RESTARTING HEAVY BALL: SHARP INTERVALS FOR MOMENTUM RETENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Resetting momentum too frequently can destabilize a convergent heavy-ball method. We characterize restart intervals over the complete scalar class with spectral radius at most $s < 1$. Building on the known repeated-root peak effect, we show that full reset is uniformly stable exactly when $s^h[1 + h(1 + s)] < 1$. Retaining a fixed fraction of momentum changes the exact answer: the optimal retention tends to $(3-\sqrt{5})/2$, and we determine the restart interval through its constant-order asymptotic term. Exact rational exclusions and independently checked interval certificates establish exact results at finite margins: at $1 - s = 0.05, 0.02, 0.01$, partial retention permits shorter optimal common intervals than full reset. A secondary gradient-only checkpoint construction preserves an accelerated exponent on fixed quadratics. The experiments separate rejected work from delays in checking the stopping target: observed-target stopping removes all gradient-query overhead when the proposal equals the fallback. A delayed-gradient handoff illustrates a distinct use for velocity control: endpoint loss can be small while a subsequent position limit is violated. Controlled optimization experiments find no general speed advantage from joint-energy certification. The main contribution is a sharp restart design rule, with the scope and costs of additional certificates made explicit.

1 INTRODUCTION

Momentum makes an optimization rule stateful. After its step size or momentum coefficient changes, the new rule inherits both the model parameters and a velocity produced by the old rule. Resetting that velocity is a natural response, but it can also remove information needed for stable dynamics. We ask: *how often can a heavy-ball rule be restarted, and can partial retention reduce the required interval?*

The question connects two established literatures. Repeated-root recurrences produce large transient peaks even when every eigenvalue lies inside the unit disk (Polyak et al., 2018). Danilova et al. (2020) analyze this phenomenon for heavy ball, including Polyak's extremal modes and the equal-initial-iterate trajectory used here. Restart strategies are also central to accelerated optimization (O'Donoghue & Candès, 2015; Roulet & d'Aspremont, 2020). Our full-reset formula uses this known transient; the new claims concern its supremum over a complete margin class, its exact implication for repeated resets, and the best constant-retention policy under arbitrary switching.

Classical switched-system analysis already supplies common Lyapunov functions, dwell-time conditions, and norm-based certificates (Branicky, 1998; Hespanha & Morse, 1999; Liberzon, 2003; Boyd et al., 1994; Ahmadi et al., 2014). Heavy-ball reset and damping policies have more specific analyses (Le & Teel, 2021). A generic existence theorem, however, does not give the exact restart interval, and resetting everything is not the optimal static policy in the class we study.

Contributions. First, we derive a necessary and sufficient full-reset interval and identify its connection to optimally tuned heavy ball. Second, we sharpen the competing-mode lower bound to an exact asymptotic optimum, including the limiting retained fraction. For three fixed margins, complete interval certificates also establish the exact finite optimum. Third, we separate two uses of certification: preserving a convergence contract under speculative rule changes, and enforcing a task-level velocity requirement. The checkpoint result is a conservative wrapper, not a new accelerator; the experiments compare it with unguarded heavy ball and direct checking.

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

Optimizer-state handling also arises in switching from adaptive moment estimation (Adam) to stochastic gradient descent (SGD), warm starts, optimizer resets, and dynamic interleaving (Keskar & Socher, 2017; Ash & Adams, 2020; Asadi et al., 2023; Huang et al., 2025; Chen et al., 2026). Those applications motivate the question but do not extend the mathematical model. Our results concern deterministic heavy-ball velocity, not adaptive second moments or general recursive self-improvement.

2 Model and scope

On a fixed quadratic $f(x) = \frac{1}{2}(x - x_\star)^T H(x - x_\star)$, let $\mu I \preceq H \preceq LI$ and $e = x - x_\star$. Heavy ball (HB) updates

$$v^+ = \beta v - \alpha \nabla f(x), \qquad x^+ = x + v^+. \tag{1}$$

In a scalar Hessian mode, $a = \alpha\lambda$ and $z = (e, v)$ obey

$$A(a, \beta) = \begin{pmatrix} 1 - a & \beta \\ -a & \beta \end{pmatrix}, \qquad J(r) = \mathrm{diag}(1, r), \quad 0 \le r \le 1. \tag{2}$$

A boundary first replaces v by rv and then executes h updates, giving $A^h J(r) z$. A frozen rule is Schur stable if all its eigenvalues have modulus below one. With $\beta \ge 0$, this means $\beta < 1$ and $0 < a < 2(1 + \beta)$. A fixed periodic product and an arbitrary sequence of allowed products impose different stability requirements.

Scope and information. The principal theorems concern a complete scalar spectral-margin class, with a modewise extension in a common orthonormal basis. The checkpoint theorem assumes a fixed quadratic, exact gradients, and valid supplied spectral bounds. Exact-Hessian experiments are mechanism diagnostics: that oracle could solve the quadratic directly. Nonlinear guards are studied separately. We make no stochastic-gradient, Adam-state, neural-optimizer, or recursive-capability claim.

For $s = 1 - \eta \in (0, 1)$, define

$$\mathcal{A}_s = \{A(a, \beta) : \beta \ge 0,\ \rho(A) \le s\}. \tag{3}$$

It is equivalently parameterized by $0 \le \beta \le s^2$ and $(1-s)(1-\beta/s) \le a \le (1+s)(1+\beta/s)$. We seek stability for *every* switching sequence in this compact class. More precisely, for a fixed policy (h, r) there must be constants $C < \infty$ and $\gamma < 1$ such that every length-n product of allowed block maps has norm at most $C\gamma^n$. Constants may depend on s, h, r but not on the chosen sequence. This is a worst-case design problem, not a claim that an optimizer normally selects the most harmful sequence. Because the class is compact and h is fixed, bounded internal powers transfer a boundary guarantee to the full trajectory. Its two repeated-root rules are $A_- = A((1 + s)^2, s^2)$ and $A_+ = A((1 - s)^2, s^2)$.

3 Full reset: an exact restart-interval rule

Theorem 1 (Sharp full-reset interval)**.** *For every integer $h \ge 1$,*

$$\sup_{A \in \mathcal{A}_s} |(A^h)_{11}| = D_s(h) := s^h[1 + h(1 + s)]. \tag{4}$$

The supremum is attained by A_-. Resetting momentum at each h-step boundary is uniformly exponentially stable under arbitrary switching in $\mathcal{A}_s$ if and only if $D_s(h) < 1$. Thus the exact design rule is

$$h_{\mathrm{reset}} = \min\{h \in \mathbb{N} :\ h \log(1/s) > \log[1 + h(1 + s)]\}. \tag{5}$$

As $\eta \downarrow 0$,

$$h_{\mathrm{reset}}(\eta) = \frac{\log(1/\eta) + \log\log(1/\eta) + \log 2 + o(1)}{\eta}. \tag{6}$$

Why the condition is exact. Let $u_h = \sum_{j=0}^{h-1} \lambda_1^{h-1-j} \lambda_2^j$. Cayley–Hamilton gives $A^h = u_h A - \beta u_{h-1} I$, and $|u_h| \le h s^{h-1}$. In particular, $(A^h)_{11} = u_{h+1} - \beta u_h$, whose two terms attain their combined absolute bound when both roots are $-s$. Resetting reduces the next boundary position to this scalar factor. Repeating the extremal rule proves necessity, including failure at equality. Appendix B gives the complete proof.

Relation to Polyak's modes and restart literature. For $\alpha_\star = 4/(\sqrt{L}+\sqrt{\mu})^2$ and $s = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$, where $\kappa = L/\mu$, the two extreme modes are precisely $\alpha_\star L = (1+s)^2$ and $\alpha_\star \mu = (1-s)^2$ (Polyak, 1964; Danilova et al., 2020). Consequently, unchanged, optimally tuned heavy ball on a quadratic containing its top mode, with velocity reset every h steps, converges for every initialization exactly under (5). The trajectory attaining the maximum is the known negative repeated-root peak after an index shift; the complete-class supremum and restart equivalence are the statements established here.

The interval has order $\sqrt{\kappa}\log\kappa$, unlike a restart based on a uniform accelerated function-error bound. For example, if a method guarantees $f(x_h) - f_\star \le CL\|e_0\|^2/h^2$ for a universal C, strong convexity gives a restart-gap ratio at most $2C\kappa/h^2$. Taking $h \ge 2\sqrt{C\kappa}$ halves that gap. This is the $O(\sqrt{\kappa})$ restart mechanism used in sharpness-based analyses (Roulet & d'Aspremont, 2020); it is not the repeated-root transient of Polyak heavy ball. Adaptive restarting also uses trajectory information absent from a static policy (O'Donoghue & Candès, 2015). For implementation, use the strict integer inequality (5), not the asymptotic expansion. At $\eta = 0.05$, the exact answer is 104; the leading asymptotic expression is only about 96.

Reading the design rule. The factor $D_s(h)$ measures the size of the position at the next reset, not the largest value reached during the block. A rule may converge without resetting, yet repeated resets can keep exciting its negative-root transient before it has decayed. Increasing h past the descending crossing makes that repeated excitation harmless. Neither eigenvalues inside the unit disk for each fixed rule nor a one-step test from zero momentum answers the same question.

For a bank whose normalized pairs (a,β) lie in the declared class, the rule gives a guaranteed interval without numerically finding a common Lyapunov matrix. One can check membership using the two scalar parameter inequalities in Section 2. If all modes share one orthonormal basis, applying the scalar bound in each mode yields the same conclusion in any finite dimension. Changing eigenvectors or noncommuting preconditioners is a different class; the scalar theorem does not cover that by dimension counting.

4 Partial retention: the optimal interval

Retaining some velocity can offset the negative-root peak, but it creates an opposing switching cost. Let $h_\star(s)$ be the smallest common interval for which *some fixed* $r \in [0,1]$ stabilizes every sequence in $\mathcal{A}_s$. Optimality here means the smallest uniformly stable interval, not fastest convergence or minimum finite-budget loss.

Theorem 2 (Sharp constant-retention asymptotics)**.** *Let $\delta = -\log s$ and $\varphi = (1+\sqrt{5})/2$. Then*

$$\delta h_\star(s) = \log(1/\delta) + \log\log(1/\delta) + \log(2/\varphi) + o(1). \tag{7}$$

Any sequence of stabilizing retentions attaining these minimum intervals satisfies

$$r \longrightarrow r_\star = \varphi^{-2} = (3-\sqrt{5})/2. \tag{8}$$

Compared with full reset, the optimal static policy saves $[\log\varphi + o(1)]/\delta$ steps per boundary. The leading logarithmic cost remains.

