + \sum_{s<a} h_s$, using $\Delta_t \leq \sum_{s<t} h_s$ and $F_t\rho_{t,b} \geq \gamma X_t$.

Let A be the maximum of the finite analytic completed-gain products just constructed, each of which has a static realization, and set $Z = 2A/\gamma$. Equation (22) gives

$$X_t Q_t \leq Z(c + X_{t-1}) \qquad (a \leq t \leq b). \tag{23}$$

If $w_j < \varepsilon c$, apply this at b to obtain $w_j \leq Z(1+\varepsilon)c$. Otherwise let t_0 be the first index with $X_{t_0} \geq \varepsilon c$. Since $X_{t_0-1} < \varepsilon c$ and $Q_{t_0} \geq 1$,

$$X_{t_0} \leq Z(1+\varepsilon)c, \qquad Z \geq \frac{\varepsilon}{1+\varepsilon}, \qquad Q_{t_0} \leq Z(1+1/\varepsilon) =: \Lambda.$$

The steps strictly after t_0 form a Λ-admissible sequence, and $\Lambda \geq 1$. Theorem 2, with arbitrary budget, gives

$$\sum_{t=t_0+1}^{b} h_t \leq C_0 d_j Z(1+1/\varepsilon)\sqrt{n_j}.$$

An empty suffix contributes zero. In both cases,

$$w_j \leq Z\big[(1+\varepsilon)c + C_0(1+1/\varepsilon)d_j\sqrt{n_j}\big] = Z(1+\varepsilon)(c+B_j).$$

This is (19) or (20), respectively. A block of zero weight can simply be omitted from a path. □

5 Weighted capacity with a transfer coefficient

Let $w_1, \ldots, w_J \geq 0$ be candidate weights separated by positive primitive gaps $g_0, \ldots, g_J$. Omitting a candidate merges its weight and adjacent gaps additively. For a selection $\mathcal{I} = \{i_1 < \cdots < i_k\}$, denote the successive gaps before its selected candidates by $s_1, \ldots, s_k$ and the final gap by s_f. Define

$$P(\mathcal{I}) = \prod_{r=1}^{k} \frac{w_{i_r}}{\kappa s_r}, \qquad D = \left(\sum_{j=0}^{J} g_j^{2/3}\right)^{3/2}.$$

The empty selection has $P = 1$ and one gap containing the entire list.

Lemma 4. *For every* $0 < \kappa \leq 8 - 4\sqrt{2}$,

$$\max_{\mathcal{I}} \frac{P(\mathcal{I})^2}{4s_f} \geq \frac{1}{8D}.$$

Proof. First bound the one-power minimum $V = \min_{P>0} s_f/P$ by D. We use the extremal-list reduction in [1, Appendix B.1, steps 1–4], replacing the coefficient 2 by κ. Its factorization inequality becomes

$$V \le \frac{\kappa V_{\mathrm{left}} V_{\mathrm{right}}}{w_j},$$

and its insertion multiplier becomes $\kappa xy/[w(x+w+y)]$. The multiplier remains strictly increasing in each neighboring gap. Thus the compactness, zero-weight merging, log-submodularity, and tight-set perturbations in that reduction are unchanged. At a positive maximizing list there is a common tight singleton, and the two tight costs give

$$V = x + y + w_j = \frac{\kappa xy}{w_j} = \frac{x + y + \sqrt{(x+y)^2 + 4\kappa xy}}{2}.$$

The induction therefore closes if this last expression to the 2/3 power is at most $x^{2/3} + y^{2/3}$.

Normalize $x = u^{3/2}$, $y = (1-u)^{3/2}$. The required inequality is $x + y + \kappa xy \le 1$. For $0 < u < 1$, put $s = \sqrt{u} + \sqrt{1-u} \in (1, \sqrt{2}]$. Direct factorization gives

$$\frac{1 - u^{3/2} - (1-u)^{3/2}}{[u(1-u)]^{3/2}} = \frac{4(s+2)}{(s-1)(s+1)^3}.$$

This decreases with s, since its logarithmic derivative is

$$\frac{1}{s+2} - \frac{1}{s-1} - \frac{3}{s+1} < 0.$$

Its minimum at $s = \sqrt{2}$ is $8 - 4\sqrt{2}$. This proves the scalar inequality, with endpoints by continuity, and hence $V \le D$.

