How Much Coordinate Coupling Can Cautious Momentum Tolerate?

Anonymous authors
Paper under double-blind review

Abstract

Cautious momentum suppresses coordinate updates unless the new momentum and current gradient have the same nonzero sign, while retaining the updated momentum buffer. We study the raw, unnormalized rule on time-varying quadratic objectives with a common minimizer, diagonal curvature ratio κ, and relative off-diagonal coupling strength ρ. For a fixed tuning and zero initial momentum, we prove that the method reaches relative Euclidean error ε within $O(\sqrt{\kappa}\log(2+\kappa)\log(1/\varepsilon))$ gradient evaluations under coupling $\rho = O(\kappa^{-1/2})$. This bound is independent of dimension and counts evaluations with suppressed updates. It improves on the optimal worst-case bound of $\Theta(\kappa\log(1/\varepsilon))$ gradient evaluations for linear methods with predetermined coefficients. Conversely, for sufficiently large κ and $\rho = \Omega(\kappa^{-1/2})$, we construct a predetermined Hessian sequence for which the same method diverges from a suitable initial point with zero momentum. The convergence and divergence bounds use explicit universal constants that differ by only a factor of about 2.14. Thus we characterize, up to constant factors, the robustness threshold for accelerated convergence of cautious momentum as $\Theta(\kappa^{-1/2})$. The proofs use a piecewise quadratic Lyapunov argument for convergence and an explicit construction for divergence.

1 Introduction

Momentum methods form update directions by combining past gradients (Polyak, 1964; Muehlebach & Jordan, 2021). Recent methods also use the alignment between momentum and the current gradient to modify these updates (Chang & Yuan, 2025; Liang et al., 2026; Cao et al., 2025). Cautious optimization suppresses a coordinate update unless the new momentum and current gradient have the same nonzero sign, while retaining the updated momentum buffer (Liang et al., 2026). We study the raw, unnormalized rule, which does not rescale the accepted updates. Our goal is to understand when this gate preserves accelerated convergence under changing curvature.

The effect of the gate depends on the coordinate system, and related work has emphasized such dependence for adaptive optimization more broadly (Zhang et al., 2025). For example, for a quadratic objective with a positive definite diagonal Hessian, each gradient coordinate has the same sign as the corresponding error, so every accepted update points toward that coordinate's value at the minimizer. Nonzero off-diagonal Hessian entries can reverse a coordinate's gradient sign and cause the gate to accept an update in the opposite direction, as illustrated in Figure 1(a). Such sign reversals can happen under arbitrarily small coupling, even near the minimizer. However, they do not by themselves imply divergence, since later updates may correct the outward steps. The question is whether repeated interactions between coupling and retained momentum can destroy convergence or acceleration.

To study this question, we consider time-varying quadratic objectives with a common unknown minimizer. We write their Hessians as $H_t = D_t + F_t$, where D_t is the diagonal part of H_t, with entries in $[\mu, L]$, and define the diagonal curvature ratio $\kappa = L/\mu$. We measure relative off-diagonal coupling by $\left\|D_t^{-1/2} F_t D_t^{-1/2}\right\|_{\mathrm{op}} \le \rho < 1$. The Hessians may vary arbitrarily over time as long as they satisfy these bounds, but the gate is always applied in the same coordinate system.

Main question. *How much coordinate coupling can raw cautious momentum tolerate while retaining a uniform accelerated convergence rate under arbitrary changes in the Hessian?*

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.

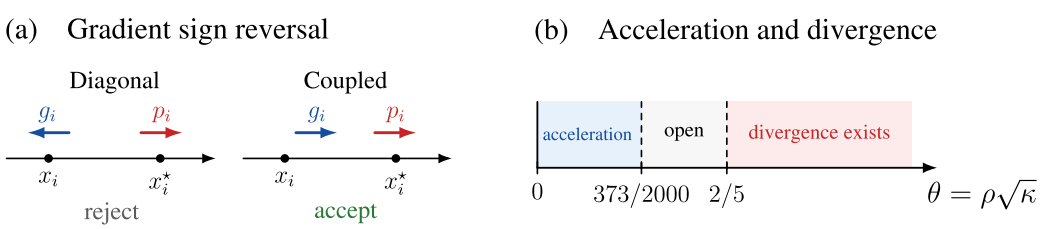


Figure 1: **Effect of coordinate coupling on cautious momentum.** (a) For $x_i < x_i^\star$, positive new momentum $p_i > 0$ proposes a leftward step, away from $x_i^\star$. With a diagonal Hessian, the gradient satisfies $g_i < 0$, so the gate rejects this step. Coupling can make $g_i > 0$, so the outward step is accepted. (b) For the tuning (2.3) and zero initial momentum, Theorem 2.1 guarantees uniform accelerated convergence when $\theta = \rho\sqrt{\kappa} \le 373/2000$. For $\theta \ge 2/5$ and sufficiently large κ, there exists an admissible Hessian sequence fixed in advance for which the method diverges. The behavior between these bounds remains open.