The competing conditions. Write $p = s^h$ and $H_h = h(1+s)$. Repeating $A_-^h J(r)$ requires

$$F_1 = (1-p)(1-rp) - pH_h(1-r) > 0. \tag{9}$$

Alternating the positive- and negative-root blocks also requires

$$\begin{aligned} F_2 &= 1 - p^2 T_h + r^2 p^4 > 0, \\ T_h &= [(1-s^2)(1-r)^2 + 4r]h^2 + 2(1-r^2)h + 1 + r^2. \end{aligned} \tag{10}$$

These are exact Jury margins. They already apply to the *two-rule bank* $\{A_-, A_+\}$; a finite bank containing these witnesses does not evade the lower bound. Alternating the two maps in one coordinate is different from running the unchanged Polyak rule on separate Hessian modes.

If $hs^h \to c > 0$, the two constraints imply $2c(1-r) \le 1$ and $2c\sqrt{r} \le 1$. Their best balance is $1 - r = \sqrt{r}$, giving $r_\star$ and $c_\star = \varphi/2$. The sufficiency proof requires more than these examples. We

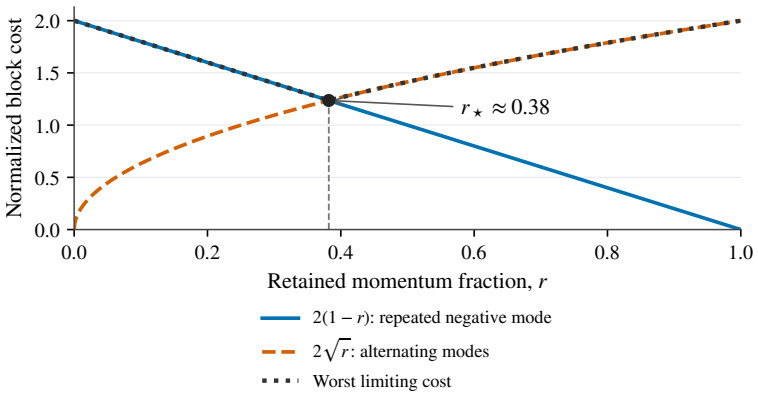

Figure 1: The two competing normalized costs. Erasure controls $2(1-r)$; alternating retained states controls $2\sqrt{r}$. Their maximum is minimized at $r_\star \simeq 0.38$. This is the complete-class limiting norm in (11), not a fitted empirical curve.

prove that the entire matrix class collapses, in this scaling limit, onto two rank-one directions. In the norm $\|S_r^{-1} z\|_1$, with

$$S_r = \begin{pmatrix} \sqrt{r} & 1 \\ 0 & 2 \end{pmatrix}, \qquad \sup_{A \in \mathcal{A}_s} \|S_r^{-1} A^h J(r) S_r\|_1 \longrightarrow 2c \max\{1-r, \sqrt{r}\}, \tag{11}$$

any strict $c < c_\star$ contracts every allowed block at $r = r_\star$. This matches the lower bound through the constant term in (7). Appendix B.6 proves the uniform limit, rather than inferring it from a grid of eigenvalues.

Why the limiting norm covers more than two examples. The lower bound alone would not establish an optimal policy: another admissible root configuration could be worse. The uniform-limit argument rules this out. Consider any sequence of matrices that could maximize the block norm as $s \uparrow 1$ and $hs^h \to c$. If the limiting roots are distinct, the divided difference in u_h tends to zero. If both roots stay strictly inside the unit circle, geometric decay also forces the block to zero. Only roots that coalesce at $+1$ or -1 can retain a nonzero transient on this scale.

Those two cases converge to the rank-one directions

$$\begin{pmatrix} 0 & r \\ 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 2 & -r \\ 4 & -2r \end{pmatrix},$$

up to scalar factors of magnitude at most c and an overall sign that does not affect the norm. The explicit basis S_r makes their largest column sums exactly the two terms in Figure 1. Compactness then upgrades the statement about every subsequential limit to the supremum over the entire class. This proves sufficiency: below the critical c, one common norm contracts every allowed block, so no untested switching word can invalidate the asymptotic guarantee.

Closing the finite-margin gap. At fixed s, F_1 is increasing and F_2 is strictly decreasing in r. We exclude every smaller integer h by exact rational evaluation at the root of F_1. At the first surviving interval, we supply a separate common-norm certificate over *all* of $\mathcal{A}_s$, not only the three witness products. Table 1 therefore reports exact optimal common intervals, not just necessary ones.

The certificates split spectra into real pairs and conjugate pairs, with both possible signs of the real part. Matrix powers are evaluated through cancellation-resistant formulas. The supplied 407 cover boxes were checked once with outward-rounded interval arithmetic and independently with exact rational arithmetic, Taylor remainders, and a checked enclosure of π. In particular, at $\eta = 0.05$, $r = 3727/10000$ and $h = 93$ work for arbitrary switching: partial retention reduces the required interval from 104 to 93. Appendix B.7 states the covering and arithmetic argument.

From a candidate basis to a verifiable design. Finding a good two-dimensional norm numerically is inexpensive, but numerical optimization alone is not used to establish sufficiency. For each finite

Table 1: Exact finite-margin optima. Rational Jury exclusions prove that every smaller h fails for every constant r. At the displayed r, a complete interval cover verifies a common induced norm below one for the entire continuous class. Boxes count the supplied cover cells.

η	Optimal $h_\star$	Full reset	Certified r	Boxes
0.05	93	104	3727/10000	172
0.02	292	320	189/500	125
0.01	669	724	19/50	110

margin, we freeze the proposed retention and basis as decimal rational numbers. We then enclose the induced norm on every cell of a complete root-parameter cover. A cell is split only when its upper bound is too large. The accepted cells form a complete cover, and their upper bounds are all strictly below one. This turns a search result into finitely many checkable inequalities.

The necessity proof also avoids a retention grid. For an interval shorter than full reset, the negative-root margin demands $r > \ell_h$, where ℓ_h is an explicit rational number. The alternating-pair margin decreases throughout $[0, 1]$. When it is already nonpositive at ℓ_h, increasing retention cannot make both tests pass. Repeating this exact sign test for every smaller integer h gives the exclusion half of Table 1. At the first surviving interval, the whole-class certificate supplies the other half. Thus the reported minimum is not an extrapolation from a fine sample of r values.

Other schedules. Retentions or commit lengths chosen from observed states are outside the static optimum. A simple sufficient envelope for varying retention is still available:

$$|A^h J(r)| \le M_s(h, r) = \begin{pmatrix} D_s(h) & rhs^{h+1} \\ h(1+s)^2 s^{h-1} & rs^h[h(1+s)-1] \end{pmatrix}. \tag{12}$$

Testing $\rho(M_s(h, r)) < 1$ permits arbitrary retentions up to r. This envelope loses sign cancellation and is not the method used to certify the optimal finite intervals. Periodic retention sets can also be disconnected; Appendix A gives an exact example and tests for specified unequal block lengths.

5 Observable checkpoints and their limits

A static restart rule protects against every allowed sequence. A different approach tests a proposed block, then accepts or rolls back after observing its actual gradient. This is standard safeguard logic (Prémont-Schwarz et al., 2022; Heaton et al., 2023), here coupled to a heavy-ball fallback that allows nonmonotone transients.

Use supplied bounds $0 < \mu < L$ containing the spectrum of the fixed quadratic. They need not be tight. With Polyak's $(\alpha_\star, \beta_\star)$, the observable state $y = (g, Lv)$ and energy $W = \|y\|^2$ have modal matrix

$$B_\lambda = \begin{pmatrix} 1 - \alpha_\star \lambda & \beta_\star \lambda / L \\ -\alpha_\star L & \beta_\star \end{pmatrix}. \tag{13}$$

Its radius is s and its norm is at most $\sqrt{27}$. Define

$$B_s(h) = hs^{h-1}\sqrt{27} + (h-1)s^h, \quad \bar{s} = (1+s)/2, \quad q = \bar{s}^2, \tag{14}$$

and choose h with $B_s(h) \le \bar{s}^h$.

Theorem 3 (Gradient-only accelerated checkpoint)**.** *For $0 < \varepsilon < W_0$, save (x, v, g) and test a block costing at most h gradient queries. Accept its output only if $W^+ \le q^h W$; otherwise roll back and run h Polyak heavy-ball steps. Every completed block contracts by q^h, for arbitrary history-dependent proposals. The cost to reach $W \le \varepsilon$, including rejected trials, is*

$$O\big(1 + \sqrt{\kappa}[\log \kappa + \log(W_0/\varepsilon)]\big). \tag{15}$$

At any already evaluated trial or fallback state with $W \le \varepsilon$, the algorithm may instead terminate immediately; this cannot increase the query bound.

There are three separate statements: an accelerated exponent at completed checkpoints, an additive $O(\sqrt{\kappa}\log\kappa)$ transient allowance, and possible practical overhead. Only the first two are guarantees.

Theorem 2 does not prove that this wrapper's logarithmic allowance is necessary. Observed-target termination also is not a claim of contract satisfaction at an unfinished checkpoint; it simply returns a state meeting the requested target.

The limitation is the fallback, not rollback bookkeeping. Polyak parameters do not guarantee acceleration on general strongly convex nonlinear functions (Lessard et al., 2016). A nonlinear accelerated extension would require an appropriate observable Lyapunov function for its fallback (Wilson et al., 2021; Su et al., 2016; Wibisono et al., 2016). Likewise, inward errors in supplied spectral bounds can invalidate the certificate. Underestimating μ or overestimating L preserves enclosure; the reverse directions need checking. For the chosen parameters, frozen stability requires $\lambda < L + \mu$. Thus an underestimated L can cause outright instability, while an overestimated μ can invalidate the rate even when the frozen rule remains stable. Appendix E.1 gives explicit examples.

6 When does a velocity contract serve a task?

For an exact linear trial block $C = A^h$, scalar retention gives

$$CJ(r)z = p + rb, \qquad p = C(e, 0), \quad b = C(0, v). \tag{16}$$

Both endpoint loss and joint energy are quadratic in r. A joint-energy sublevel is an interval, so minimizing loss on it is inexpensive. Without the constraint, minimizing over all of $[0, 1]$ has no greater loss at the same incoming state. This is a direct consequence of set inclusion, not a convergence comparison between future trajectories.

Loss convergence already implies eventual velocity convergence for one-step updates: $v_{t+1} = x_{t+1} - x_t$. The reason to impose a stronger contract must therefore be a separate requirement. Consider a scalar deployed state that must stay within $|e + jv| \le R$ for $j = 0, \dots, K$ velocity-only ticks while the next gradient is delayed. A sufficient interface condition is

$$e^2 + v^2 \le \frac{R^2}{1 + K^2}, \tag{17}$$

because $|e + jv| \le \sqrt{1 + j^2}\sqrt{e^2 + v^2}$. The threshold is sharp among Euclidean energy balls. Exact path checking, when available, uses the two strips $|e| \le R$ and $|e + Kv| \le R$ and is less conservative.