Add $\lambda = D$ to the last primitive gap. The modified one-power minimum satisfies

$$V_\lambda \le (D^{2/3} + \lambda^{2/3})^{3/2} = 2^{3/2} D.$$

For a minimizing selection, P is unchanged and $(s_f + \lambda)^2 \ge 4 s_f \lambda$. Consequently

$$\max_{\mathcal{I}} \frac{P(\mathcal{I})^2}{4 s_f} \ge \frac{\lambda}{V_\lambda^2} \ge \frac{1}{8D}.$$

□

Our coefficient $\kappa = 20/9$ is below $8 - 4\sqrt{2}$. The constant in the virtual-block transfer is consequential: an arbitrarily large coefficient would not support the 2/3-power bound.

6 Global embedding and proof of the theorem

Use every retention block as a candidate, with weight w_j, and set

$$\begin{aligned} g_0 &= 2 + B_1, \\ g_j &= 1 + 2m_{b_j} + B_{j+1} \quad (1 \le j < J), \\ g_J &= 1 + 2m_{b_J}. \end{aligned}$$

For any selected list of blocks, its first merged gap covers the denominator $2 + \sum_{j<i_1} w_j + B_{i_1}$ in Lemma 3. A later merged gap, between selected blocks $i < j$, covers

$$2m_{b_i} + \sum_{i<\ell<j} w_\ell + B_j.$$

Its final merged gap covers $1 + \sum_{j>i_k} w_j$. The selected virtual transfers concatenate chronologically; waits at virtual endpoints use no new links. Lemma 1 permits the final penalty to be installed at the last selected block endpoint. Thus one static chain realizes error at least $P(\mathcal{I})^2/(4s_f)$. For the empty selection, the scalar certificate $1/[4(1+H)]$ from [1] suffices, since its merged gap is at least $1 + H$.

It remains to bound D. With $q = 2/3$, subadditivity yields

$$D^q \le 2^q + J + 2^q \sum_{j=1}^{J} m_{b_j}^q + \sum_{j=1}^{J} B_j^q.$$

We have $J \le T$, $d_j \ge 1$, and $\sum_j m_{b_j}^q \le \sum_j d_j \le 101T$. Hölder's inequality and Lemma 2 give

$$\begin{aligned}\sum_j B_j^q &= 82000^{2/3} \sum_j d_j^{2/3} n_j^{1/3} \\ &\le 82000^{2/3} \left(\sum_j d_j\right)^{2/3} \left(\sum_j n_j\right)^{1/3} \le 82000^{2/3} 101^{2/3} T.\end{aligned}$$

In particular, for the absolute constant

$$C_* = 205 + (82000 \cdot 101)^{2/3},$$

we have $D^q \le C_* T$ and $D \le C_*^{3/2} T^{3/2}$. Lemma 4 and the static embedding prove

$$f(x_T) - f_\star \ge \frac{1}{8C_*^{3/2} T^{3/2}} \qquad (L = R = 1).$$

Constants have not been optimized.

Every warm transfer uses a positive-length interval after its preceding checkpoint. There are at most $T - 1$ such transfers, so the static chain has dimension at most $T + 1$. All offsets and thresholds are fixed before the run by the chosen schedule and path, as in [1, Appendix A.4]. The final wait changes neither the dimension nor the number of updates.

For general L, R, apply the normalized result to steps $L\eta_t$ and scale its objective F by $f(x) = LR^2 F((x - b)/R)$. This gives L-smoothness, initialization distance at most R, the stated original heavy-ball updates, and the objective gap multiplied by LR^2. Theorem 1 follows with $c = (8C_*^{3/2})^{-1}$.

References

[1] Jianhao Ma and Jingzhao Zhang. *A Lower Bound for the Heavy-Ball Method on Smooth Convex Functions*. arXiv:2609.08656v1, September 8, 2026. `https://arxiv.org/abs/2609.08656`.

[2] Chang He and Shuzhong Zhang. *Heavy-Ball Method under Randomized Schedules*. arXiv:2609.09743, 2026. `https://arxiv.org/abs/2609.09743`.