1.1 Our Contributions

Our main result, Theorem 2.1, answers the above main question up to universal constant factors. For the fixed tuning studied in this paper, raw cautious momentum retains accelerated convergence under coupling of order $\kappa^{-1/2}$, while coupling of the same order can also cause divergence; see Figure 1(b).

1. **Uniform accelerated convergence.** For $\rho \le 373/(2000\sqrt{\kappa})$, we prove that the method reaches relative Euclidean error ε within $O(\sqrt{\kappa}\log(2+\kappa)\log(1/\varepsilon))$ gradient evaluations from zero initial momentum. The bound holds uniformly over all finite dimensions and admissible Hessian sequences, and counts evaluations with suppressed coordinate updates.
2. **Divergence under stronger coupling.** For sufficiently large κ and $\rho \ge 2/(5\sqrt{\kappa})$, we construct an admissible Hessian sequence in dimension fourteen, fixed in advance, for which the same method diverges from a suitable initial point with zero momentum. After a finite number of preparation steps, the Hessian sequence is periodic and the error norm grows geometrically.
3. **Acceleration over linear methods.** Cautious momentum improves the dependence on κ from linear to square root, up to a logarithmic factor, compared with linear methods whose coefficients are chosen in advance. We establish this comparison by proving that, for $\kappa \ge 2$, the optimal worst-case bound for this linear class is $\Theta(\kappa\log(1/\varepsilon))$ gradient evaluations, even on diagonal Hessian sequences (Theorem C.3, Appendix C). This class includes gradient descent, heavy-ball momentum, and Nesterov acceleration with parameter schedules chosen in advance.

The constants in the convergence and divergence bounds differ by a factor of $800/373 \approx 2.14$. The behavior between these bounds remains open.

1.2 Related Work

Gradient-dependent momentum rules. Several methods use current gradients to control momentum updates. Liang et al. (2026) introduce cautious optimizers and establish continuous-time descent and, under step-size conditions, one-step loss comparisons with the base optimizer from the same state. Their implementation normalizes the mask. MGUP assigns smaller positive step sizes to unselected coordinates, and the authors also show one-dimensional stochastic nonconvergence of Cautious Adam (Chang & Yuan, 2025). Grams combines gradient signs with momentum-based magnitudes (Cao et al., 2025), while adaptive restart resets momentum through a gradient inner-product test (O'Donoghue & Candès, 2015). For raw cautious momentum, we study when this gradient-dependent gating gives acceleration under time-varying Hessians. For the fixed tuning studied here, we characterize its tolerance to coordinate coupling up to constant factors.

Momentum dynamics and robustness. Control-theoretic analyses study convergence and robustness to perturbations. Integral quadratic constraints accommodate time-varying objectives with a common

minimizer (Lessard et al., 2016). On fixed smooth strongly convex objectives, the Robust Momentum Method trades convergence speed for robustness to relative gradient errors (Cyrus et al., 2018). Muehlebach & Jordan (2021) analyze momentum dynamics, including time-varying algorithm parameters, for fixed objectives. For cautious momentum, off-diagonal coupling acts as a gradient perturbation. We bound its effect using a piecewise quadratic Lyapunov function that tracks both error and momentum.

Acceleration and predetermined linear methods. Acceleration also depends on the objective, the gradient model, and how updates are chosen. Suitably tuned heavy-ball accelerates on strongly convex quadratics (Polyak, 1964), but no fixed tuning guarantees acceleration over all smooth strongly convex objectives (Goujaud et al., 2025). With stochastic gradients, heavy-ball can fail to improve on SGD (Kidambi et al., 2018), and SGD with and without momentum can behave similarly at small, matched effective learning rates (Wang et al., 2024). AGNES, a modified Nesterov method, achieves acceleration under unbiased multiplicative noise (Gupta et al., 2024). Arjevani & Shamir (2016) study limits on acceleration for linear methods with coefficients chosen in advance on fixed objectives. Appendix C establishes the optimal worst-case bound for linear methods with predetermined coefficients under time-varying diagonal curvature. Cautious momentum improves on this bound through its gradient-dependent gate.