The difference matters even in a simple example. Starting at $(e, v) = (1/\sqrt{5}, 0)$, two steps with $a = 2.5$, $\beta = 0.9$ reach $(0, 3/(2\sqrt{5}))$. Endpoint loss is zero, yet two subsequent velocity ticks reach $3/\sqrt{5} > 1$. A loss-only fallback test accepts this handoff; the joint-energy condition or the exact path condition rejects it. This is a constructed safety diagnostic with simulated candidate blocks and a constraint on the deployed handoff, not a physical experiment or a speed claim.

7 Experiments

The experiments comprise $15{,}968$ deterministic trajectories across six suites, including matched-method runs with methods sharing tasks and seeds. Primary progress uses

$$A_Z = \frac{1}{T} \sum_{t=1}^{T} \mathrm{clip}_{[-12,12]} \log_{10}(Z_t / Z_0), \tag{18}$$

with $Z = f - f_\star$ for exact-geometry diagnostics and $Z = W$ for observable guards. Lower is better. Useful-accuracy comparisons use the first 10^{-8} crossing, with runs that miss the target recorded separately. Paired intervals use $5{,}000$ seed-cluster bootstrap resamples (Efron & Tibshirani, 1994); distinct matrix families and proposal regimes are not pooled.

7.1 Checkpoint stopping and rejected work

We use rotated, shifted quadratics with $d \in \{8, 20\}$, $\kappa \in \{10, 100, 1000\}$, and 16 seeds: 672 runs across seven policies. The optimizer receives a gradient callable and spectral bounds, not H or $x_\star$. Commit lengths are $14, 58, 228$. The HB proposal uses *exactly the fallback's* Polyak parameters. The

Table 2: Median gradient calls to $W/W_0 \le 10^{-8}$, including rejected work and initialization. Every method uses $W = \|g\|^2 + L^2\|v\|^2$ for stopping. The one-step guard accepts using $\|g\|^2 + (\mu L/2)\|v\|^2$ and two function-value queries. At $\kappa = 1000$, gradient descent has 28.1% hits; its median is the cap plus initialization. Other rows have 100% hits.

Policy and stopping rule	$\kappa = 10$	100	1000
Gradient descent	65	462	2049
Heavy ball	21	74	255
One-step guard	22	62	172.5
HB: endpoint stop	29	117	457
HB: observed-target stop	21	74	255
Switch: endpoint stop	43	175	602
Switch: observed-target stop	35	132	400

switching proposal multiplies $\alpha_\star$ by 1.8 on attempts $1, 4, 7, \ldots$. Identical fallback blocks are never rerun.

The identical HB checkpoint rejects zero trials in all 96 configurations. Its endpoint-only overhead is entirely due to checking the target only at block endpoints: observed-target stopping exactly matches ordinary HB's query count in every pair. The switching checkpoint rejects one trial per run with either stopping rule. Median discarded work is $14, 58, 145$ gradients; the last value includes early termination of an overflowing trial. At $\kappa = 1000$, observed-target stopping reduces its median cost from 602 to 400, not to HB's 255. The one-step guard still uses the fewest gradients there, 172.5, with additional function access. All tested completed checkpoint contracts hold. The accepted trajectories have nonmonotone loss, so the construction is not enforcing per-step descent.

The cost comparison is paired at the task level. Plain HB and the identical-proposal checkpoint begin with the same state, use the same parameters, and generate the same gradient sequence. Their equality under observed-target stopping is therefore expected and is confirmed in every run. For the switching proposal, the early-stop trajectory after rollback also equals the fallback trajectory; its additional calls are exactly the discarded trial. This distinction separates an implementation choice—when to inspect an already available target—from the cost of testing an unsuccessful proposal. Counting only accepted updates would hide the latter.

Function-value access matters for interpreting the smallest gradient count. The one-step comparator uses two loss evaluations per decision as well as its gradient checks. We charge these calls separately rather than pretending that its gradient count is its entire computational budget. On these inexpensive quadratics it is still an effective alternative. The checkpoint construction instead addresses a specification that permits exact gradients and supplied spectral bounds but does not require function-value queries.

7.2 Delayed-gradient handoff

We test $1{,}024$ configurations, each with four policies: $4{,}096$ simulated handoffs. A two-step cancellation-focused proposal family creates small endpoint loss with potentially large outgoing velocity. All candidates are simulated before deployment; the independent position limit applies during the subsequent $K \in \{1, 2, 4, 8\}$ ticks. Policies share endpoint-loss minimization and the fallback. They differ only by no constraint, a contracting energy ball, the interface ball without contraction, or the exact path constraint.

At $K = 2$, loss-only selection violates the position limit in $137/256$ cases despite safe endpoints. All three constrained policies have zero failures. The contracting energy test accepts $35/256$ proposals; ball membership alone accepts $89/256$; direct path checking accepts $134/256$. Thus this task needs a velocity constraint, but not a geometric contraction rate. The controls separate the additional cost of contraction from the conservatism of the isotropic ball. Resetting outgoing velocity would also make these endpoints safe without changing their loss; that is a different output intervention, not a demonstrated advantage of certification over resetting. The rates describe the constructed stress family, not the prevalence of unsafe switches in training.

Table 3: Delayed-gradient stress diagnostic; all entries except K are percentages over 256 configurations per row. Failure means an actual deployed position exceeds $R = 1$, not merely violation of an energy inequality. Every returned endpoint is initially inside the allowed region; all three constrained policies have zero path failures.

K	Loss-only fail	Contract accept	Ball accept	Path accept
1	23.4	13.7	34.8	64.1
2	53.5	13.7	34.8	52.3
4	57.0	13.7	34.8	42.2
8	58.2	13.7	34.8	37.5

Safety is evaluated on the returned trajectory, not on whether a policy satisfies its own acceptance inequality. In particular, the loss-only policy's selected endpoint is within the position limit in every configuration, and every proposed frozen rule is stable. Failures appear only during the independently specified delayed-gradient ticks. The resulting experiment distinguishes a real constraint violation from an abstract energy-budget violation. Ball membership is a reusable sufficient interface condition, while the direct-path comparator exploits the complete known handoff dynamics. Its higher acceptance demonstrates the cost of using a generic joint-state bound instead of that task-specific information.

7.3 Optimization controls and curvature costs

Exact-Hessian replay comprises 5,376 runs with identical proposal sequences, two-probe retention searches, and matched fallback comparisons. Certified minus unconstrained retention has mean clipped progress differences $+0.02, +0.10, +0.08, +0.68$ across commuting/noncommuting and conservative/aggressive strata; all four paired intervals are positive. Additional search over retention, not the energy constraint, explains gains over testing only full retention.

A separate 1,024-run comparison tunes $P_c = \mathrm{diag}(H, cH)$ on disjoint calibration seeds. Tuning raises uniform coverage from 36.2% to 60.1%; statewise checking in that same metric reaches 89.7%. This comparison concerns that metric family. A 960-run two-scale proposal-control experiment is included as an evolution-strategy diagnostic, without learned-optimizer or recursive-capability interpretation.

For nonlinear observable guards, 3,840 runs compare local Hessian-error bounds, spectral-interval bounds, and actual-gradient checks. With $d(r) = -\alpha g + \beta r v$ and $\tau = \mu L/2$, the local certificate is

$$U_{\mathrm{loc}}(d) = \left(\|g + \widehat{H}d\| + \epsilon\|d\| + \tfrac{M}{2}\|d\|^2\right)^2 + \tau\|d\|^2. \tag{19}$$

Available spectral information also gives

$$U_{\mathrm{sp}}(d) = \left(\|g + cd\| + b\|d\|\right)^2 + \tau\|d\|^2, \quad c = (L+\mu)/2,\ b = (L-\mu)/2. \tag{20}$$

Their convex sublevels can be solved separately and combined as a union. On logistic tasks refreshed every eight steps, this increases acceptance from 25.0% to 60.3%, but does not improve the stale-curvature target-hit rate. Postchecking remains faster and more successful on these tasks. Appendix D defines the uncertainty sets and proves the corresponding bounds. These experiments do not establish optimizer superiority.

7.4 Three distinct uses of the results

The experiments and certificates answer different questions, so a single ranking by final loss would be misleading. For a static rule bank, the input is a spectral-margin specification and the output is a pair (h, r) valid for every allowed switching sequence. Table 1 provides three such designs together with proofs of minimality. The full-reset inequality is also a direct design rule at any specified margin. A larger interval should not be assumed safe at the same partial retention without checking it: the finite certificates concern the reported block length, whereas the full-reset factor has a single descending crossing. These results avoid validating each possible future switching sequence separately, but optimize the minimum stable interval rather than the rate of loss reduction.

A checkpoint addresses a different information regime. The proposed rule need not belong to a known margin class; instead its realized output is tested and a failed trial can be rolled back. This requires the ability to save state and pay for speculative work. The bound protects progress at completed checkpoints, not the largest intermediate position during a trial. Observed-target stopping uses the same available gradients and removes unnecessary work after success. It cannot recover the queries already spent on an unsuccessful proposal. The query accounting therefore separates saved-state fallback cost, rejected work, and the use of function values by the one-step comparator.

An independently constrained handoff calls for a third contract. A small endpoint objective says where the model is, but not how far an executor will move before receiving another gradient. The ball and path controls use that additional requirement directly. The ball-only ablation shows that a geometric energy decrease is stronger than this one-time safety specification needs. A direct path test or an allowed outgoing-velocity reset can be preferable. The positive observation is that endpoint loss alone misses the failure mechanism, not that the most restrictive energy certificate is the best controller.

Finally, the finite-margin proof and the trajectory measurements support different kinds of claims. Complete interval covers establish a continuous-class statement; they do not estimate a failure probability from random tests. Conversely, the handoff rates and optimizer timings characterize stated finite experimental designs. The repeatability check reruns all six suites and compares every non-timing scientific measurement for exact agreement. This checks the implementation and table provenance, while the analytic arguments and independently replayed rational inequalities support the mathematical claims.

8 Discussion

The main result is a sharp answer for static heavy-ball restarts. Partial retention improves the constant-order term, and independent whole-class certificates make that improvement exact at three finite margins. The repeated-root transient itself is established prior work. The switching optimum, not a new peak formula, is what the theory adds.

Certification serves different purposes in the two operational examples. Checkpoints preserve a fallback's convergence contract under untrusted proposals; early stopping removes avoidable granularity cost but not discarded work. A handoff energy bound protects a separate position requirement, while direct path checking is less conservative. Neither example supports a general optimization speedup. Extending the static theory beyond common modal coordinates, or replacing the quadratic checkpoint fallback by an observable nonlinear accelerated certificate, are distinct mathematical problems. Stochastic gradients and adaptive second-moment state require further modeling rather than a change of terminology.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amir Ali Ahmadi, Raphaël M. Jungers, Pablo A. Parrilo, and Mardavij Roozbehani. Joint spectral radius and path-complete graph lyapunov functions. *SIAM Journal on Control and Optimization*, 52 (1):687–717, 2014. doi: 10.1137/110855272. URL https://arxiv.org/abs/1111.3427.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1606.04474.

Kavosh Asadi, Rasool Fakoor, and Shoham Sabach. Resetting the optimizer in deep RL: An empirical study. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72284–72324, 2023. doi: 10.52202/075280-3164. URL https://arxiv.org/abs/2306.17833.

Jordan T. Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/1910.08475.

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998. doi: 10.1287/moor.23.4.769.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics*. SIAM, 1994. URL https://stanford.edu/~boyd/lmibook/.

Michael S. Branicky. Multiple lyapunov functions and other analysis tools for switched and hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):475–482, 1998. doi: 10.1109/9.664150.

Yile Chen, Zeyi Wen, Jian Chen, and Jin Huang. Towards dynamic interleaving optimizers. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AII8ADdDHt.

Marina Danilova, Anastasiia Kulakova, and Boris Polyak. Non-monotone behavior of the heavy ball method. In *Difference Equations and Discrete Dynamical Systems with Applications*, pp. 213–230. Springer, 2020. doi: 10.1007/978-3-030-35502-9_9. URL https://arxiv.org/abs/1811.00658.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL https://jmlr.org/papers/v12/duchi11a.html.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1994. URL https://www.routledge.com/link/link/p/book/9780412042317.

James Harrison, Luke Metz, and Jascha Sohl-Dickstein. A closer look at learned optimization: Stability, robustness, and inductive biases. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.11208.

Howard Heaton, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Safeguarded learned convex optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7848–7855, 2023. doi: 10.1609/aaai.v37i6.25950. URL https://ojs.aaai.org/index.php/AAAI/article/view/25950.

João P. Hespanha and A. Stephen Morse. Stability of switched systems with average dwell-time. In *Proceedings of the 38th IEEE Conference on Decision and Control*, volume 3, pp. 2655–2660, 1999. doi: 10.1109/CDC.1999.831330. URL https://www.eng.yale.edu/controls/1999/average.pdf.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.06842.

Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from Adam to SGD. arXiv:1712.07628, 2017. URL https://arxiv.org/abs/1712.07628.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Louis Kirsch and Jürgen Schmidhuber. Eliminating meta optimization through self-referential meta learning. arXiv:2212.14392, 2022. URL https://arxiv.org/abs/2212.14392.

Justin H. Le and Andrew R. Teel. Hybrid heavy-ball systems: Reset methods for optimization with uncertainty. In *2021 American Control Conference*, pp. 2236–2241, 2021. doi: 10.23919/ACC50511.2021.9482790. URL https://arxiv.org/abs/2009.13770.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. doi: 10.1137/15M1009597. URL https://arxiv.org/abs/1408.3595.

Daniel Liberzon. *Switching in Systems and Control*. Birkhäuser, 2003. doi: 10.1007/978-1-4612-0017-8.

Luke Metz, Niru Maheswaranathan, Jeremy Nixon, C. Daniel Freeman, and Jascha Sohl-Dickstein. Understanding and correcting pathologies in the training of learned optimizers. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4556–4565, 2019. URL https://proceedings.mlr.press/v97/metz19a.html.

Luke Metz, Niru Maheswaranathan, C. Daniel Freeman, Ben Poole, and Jascha Sohl-Dickstein. Tasks, stability, architecture, and compute: Training more effective learned optimizers, and using them to train themselves. arXiv:2009.11243, 2020. URL https://arxiv.org/abs/2009.11243.

Luke Metz, C. Daniel Freeman, Niru Maheswaranathan, and Jascha Sohl-Dickstein. Training learned optimizers with randomly initialized learned optimizers. arXiv:2101.07367, 2021. URL https://arxiv.org/abs/2101.07367.

Luke Metz, C. Daniel Freeman, James Harrison, Niru Maheswaranathan, and Jascha Sohl-Dickstein. Practical tradeoffs between memory, compute, and performance in learned optimizers. In *Proceedings of the 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 142–164, 2022a. URL https://proceedings.mlr.press/v199/metz22a.html.

Luke Metz, James Harrison, C. Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, and Jascha Sohl-Dickstein. VeLO: Training versatile learned optimizers by scaling up. arXiv:2211.09760, 2022b. URL https://arxiv.org/abs/2211.09760.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. doi: 10.1007/s10107-006-0706-8.

Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015. doi: 10.1007/s10208-013-9150-3. URL https://arxiv.org/abs/1204.3982.

Boris 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.

Boris T. Polyak, Pavel S. Shcherbakov, and Georgi Smirnov. Peak effects in stable linear difference equations. *Journal of Difference Equations and Applications*, 24(9):1488–1502, 2018. doi: 10.1080/10236198.2018.1504930. URL https://arxiv.org/abs/1803.00808.

Isabeau Prémont-Schwarz, Jaroslav Vítků, and Jan Feyereisl. A simple guard for learned optimizers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17910–17925, 2022. URL https://proceedings.mlr.press/v162/premont-schwarz22a.html.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020. doi: 10.1137/18M1224568. URL https://arxiv.org/abs/1702.03828.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016. URL https://jmlr.org/papers/v17/15-084.html.

Tao Sun, Penghang Yin, Dongsheng Li, Chun Huang, Lei Guan, and Hao Jiang. Non-ergodic convergence analysis of heavy-ball algorithms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):5033–5040, 2019. doi: 10.1609/aaai.v33i01.33015033. URL https://ojs.aaai.org/index.php/AAAI/article/view/4435.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113. URL https://arxiv.org/abs/1603.04245.

Olga Wichrowska, Niru Maheswaranathan, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Nando de Freitas, and Jascha Sohl-Dickstein. Learned optimizers that scale and generalize. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3751–3760, 2017. URL https://proceedings.mlr.press/v70/wichrowska17a.html.

Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021. URL https://www.jmlr.org/papers/v22/20-195.html.

A PROOFS FOR PERIODIC SCHEDULES

A.1 SCALAR STABILITY AND THE DISCONNECTED FRONTIER

The characteristic polynomial of (2) is $\lambda^2-(1-a+\beta)\lambda+\beta$. For a real monic quadratic $\lambda^2-T\lambda+D$, both roots have modulus below one exactly when

$$1-D>0, \qquad 1-T+D>0, \qquad 1+T+D>0. \tag{21}$$

One direct derivation maps the open unit disk to the left half-plane by $\lambda=(1+w)/(1-w)$. After multiplication by $(1-w)^2$, the polynomial has coefficients $1+T+D$, $2(1-D)$, and $1-T+D$, whose strict positivity is the real second-order Hurwitz condition. Degenerate boundary cases follow by continuity or direct substitution; they are not asymptotically stable. Applying (21) to A with $\beta\geq 0$ gives $\beta<1$, $a>0$, and $a<2(1+\beta)$.

The identity $A(a,\beta)J(r) = A(a,\beta r)$ gives a direct trace calculation. For $C = A(a_2,\beta_2)J(r_2)A(a_1,\beta_1)J(r_1)$, write $c_i=\beta_i r_i$. Then

$$\operatorname{tr} C=(1-a_1)(1-a_2)-a_1c_2-a_2c_1+c_1c_2, \qquad \det C=c_1c_2. \tag{22}$$

The third margin in the symmetric case is

$$1-d^2-2\beta r+2\beta^2r^2>0. \tag{23}$$

In the symmetric family $a_1=1-d$, $a_2=1+d$ with common $c=\beta r$, the trace is $-d^2-2c+c^2$ and determinant c^2. The three inequalities become $1-c^2>0$, $1+d^2+2c>0$, and (23). When $d^2>1/2$, the last polynomial has roots

$$r_\pm=\frac{1\pm\sqrt{2d^2-1}}{2\beta}. \tag{24}$$

Intersect the two exterior open intervals with $[0,1]$. At $d=0.9$, $\beta=0.95$, the endpoints are approximately 0.11 and 0.94. At $\beta=0.5$, only the first component remains: $r<1-\sqrt{0.62}\approx 0.21$. The comparison is for an alternating two-step schedule. It does not justify arbitrary switching among those modes.

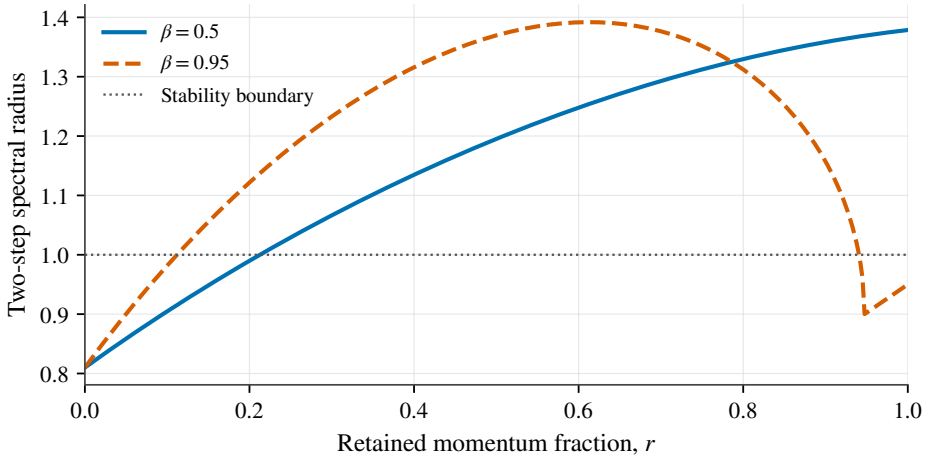

Figure 2: Exact two-step spectral radius as a function of retention. At momentum coefficient 0.95, both complete reset and complete retention are stable, but an intermediate band is unstable. This figure illustrates a periodic product, not a common-metric certificate.

A.2 ARBITRARY PRESCRIBED FINITE COMMIT LENGTHS

For fixed integers h_1,h_2, set $C_i=A_i^{h_i}$, and let their entries be $C_i=\left(\begin{smallmatrix}p_i & q_i\\ u_i & w_i\end{smallmatrix}\right)$. Then

$$T=p_2p_1+r_2q_2u_1+r_1u_2q_1+r_1r_2w_2w_1, \tag{25}$$

$$D=r_1r_2\beta_1^{h_1}\beta_2^{h_2}. \tag{26}$$

Thus (21) is an exact finite polynomial test. With $r_1 = r_2 = r$, each margin is at most quadratic in r. Find every real root in $[0, 1]$, sort the roots and interval endpoints, and test one point in each component; evaluate endpoints separately. This does not assume that the stable region is a prefix interval. Matrix powers may instead be computed using the polynomial recurrence in Appendix B.

A.3 Why local checks are not sequence guarantees

Let $A = A(1/10, 1/2)$ and $B = A(13/10, 1/2)$. The traces of A, B are $7/5, 1/5$, with determinant $1/2$; both are stable. Squares are stable because their eigenvalues are squared. The products AB, BA have trace $-18/25$ and determinant $1/4$, so both also pass (21). However,

$$\operatorname{tr}(BA^3) = -1489/1250, \qquad \det(BA^3) = 1/16, \tag{27}$$

and its third Jury margin is $-1287/10000$. Repeating that four-step word diverges on a generic initial state. These equalities are checked in exact rational arithmetic. The example is a small specialization of the general distinction between finite-word tests and arbitrary-switching stability; path-complete methods address the broader problem (Ahmadi et al., 2014).

B Complete quantitative margin analysis

B.1 Parameterization of the full margin class

Apply the closed-disk Schur criterion to A/s. With $\beta \geq 0$, its determinant is β/s^2, and its trace is $(1 - a + \beta)/s$. The resulting inequalities are

$$0 \leq \beta \leq s^2, \qquad |1 - a + \beta| \leq s + \beta/s. \tag{28}$$

Solving the second pair for a gives (3)'s stated parameterization. Boundary points are included because $s < 1$; a repeated root at s or $-s$ is still strictly inside the unit disk. The parameter set is compact for every fixed s.

B.2 Powers and their entrywise extremizer

Let λ_1, λ_2 be the eigenvalues, possibly equal or complex. Define

$$u_0 = 0, \quad u_1 = 1, \quad u_h := \sum_{j=0}^{h-1} \lambda_1^{h-1-j} \lambda_2^j \quad (h \geq 1). \tag{29}$$

No division by a root gap is needed. For distinct roots it equals $(\lambda_1^h - \lambda_2^h)/(\lambda_1 - \lambda_2)$. Cayley–Hamilton and induction give

$$A^h = u_h A - \beta u_{h-1} I, \qquad |u_h| \leq h s^{h-1}. \tag{30}$$

Using $u_{h+1} = (1 - a + \beta)u_h - \beta u_{h-1}$, the entries satisfy

$$|(A^h)_{11}| = |u_{h+1} - \beta u_h| \leq (h + 1)s^h + h s^{h+1}, \tag{31}$$

$$|(A^h)_{12}| = |\beta u_h| \leq h s^{h+1}, \tag{32}$$

$$|(A^h)_{21}| = |a u_h| \leq h(1 + s)^2 s^{h-1}, \tag{33}$$

$$|(A^h)_{22}| = |\beta(u_h - u_{h-1})| \leq h s^{h+1} + (h - 1)s^h. \tag{34}$$

For $h = 1$, $u_0 = 0$ makes the final expression exact as well. At $A_- = A((1 + s)^2, s^2)$ both eigenvalues are $-s$ and $u_h = h(-s)^{h-1}$. Opposite signs make every displayed triangle inequality an equality. This proves (12) and (4) simultaneously, including the repeated-root case.

B.3 Proof of Theorem 1

After a full reset, the incoming state is $(e, 0)$, and the next boundary position is $(A^h)_{11}e$. Uniformly over any sequence, $|e_n| \leq D_s(h)^n |e_0|$. The outgoing velocity is bounded by $h(1 + s)^2 s^{h-1} |e_{n-1}|$. For fixed s, h, the finitely many within-block powers are uniformly bounded on the compact parameter class. Hence $D_s(h) < 1$ implies uniform exponential stability of the entire trajectory. If $D_s(h) \geq 1$,

repeatedly use A_- and any nonzero initial position. Its reset-boundary magnitude is multiplied by exactly $D_s(h)$, proving necessity and the exclusion of equality.

For the asymptotic commit length, write $\delta = -\log s$ and $t = \delta h$. At the positive crossing $D_s(h) = 1$,

$$e^t = 1 + \frac{1+s}{\delta} t. \tag{35}$$

As $\delta \downarrow 0$, the relevant crossing has $t \to \infty$. Taking logarithms and then substituting $t/\log(1/\delta) \to 1$ gives

$$t = \log(1/\delta) + \log\log(1/\delta) + \log(1+s) + o(1). \tag{36}$$

The integer rounding changes h by at most one. Since $\delta = \eta + O(\eta^2)$ and $1+s \to 2$, (6) follows. The ratio $D_s(h+1)/D_s(h) = s[1 + (1+s)/(1+h(1+s))]$ decreases with h, so there is no subsequent return above one after the descending crossing.

B.4 Retention envelope and norm upper bound

If $\rho(M_s(h,r)) < 1$, then $|z_n| \le M_s(h,r)^n |z_0|$ entrywise for arbitrary admissible rules and retentions in $[0, r]$. A stable fixed nonnegative matrix has exponentially decaying powers. The same argument applies to a fixed repeating product of two envelopes. For one envelope, put $D = M_{11}$, $B = M_{12}(h,1)$, $C = M_{21}$, and $E = M_{22}(h,1)$. A nonnegative 2×2 matrix $\left(\begin{smallmatrix} D & rB \\ C & rE \end{smallmatrix}\right)$ has radius below one iff $D < 1$, $rE < 1$, and $(1-D)(1-rE) > rBC$. The last strict inequality, together with $D < 1$, implies $rE < 1$ and gives the explicit threshold

$$r < \frac{1-D}{E(1-D) + BC}. \tag{37}$$

A clipped endpoint equal to one is included only if the original strict inequality still holds.

The parameter bounds imply $|1-a| \le 3$, $a \le 4$, and $\beta \le 1$, so $\|A\|_2 \le \|A\|_F \le \sqrt{27}$. Equation (30) gives $\left\|A^h\right\| \le B_s(h)$, with its second summand zero at $h = 1$. Since $\|J(r)\| \le 1$, $B_s(h) < 1$ contracts every allowed block in the same Euclidean norm, even when retention changes arbitrarily from block to block. Taking $h = C\delta^{-1}\log(1/\delta)$ for any fixed $C > 1$ makes $B_s(h) \to 0$. This proves the sufficient order; the sharp constant is established below.

B.5 Competing witnesses and a coarse localization

Define nilpotent matrices by $A_+ = s(I + N_+)$ and $A_- = -s(I + N_-)$. Direct substitution gives $N_+^2 = N_-^2 = 0$, so $A_+^h = s^h(I + hN_+)$ and $A_-^h = (-s)^h(I + hN_-)$. Put $p = s^h$, $H_h = h(1+s)$. Repeating the negative-root block with constant retention requires

$$F_1 = (1-p)(1-rp) - pH_h(1-r) > 0. \tag{38}$$

Indeed, after removing the harmless sign $(-1)^h$, its trace is $p[1 + r + H_h(1-r)]$ and determinant rp^2. This is its $1 - T + D$ Jury inequality.

Alternating the positive- and negative-root blocks requires

$$F_2 = 1 - p^2 T_h + r^2 p^4 > 0, \tag{39}$$

where exact multiplication gives

$$T_h = \left[(1-s^2)(1-r)^2 + 4r\right]h^2 + 2(1-r^2)h + 1 + r^2. \tag{40}$$

The pair determinant is r^2p^4. These trace and determinant identities are included in the symbolic verifier.

For this preliminary asymptotic bound, assume $0 < \delta = -\log s \le 0.1$. Let $t_\delta > 1$ denote the root of $te^{-t} = 2\delta$. Assume first that $h\delta \le 1$. If $r \le 1/2$, the subtracted term in (38) is at least $h/(2e)$, whereas $(1-p)(1-rp) \le 1 - p \le h\delta \le 0.1h$, a contradiction. If $r \ge 1/2$ and $h \ge 3$, then $p^2 T_h \ge 2h^2/e^2 > 2$, whereas $1 + r^2p^4 \le 2$, contradicting (39). For $h = 1, 2$, $p^2 \ge e^{-0.4}$ and $T_h \ge 4.75$ for $r \in [1/2, 1]$, again giving $p^2 T_h > 2$. To see the last bound, drop the nonnegative $(1-s^2)$ term and minimize $4rh^2 + 2(1-r^2)h + 1 + r^2$: it increases in r on this range and in integer $h \ge 1$, attaining 4.75 at $(r,h) = (1/2, 1)$. Therefore any valid schedule has $h\delta > 1$.

For $r \le 1/2$, (38) now implies $ph < 2$. For $r \ge 1/2$, (39) and $T_h \ge 2h^2$ imply $ph < 1$. In either case $t = h\delta > 1$ satisfies $te^{-t} < 2\delta$. Because te^{-t} is strictly decreasing on $(1, \infty)$, $t > t_\delta$, proving the lower bound. The root exists because $2\delta \le 0.2 < e^{-1}$.

More precisely, the lower root obeys

$$t_\delta = \log(1/\delta) + \log\log(1/\delta) - \log 2 + o(1). \tag{41}$$

Together with full reset as an admissible constant-retention design, this brackets the best such design between the same leading logarithm and log-log terms. It does not constrain state-adaptive reset policies or every nonuniform schedule. For the special case $r = 1$, the exact critical-pair radius is

$$\rho(A_-^h A_+^h) = \left[s^h(h + \sqrt{h^2+1})\right]^2, \tag{42}$$

which follows by solving the quadratic with normalized trace $2 + 4h^2$ and determinant one.

The two witnesses suffice for the lower bounds below. No infinite or adversarially growing rule bank is needed.

B.6 Sharp limiting norm and optimal constant retention

We prove Theorem 2 in three steps. All matrix norms in this subsection are fixed-dimensional; compactness therefore turns the subsequence argument into a uniform supremum statement.

Step 1: collapse of the whole class to two directions. Let $s_n \uparrow 1$, let integers $h_n \to \infty$ satisfy $h_n s_n^{h_n} \to c \in (0, \infty)$, and fix $r \in (0, 1)$. Consider any sequence $A_n \in \mathcal{A}_{s_n}$. Its parameters lie in a common compact set, so pass to a subsequence on which A_n and the unordered eigenvalue pair converge. If the limiting eigenvalues are distinct, the divided-difference expression gives

$$|u_{h_n}| \le \frac{2s_n^{h_n}}{|\lambda_{1,n} - \lambda_{2,n}|} \longrightarrow 0,$$

and the same holds for u_{h_n-1}. Equation (30) implies $A_n^{h_n} \to 0$. If the limiting roots coincide at a point strictly inside the unit disk, both roots are eventually bounded by some $\rho_0 < 1$, and $h_n \rho_0^{h_n-1} \to 0$ gives the same conclusion.

A nonzero cluster limit can therefore occur only when both roots tend to $\epsilon \in \{-1, 1\}$. There is no other repeated unit-modulus root for a real second-order matrix. In this case $A_n \to A(0, 1)$ for $\epsilon = 1$, or $A_n \to A(4, 1)$ for $\epsilon = -1$. The coefficients u_{h_n} are real and bounded by $h_n s_n^{h_n-1} \to c$; extract a further subsequence with $u_{h_n} \to t$, $|t| \le c$. The recurrence in the form $u_h = \lambda_1 u_{h-1} + \lambda_2^{h-1}$ gives

$$\epsilon u_{h_n} - \beta_n u_{h_n-1} = \lambda_{1,n}(\epsilon - \lambda_{2,n}) u_{h_n-1} + \epsilon \lambda_{2,n}^{h_n-1} \longrightarrow 0. \tag{43}$$

Here u_{h_n-1} is bounded and $s_n^{h_n-1} \to 0$. Thus every nonzero cluster limit of $A_n^{h_n} J(r)$ is $tN_+ J(r)$ or $-tN_- J(r)$, where

$$N_+ J(r) = \begin{pmatrix} 0 & r \\ 0 & 0 \end{pmatrix}, \qquad N_- J(r) = \begin{pmatrix} 2 & -r \\ 4 & -2r \end{pmatrix}. \tag{44}$$

The minus sign is immaterial for any norm. The repeated endpoint rules attain the magnitudes $|t| = c$ (possibly after a parity subsequence). Hence the cluster bound gives the actual limit of the supremum, not only a one-sided estimate.

Step 2: a common norm attains the competing-mode balance. For S_r in (11), direct multiplication yields

$$S_r^{-1} N_+ J(r) S_r = \begin{pmatrix} 0 & 2\sqrt{r} \\ 0 & 0 \end{pmatrix}, \quad S_r^{-1} N_- J(r) S_r = \begin{pmatrix} 0 & 0 \\ 2\sqrt{r} & 2(1-r) \end{pmatrix}. \tag{45}$$

Their largest absolute column sum is $2\max\{\sqrt{r}, 1 - r\}$. Step 1 and continuity of this fixed norm now prove (11). To see explicitly why the statement is uniform, a failure of the upper bound would supply a sequence of maximizing or near-maximizing matrices. Compactness would give a cluster limit from Step 1 with a larger norm, a contradiction. Endpoint rules supply the lower bound.

The minimum over $r \in (0, 1)$ is attained uniquely at $1 - r = \sqrt{r}$. Writing $\varphi = (1 + \sqrt{5})/2$, this is $r_\star = \varphi^{-2}$, with limiting multiplier $2c/\varphi$. Any fixed $c < \varphi/2$ therefore gives a strict common-norm contraction for all sufficiently large n at $r = r_\star$. Arbitrary products then contract in that norm; bounded within-block powers give full-trajectory stability for each fixed s, h.

Step 3: the optimum and the constant term. The coarse bounds above and full reset localize every optimal interval to

$$\delta h_\star = \log(1/\delta) + \log\log(1/\delta) + O(1).$$

Consequently $c_s = h_\star s^{h_\star}$ stays bounded and bounded away from zero, while $p = s^{h_\star} \to 0$. Along any convergent subsequence $c_s \to c$ and $r_s \to r$, the exact necessary margins (38)–(40) give

$$2c(1-r) \le 1, \qquad 4c^2 r \le 1. \tag{46}$$

For the pair margin, the term $p^2(1-s^2)h^2$ tends to zero, the term $4rp^2h^2$ tends to $4rc^2$, and all remaining terms vanish. The best simultaneous bound is $c \le c_\star = \varphi/2$, with equality only at $r = r_\star$. This includes the endpoint retentions by continuity.

Conversely, choose any fixed $c < c_\star$ and integers $h(s)$ on the descending branch of hs^h such that $h(s)s^{h(s)} \to c$. Such integers exist because $\delta \to 0$, so neighboring points on that branch have asymptotically equal values. Step 2 proves that this interval and $r_\star$ work for all sufficiently small δ. It follows that every subsequential limit of c_s is at least c, since $h_\star \le h(s)$ and hs^h is decreasing when $h\delta > 1$. Letting $c \uparrow c_\star$ and using the necessity bound proves $c_s \to c_\star$. The unique equality case in (46) proves $r_s \to r_\star$ for every sequence attaining the minimum interval.

Finally $c_s = (\delta h_\star)e^{-\delta h_\star}/\delta$. Taking logarithms and using the coarse localization gives

$$\delta h_\star = \log(1/\delta) + \log\log(1/\delta) - \log c_\star + o(1),$$

which is (7). Subtracting it from the full-reset expansion gives the claimed $\log\varphi/\delta$ saving. This result is a static common-interval optimum; it is not a lower bound on feedback restarts or on the checkpoint overhead.

B.7 Exact finite-margin optima: exclusion and verification

There are two logically separate parts to each finite-margin proof. The necessary part excludes *every* r at each smaller integer h. The sufficient part checks *every* matrix in the continuous class at one fixed feasible r. Neither part substitutes a grid for a continuum.

Exact rational exclusion. For $h < h_{\mathrm{reset}}$, F_1 is strictly increasing in r with positive slope $p[H_h - 1 + p]$, and its root is

$$\ell_h = \frac{p(1+H_h)-1}{p(H_h-1+p)} > 0. \tag{47}$$

In all excluded cases it lies in $[0,1]$; a root at or above one would already exclude every retention. Write $T_h = A_0 + A_1 r + A_2 r^2$, where

$$A_0 = (1-s^2)h^2 + 2h + 1, \quad A_1 = 2(1+s^2)h^2, \quad A_2 = (1-s^2)h^2 - 2h + 1.$$

The derivative of F_2 is $-p^2[A_1 + 2(A_2 - p^2)r]$. The bracket is positive at $r = 0$ and at $r = 1$, where it equals $4h^2 - 4h + 2 - 2p^2 > 0$. It is affine, so F_2 is strictly decreasing on $[0,1]$. Thus $F_2(\ell_h) \le 0$ proves that no r can satisfy both strict margins. All inputs are rational for the three reported s, so this sign is checked with integer numerators and denominators, without rounding.

The checker verifies $F_2(\ell_h) \le 0$ for every $h < 93,292,669$, respectively. At the first surviving h, the exact open retention intervals summarized in Table 4 pass the two repeated-block tests and the alternating-pair test. The other negative and pair Jury margins are automatically positive. For the positive block, the affine trace is $p[1 + r - h(1-s)(1-r)]$, whose absolute value is bounded by $p\max\{|1 - h(1-s)|, 2\} < 1$ over the entire interval in each reported case. Its determinant is $rp^2 < 1$. Roots delimiting the upper endpoint of the feasible interval are isolated by rational bisection; displayed endpoints are rounded summaries, not exact proof inputs.

Whole-class sufficient cover. Fix a case (s, h, r, B) using exact decimal-rational inputs, with rounded summaries in Table 4. The norm is $\|B^{-1}z\|_1$. Four copies of the closed unit square cover the entire spectrum class: real roots $(\sigma su, \sigma sv)$, and conjugate roots $\sigma su\exp(\pm i\pi v/2)$, with $u, v \in [0,1]$ and $\sigma \in \{-1, 1\}$. Real roots have the same sign because $\beta \ge 0$; zeros and repeated roots are included. The spectrum uniquely specifies the trace, determinant, and hence A.

Table 4: Witness-feasible intervals and bases, rounded for display. Retention endpoints are shown as percentages. In every case $B_{12} = 1$. Certification uses exact decimal-rational inputs; these rounded summaries must not be used as certificate inputs.

η	Retention interval (%)	B_{11}	B_{21}	B_{22}
0.05	(35.71, 38.84)	0.63	-0.03	2.04
0.02	(37.15, 38.49)	0.62	-0.01	2.02
0.01	(37.62, 38.40)	0.62	-0.01	2.01

For each cell the verifier encloses the matrix power before testing

$$\sup_{(u,v) \text{ in cell}} \|B^{-1}A^h J(r)B\|_1 < 1. \tag{48}$$

To avoid subtracting large correlated interval terms, write

$$A^h = c_0 I + c_1(A - \tfrac{1}{2}\operatorname{tr}(A)I).$$

For real roots, $c_0 = (\lambda_1^h + \lambda_2^h)/2$ and $c_1 = u_h$; the latter is a sum of nonnegative monomials up to a common sign. For conjugate roots $\sigma\rho e^{\pm i\theta}$,

$$c_0 = \sigma^h \rho^h \cos(h\theta), \quad c_1 = \sigma^{h-1}\rho^{h-1} h \frac{\operatorname{sinc}(h\theta)}{\operatorname{sinc}(\theta)}.$$

Here $\operatorname{sinc}(0) = 1$ removes the repeated-root singularity and $0 \le \theta \le \pi/2$ keeps the denominator positive. Conjugating the matrix factors before multiplying by c_0, c_1 further reduces dependency overestimation. Ordinary interval enclosure then supplies a rigorous upper bound on each absolute column sum.

Cells are generated by bisecting the longer side (first coordinate on ties). Each binary leaf path therefore specifies exact dyadic endpoints. For each of the four squares, the supplied paths are prefix-free and their exact area sum is one. This certifies complete coverage. There are $172, 125, 110$ leaves in the three cases; the largest verified norm bound in each case is below $1 - 10^{-3}$. Because one fixed norm contracts every matrix, all switching sequences contract. Compactness also bounds the finitely many internal powers, completing the sufficiency argument. Matching the exact exclusions proves the finite optimal intervals.

Independent arithmetic replay. The first checker uses outward interval arithmetic; the second recomputes every enclosure using exact rational endpoints. Sine and cosine endpoint values are enclosed by their Taylor polynomials with Lagrange remainder of orders 39 and 38, respectively, after range reduction. The range-reduction error is included using the unit Lipschitz bound. A rational enclosure of π is checked independently with Machin's identity and alternating arctangent remainders. All possible trigonometric extrema inside each argument interval are included. The second checker uses no stored upper bounds from the first and no floating-point sign decision. Floating-point searches found the candidate bases, but only the reproduced rational inequalities are used as proof.

B.8 The marginal impossibility as a limiting statement

For completeness, fix $r \in [0, 1]$ and positive integers h_1, h_2. At the boundary pair $A(0, 1), A(4, 1)$, the absolute trace of $A(4, 1)^{h_2} J(r) A(0, 1)^{h_1} J(r)$ exceeds $1 + r^2$ by

$$2h_2(1 + 2h_1 r - r^2) > 0. \tag{49}$$

Perturb to $A(\epsilon, 1 - \epsilon)$ and $A(4 - 3\epsilon, 1 - \epsilon)$. Both frozen rules are strictly stable for small positive ϵ, while continuity preserves the product's strict instability. This explains why no finite rule-independent schedule covers the entire frozen-stable class. It is a nonuniform-stability example, consistent with classical dwell-time theory (Liberzon, 2003), not the principal theoretical advance.

C LINEAR DIAGNOSTIC CERTIFICATES AND MEMORY-REPAIR LIMITS

C.1 TWO PROBES, THE INTERVAL, AND THE IDENTICAL-INFORMATION ABLATION

For $C = A^h$, compute $p = C(e, 0)$ and $s = C(0, v)$. A fixed quadratic budget gives

$$a_r r^2 + 2b_r r + c_r \le 0, \qquad a_r = s^T P s \ge 0, \quad b_r = p^T P s, \quad c_r = p^T P p - b. \tag{50}$$

Intersect its sublevel with $[0, 1]$. If $a_r = 0$, solve the linear or constant inequality; otherwise use its two roots when real, preserving a possible positive lower endpoint. For endpoint quadratic loss, put $a_f = s_e^T H s_e$ and $b_f = p_e^T H s_e$. If $a_f > 0$, the minimizer on an interval $[l, u]$ is $\mathrm{clip}_{[l,u]}(-b_f/a_f)$; otherwise compare endpoints, or take the declared lower endpoint when the loss is constant. The two probes are simply superposition for a fixed linear map. Normalization, clipping, and general nonlinear update blocks need not preserve this identity.

For the ablation, the unconstrained method minimizes the identical endpoint loss over $[0, 1]$, then compares with the identical fallback. The certified interval is a subset, so at the same incoming state the unconstrained endpoint loss is no larger, including the fallback option. This argument says nothing about which closed-loop trajectory subsequently has lower loss. A test enumerates random same-state comparisons; the actual closed-loop experiment reports both equality frequency and the signed effect.

C.2 THE EXACT-GEOMETRY FALLBACK AND ITS INFORMATIONAL STATUS

The diagnostic metric is $P = \mathrm{diag}(H, H)$. Let F be gradient descent with step $1/(2L)$, discarding incoming velocity. In a Hessian eigenmode λ its squared P-energy factor on the position part is $(1 - \lambda/(2L))^2 + (\lambda/(2L))^2$, whose maximum on $[\mu, L]$ is

$$q_0 = 1 - \mu/L + \mu^2/(2L^2). \tag{51}$$

Thus $F^T P F \preceq q_0 P$ and F^h contracts by q_0^h. The diagnostic guard uses $q = (1+q_0)/2$ and compares true endpoint loss with $F^h z$.

Knowing H and g makes $e = H^{-1} g$ observable, but it also gives $x_\star = x - H^{-1} g$. For the small dense matrices tested here, solving is not an expensive unavailable operation relative to the certificate. We do not defend an optimization-speed interpretation of this benchmark. Its purpose is controlled causal accounting of the retention search versus the additional energy constraint. No comparison in this section establishes an advantage over a direct solve, or a reason to use this guard on a known quadratic.

C.3 ALL-STATE VERSUS STATEWISE CERTIFICATION

Write $Q = C^T P C$ and partition it into position and memory blocks. For $P_c = \mathrm{diag}(H, cH)$, the all-state condition is $J(r)^T Q J(r) \preceq q^h P_c$. The largest eigenvalue of

$$P_c^{-1/2} J(r)^T Q J(r) P_c^{-1/2} \tag{52}$$

must be at most q^h. If a positive r is feasible, zero is feasible: apply the same inequality only to inputs with zero incoming velocity to get $Q_{ee} \preceq q^h H$, which is exactly the zero-retention condition. The set is convex in r, because for every fixed z the left energy is a convex quadratic; its intersection over all z is convex. Hence the all-state set is a prefix interval. A statewise interval can exclude zero. This elementary fact concerns this block-diagonal metric and scalar transfer class; it is not an impossibility for more general common certificates.

C.4 SHARP UNRESTRICTED MEMORY REPLACEMENT

For a positive-definite block matrix Q, define $\mathcal{S}(Q) = Q_{ee} - Q_{ev} Q_{vv}^{-1} Q_{ve}$. Completing the square gives

$$(e, v')^T Q (e, v') = e^T \mathcal{S}(Q) e + (v' + Q_{vv}^{-1} Q_{ve} e)^T Q_{vv} (v' + Q_{vv}^{-1} Q_{ve} e). \tag{53}$$

Therefore *any* memory replacement preserving e can satisfy budget b exactly when $e^T \mathcal{S}(Q) e \le b$. If replacement may depend arbitrarily on (e, v), the sharp worst-case ratio relative to an incoming

positive-definite metric P is

$$\inf_T \sup_{z \neq 0} \frac{T(z)^T Q T(z)}{z^T P z} = \lambda_{\max}(\mathcal{S}(Q), \mathcal{S}(P)), \tag{54}$$

where $T(e, v)$ must preserve e. For the upper bound, choose $v' = -Q_{vv}^{-1} Q_{ve} e$ and use $z^T P z \geq e^T \mathcal{S}(P) e$. For the lower bound, choose incoming $v = -P_{vv}^{-1} P_{ve} e$ and a maximizing generalized eigenvector e; every outgoing memory has energy at least the floor in (53). Thus nonlinear dependence on incoming memory cannot improve the worst-case optimum. This is a Schur-complement consequence, credited to standard quadratic optimization rather than presented as a new general minimax principle (Boyd & Vandenberghe, 2004).

When $Q = A^T P A$ for one momentum step with invertible momentum coefficient and $P = \mathrm{diag}(H, H)$, outgoing velocity u can be freely chosen. Minimizing $(e+u)^T H (e+u) + u^T H u$ gives $u = -e/2$ and $\mathcal{S}(Q) = H/2$. In particular, a transfer can write an objective-informed half-Newton step into the momentum. Such a comparison has more power than scalar erasure; it is not evidence that preserving old memory is useful.

The minimum-change replacement projects the old memory onto the ellipsoid in (53). Diagonalize $Q_{vv} = U \Lambda U^T$, let $m = -Q_{vv}^{-1} Q_{ve} e$ and $b' = b - e^T \mathcal{S}(Q) e$. If $b' < 0$, no repair exists. If $b' = 0$, the feasible set is the singleton $\{m\}$: return m directly. If the incoming memory differs from m, this case corresponds to a multiplier tending to infinity, not a finite root. For $b' > 0$, an already feasible memory is unchanged. Otherwise the projection is $m + U(I + t\Lambda)^{-1} U^T (v - m)$, where $t > 0$ is the unique finite root making the ellipsoid energy equal b'. Strict monotonicity of the residual from its positive initial value to zero supplies a bracketed solver. Unit tests cover negative, zero, and positive residual budgets, including nonzero incoming displacement from the center.

Infeasible repair instances. We reconstructed all 3,000 diagnostic cases with seed 622 and the declared dimension, step, momentum, and commit distributions. All feasibility labels agree with $e^T \mathcal{S}(Q) e \leq b$. For the eleven infeasible cases, the minimum energy exceeds budget by 7.45% to 42.34%. They are not failed projections, missing optimizations, or floating-point boundary cases.

D Proofs for observable nonlinear certificates

Assume f is twice differentiable on each tested segment, $\mu I \preceq \nabla^2 f \preceq L I$, and $\left\| \nabla^2 f(x) - \widehat{H} \right\| \leq \epsilon$. With an M-Lipschitz Hessian, integration gives

$$\nabla f(x + d) = g + \nabla^2 f(x) d + R, \tag{55}$$

$$\|R\| \leq \int_0^1 M t \|d\|^2 \, dt = \tfrac{M}{2} \|d\|^2 . \tag{56}$$

Triangle inequality gives (19). This is the usual model-remainder technique underlying curvature-controlled methods (Nesterov & Polyak, 2006; Ben-Tal & Nemirovski, 1998). The expression before squaring is nonnegative and convex in affine $d(r)$; composing it with the increasing convex square on $[0, \infty)$ preserves convexity. Adding $\tau \|d(r)\|^2$ therefore preserves a convex sublevel problem in r.

D.1 What each sharpness statement actually covers

For a quadratic and the *full* symmetric ball $\{\Delta = \Delta^T : \|\Delta\| \leq \epsilon\}$,

$$\sup_{\Delta} \left\| g + (\widehat{H} + \Delta) d \right\| = \left\| g + \widehat{H} d \right\| + \epsilon \|d\| . \tag{57}$$

For nonzero d and nominal outgoing vector, let $u = d / \|d\|$ and $w = (g + \widehat{H} d) / \left\| g + \widehat{H} d \right\|$. A Householder reflection $R = I - 2aa^T$, with $a = (u - w)/\|u - w\|$, sends u to w. If $u = w$, use $R = I$; if the nominal vector is zero, any unit target suffices. Then $\Delta = \epsilon R$ is symmetric, has norm ϵ, and attains the bound. Zero displacements or zero radii are immediate. The attaining matrix need not obey $\mu I \preceq \widehat{H} + \Delta \preceq L I$. Accordingly this sharpness claim alone cannot justify using the larger ball when an interval is also known.

For the interval itself, every symmetric H with $\mu I \preceq H \preceq LI$ has $H = cI + bR$ with $\|R\| \le 1$, where $c = (L+\mu)/2$, $b = (L-\mu)/2$. The same reflection construction now gives

$$\sup_{\mu I \preceq H \preceq LI} \|g + Hd\| = \|g + cd\| + b\,\|d\|\,. \tag{58}$$

Unlike the previous witness, the attaining $cI + bR$ is guaranteed to lie in the interval. This proves the quadratic sharpness of (20). For nonlinear f, the average Hessian $\int_0^1 \nabla^2 f(x+td)dt$ also lies in this interval, so the bound remains valid without a Hessian-remainder term.

The cap follows from $\mu I - \widehat{H} \preceq H - \widehat{H} \preceq LI - \widehat{H}$. Both extreme eigenvalues of the difference are bounded in absolute value by $\max\{L - \lambda_{\min}(\widehat{H}), \lambda_{\max}(\widehat{H}) - \mu\}$. A blind $L - \mu$ cap is justified only when the estimate is also in the interval; our implementation uses the more general expression because controlled noisy estimates can leave it.

D.2 Combining available upper bounds

Let $I_{\mathrm{loc}}, I_{\mathrm{sp}}$ be the two feasible sublevels in $[0,1]$. The safe set for our combined sufficient test is $I_{\mathrm{loc}} \cup I_{\mathrm{sp}}$. Each is an interval, but their union need not be. We minimize the same nominal quadratic loss independently on each component and compare the candidate minima, then use the common true-loss fallback comparison. Filling the gap between two disjoint components would be incorrect. Both bounds cover supersets of the actual joint uncertainty set; using their minimum can tighten a valid certificate without being the exact optimization over the intersection. These bounds do not solve the optimization problem over the intersection.

D.3 Fallback contraction and the limited strength of the one-step guard

For $x_F = x - g/L$, the mean-value formula gives $g_F = (I - \bar{H}/L)g$ with $\mu I \preceq \bar{H} \preceq LI$, hence $\|g_F\| \le (1-\mu/L)\,\|g\|$. With $v_F = -g/L$ and $\tau = \mu L/2$,

$$W_F \le [(1-\mu/L)^2 + \mu/(2L)]\,\|g\|^2 \le q_0 W. \tag{59}$$

A certified accepted candidate has $W^+ \le qW$, while a rejected one uses this fallback. Multiplying the inequalities proves contraction path by path, including history-dependent proposals. Strong convexity gives $f(x) - f_\star \le \|g\|^2/(2\mu)$, so the energy implies convergence. Adding the true-loss comparison gives the usual gradient-descent objective bound. These are standard safeguard deductions (Prémont-Schwarz et al., 2022; Heaton et al., 2023).

The displayed worst-case exponent is gradient-descent-like. It is not a certificate of optimal heavy-ball or Nesterov acceleration. Enforcing a one-step true-loss fallback comparison can remove nonmonotone accelerated steps, but it does not prove that every accepted trajectory is slow: the one-step comparator in Table 2 is empirically strong. The separate checkpoint result is needed for a uniform accelerated exponent while explicitly allowing those transients.

E Proof and implementation of accelerated checkpoints

On a fixed quadratic, diagonalize H for analysis only. In mode λ, the state (g, Lv) follows (13). This matrix is similar to $A(\alpha_\star\lambda, \beta_\star)$, so its determinant is s^2 and its trace lies in $[-2s, 2s]$. Its eigenvalues therefore have modulus s, including repeated endpoints. The entries obey $|1 - \alpha_\star\lambda| \le 3$, $\beta_\star\lambda/L \le 1$, $\alpha_\star L \le 4$, and $\beta_\star \le 1$, giving the norm bound $\sqrt{27}$. The Cayley–Hamilton argument is invariant to this change of coordinates, so

$$\left\|B_\lambda^h\right\| \le B_s(h) \le \bar{s}^h. \tag{60}$$

The mode basis is orthogonal in both g and Lv, making the full operator norm the maximum of its modal norms. Thus the fixed fallback block contracts W by at most $\bar{s}^{2h} = q^h$ for every admissible Hessian. No matrix decomposition is used by the algorithm.

A saved incoming state includes x, v, g. A proposed block is speculative until its outgoing gradient passes the energy test. On rejection, discard the entire trial and reuse the saved g to begin the fixed heavy-ball fallback. This proves the claimed block contraction, including arbitrarily bad or

history-dependent trial rules. A proposal using more than one gradient per internal step must be charged accordingly; the stated cost class permits at most h gradient calls per trial.

To see that a finite h exists, divide the norm bound by $\bar{s}^h$:

$$\frac{B_s(h)}{\bar{s}^h} = \left(h\sqrt{27}/s + h - 1\right)(s/\bar{s})^h. \tag{61}$$

The exponentially decaying factor beats the linear prefactor. As $\kappa \to \infty$, $1 - s = \Theta(\kappa^{-1/2})$, so the least sufficient h is $O(\sqrt{\kappa}\log\kappa)$. Also $-\log q = \Theta(\kappa^{-1/2})$. With $N = \lceil \log(W_0/\varepsilon)/(h[-\log q]) \rceil$ blocks, total gradient calls are at most $1 + 2hN$, yielding (15). This bound accounts for the initial gradient and for a rejected trial at every block. It does not rely on a probability of successful proposals.

Observed-target stopping. After any gradient is evaluated, the full observable state and hence W are available. If $W \le \varepsilon$, returning this state already solves the prescribed task. Checking this condition needs no additional query. It may occur during a speculative block or a fallback, before the next checkpoint. The per-checkpoint contraction claim concerns only completed checkpoints; it need not be extended to this unfinished final block. This modification never increases the worst-case gradient budget above $1 + 2hN$. The extra $O(h)$ term is a sufficient allowance for transients and rounding, not a necessity result derived from the static restart lower bound.

Algorithm 1 Observed-target stopping with gradient-only checkpoints

Require: Gradient callable, valid spectral bounds $0 < \mu < L$, target $\varepsilon > 0$
Compute $\alpha_\star, \beta_\star, q, h$; evaluate initial g and $W = \|g\|^2 + L^2\|v\|^2$
while $W > \varepsilon$ **do**
    Save (x, v, g, W); run a proposed block of at most h gradient queries
    After each evaluated gradient, return immediately if $\|g\|^2 + L^2\|v\|^2 \le \varepsilon$
    if the completed trial satisfies $W^+ \le q^h W$ **then**
        Keep its output and continue
    else
        Restore the saved state and run the fixed Polyak fallback for h steps
        Apply the same observed-target stopping test after each fallback gradient
        Keep the completed fallback output
    end if
end while
Return state and all query counters

If the trial rule and fallback are identical, the proved fallback bound ensures that a completed trial is accepted. Rerunning the same rejected block is therefore neither necessary nor part of the implementation. A failed contract for this identical case raises an implementation error. General failed trials are discarded with their query counts preserved. An observed-target return is recorded separately from a completed accepted or rejected checkpoint.

E.1 Supplied spectral bounds and the direction of error

The proof requires an enclosing interval, not the exact extreme eigenvalues. Write supplied values as $\widehat{\mu}, \widehat{L}$ for this paragraph. If

$$0 < \widehat{\mu} \le \lambda_{\min}(H) \le \lambda_{\max}(H) \le \widehat{L},$$

all modal and norm bounds hold with the supplied values and condition number $\widehat{L}/\widehat{\mu}$. Decreasing the lower bound or increasing the upper bound preserves the guarantee but may lengthen the certified block.

Inward errors—overestimating μ or underestimating L—remove the premise and can invalidate the claimed contraction. They do not necessarily cause divergence. With parameters tuned to $[\widehat{\mu}, \widehat{L}]$, the exact scalar frozen-stability condition is

$$0 < \lambda < \frac{2(1 + \beta_\star)}{\alpha_\star} = \widehat{L} + \widehat{\mu}. \tag{62}$$

Thus overestimating only the lower bound can preserve frozen stability while destroying the proposed accelerated factor. For instance, tuning to $[4, 10]$ and applying the rule at true curvature $\lambda = 1$ gives $h = 5$ and $\|B_1^h\|_2^2/q^h > 111$, although the frozen spectral radius remains below one. Tuning to $[1, 5]$ at true curvature 10 violates (62): its spectral radius is approximately 2.62. The checker reports a block-factor ratio above 10^{10}. Outward bounds $[0.5, 20]$ cover both actual curvatures $1, 10$ and pass the same test. Regression tests compute these examples independently from the modal matrix.

The fallback proof is for a fixed quadratic. Even exact global strong-convexity and smoothness constants do not make this particular Polyak rule accelerated on all nonlinear objectives (Lessard et al., 2016). A nonlinear checkpoint extension needs an accelerated fallback and a contracting Lyapunov quantity computable from the permitted observations. An objective-gap potential involving an unavailable $f_\star$ or $x_\star$ does not meet a gradient-only interface by itself. Integral quadratic constraint (IQC) or tailored potential analyses can supply richer certificates (Lessard et al., 2016; Wilson et al., 2021), but replacing the fixed Hessian by a local one is not such an extension (Sun et al., 2019).

F Further related work

The analytical target is the full scalar class at a prescribed spectral radius, with an exact reset extremizer and a quantitative necessary transient cost for constant retention. Le and Teel already address feedback resets/damping with LMI-based exponential guarantees and uncertainty (Le & Teel, 2021); our result is not a substitute for their broader controller analysis. Common metrics and path-dependent certificates can improve the conservative comparison envelope (Boyd et al., 1994; Ahmadi et al., 2014). The empirical diagonal sweep is not an optimized LMI benchmark.

Heavy-ball momentum originates in Polyak (1964); continuous-time, variational, and Lyapunov analyses explain why transients need not be monotone in objective or gradient norm (Su et al., 2016; Wibisono et al., 2016; Wilson et al., 2021). IQC analysis already treats algorithm dynamics through a richer certificate class (Lessard et al., 2016). Our checkpoint proof uses classical heavy-ball parameters and a transparent matrix-power bound, adding a gradient-only acceptance implementation and explicit accounting for rollback, not a new accelerated optimizer.

Adaptive gradient (Adagrad) and Adam maintain adaptive statistics with coordinate-dependent meanings (Duchi et al., 2011; Kingma & Ba, 2015). Scaling such a statistic is not equivalent to scaling heavy-ball velocity. Switching from Adam to SGD (SWATS) directly studies this optimizer change (Keskar & Socher, 2017); warm-starting examines reuse across training stages (Ash & Adams, 2020); reset studies in deep reinforcement learning (RL) and spike-aware Adam with momentum reset (SPAM) explore practical neural-training settings (Asadi et al., 2023; Huang et al., 2025). These are motivations and missing validation domains, not applications certified by our theorems. A second-moment accumulator would require a positive-state model, a specification of bias correction and normalization, and a new transfer semantics. No formal extension to that state is claimed.

Learned update rules introduce other state dynamics (Andrychowicz et al., 2016; Wichrowska et al., 2017). Their training pathologies, robustness, and memory–compute tradeoffs are studied directly in Metz et al. (2019); Harrison et al. (2022); Metz et al. (2022a); scaling learned optimization is explored in Metz et al. (2022b). Work on learned optimizers training themselves and self-referential meta-learning uses more complex mechanisms than our two-scale search (Metz et al., 2020; 2021; Kirsch & Schmidhuber, 2022). The formal results concern heavy-ball switching only. The two-scale self-adaptation diagnostic in Section 7.3 is not evidence about those mechanisms.

Finally, a velocity–gradient restart check of the form $v^T g > 0$ belongs to the gradient-direction restart family of O'Donoghue & Candès (2015), although the iterate indexing and update rule differ from their accelerated-gradient scheme. The relationship is a direction-based restart criterion; the indexing is specified in the implementation. We distinguish reproducing that restart idea from reproducing the complete analyses or algorithms of that paper or of Le & Teel (2021). More generally, accepting certified candidates and falling back otherwise is standard safeguard logic (Prémont-Schwarz et al., 2022; Heaton et al., 2023); robust convex bounds and projection calculations use established tools (Ben-Tal & Nemirovski, 1998; Boyd & Vandenberghe, 2004). The main text reserves theorem-level emphasis for the explicit margin-dependent results and the accelerated checkpoint statement.