2 Coordinate coupling and main results

2.1 Problem setting and algorithm

We consider time-varying quadratic objectives with a common unknown minimizer and bounded coupling between coordinates. To be more specific, fix $0 < \mu \le L$, a finite dimension d, and $\rho \in [0, 1)$. At iteration t, we evaluate $g_t(z) = H_t(z - x^\star)$, which is the gradient of $f_t(z) = \frac{1}{2}(z - x^\star)^\top H_t(z - x^\star)$. The common minimizer $x^\star$ is unknown to the method. We assume that

$$\begin{aligned} H_t = D_t + F_t, \qquad D_t = \operatorname{diag}(d_{t,1}, \dots, d_{t,d}), \qquad \mu \le d_{t,i} \le L, \\ F_t = F_t^\top, \qquad \operatorname{diag} F_t = 0, \qquad \|D_t^{-1/2} F_t D_t^{-1/2}\|_{\mathrm{op}} \le \rho, \qquad \kappa = L/\mu. \end{aligned} \tag{2.1}$$

Here D_t is the diagonal part of H_t, and $\|\cdot\|_{\mathrm{op}}$ is the Euclidean operator norm. The coupling condition implies $(1 - \rho)\mu I \preceq H_t \preceq (1 + \rho)LI$, so every H_t is positive definite.

We call a Hessian sequence admissible if it satisfies (2.1) at every iteration. The Hessians may change arbitrarily, which include choices based on information from previous iterations. A sequence fixed in advance is chosen before the method is executed, possibly using its tuning and initial point.

The raw cautious momentum update is

$$v_{t+1} = \beta v_t + g_t(x_t), \qquad x_{t+1,i} = x_{t,i} - \eta v_{t+1,i} \mathbf{1}\{v_{t+1,i} g_{t,i}(x_t) > 0\}, \qquad v_0 = 0. \tag{2.2}$$

The gate suppresses a coordinate update when the above product is nonpositive, while retaining the updated momentum buffer. We use the raw rule without normalization, momentum resets, or a second-moment state. Each iteration counts as one gradient evaluation, including iterations with suppressed updates. Throughout the main result, the step size and momentum parameter are

$$\eta = \frac{1}{16L}, \qquad \beta = 1 - \frac{1}{4\sqrt{\kappa}}. \tag{2.3}$$

2.2 Main result

The following theorem bounds how much coupling the method tolerates under the tuning (2.3) and zero initial momentum.

Theorem 2.1 (Acceleration and divergence under coordinate coupling)**.** *Under* (2.1)–(2.3)*, there exist universal constants $C_* > 0$ and $\kappa_0 < \infty$ such that the following hold.*

(i) **Uniform accelerated convergence.** *$\forall \kappa \ge 1$ and $\forall 0 \le \rho \le 373/(2000\sqrt{\kappa})$, the iterates satisfy*

$$\|x_t - x^\star\|_2 \le 2 \exp\left[-\frac{t}{C_* \sqrt{\kappa} \log(8\sqrt{\kappa})} \right] \|x_0 - x^\star\|_2 \tag{2.4}$$

for every finite dimension, every admissible Hessian sequence, every initial point with $v_0 = 0$, and every integer $t \geq 0$.

(ii) **Divergence under stronger coupling.** *$\forall \kappa \geq \kappa_0$ and $\forall 2/(5\sqrt{\kappa}) \leq \rho < 1$, there exist a nonzero initial point and an admissible Hessian sequence in dimension fourteen, with $x^\star = 0$ and $v_0 = 0$, for which $\|x_t\|_2 \to \infty$ as $t \to \infty$. The sequence is fixed in advance and eventually periodic, and the norm grows geometrically after a finite number of iterations.*

An explicit value of C_*, which we do not optimize, is given in Appendix A.3.

2.3 Consequences of the main result

The convergence guarantee gives a uniform bound on the number of gradient evaluations. For $0 < \varepsilon \leq 1/2$, let T_ε be the smallest integer $T \geq 0$ such that $\|x_t - x^\star\|_2 \leq \varepsilon \|x_0 - x^\star\|_2$ for every integer $t \geq T$, uniformly over all finite dimensions, admissible Hessian sequences, and initial points $x_0 \neq x^\star$ with $v_0 = 0$. Under the coupling condition in Theorem 2.1(i), (2.4) gives

$$T_\varepsilon = O\big(\sqrt{\kappa}\log(2+\kappa)\log(1/\varepsilon)\big)\,, \tag{2.5}$$

with a universal constant. This count includes evaluations with suppressed coordinate updates. This bound shows acceleration over linear methods whose coefficients are chosen before observing gradient responses. For $\kappa \geq 2$, Theorem C.3 in Appendix C proves their optimal worst-case bound of $\Theta(\kappa \log(1/\varepsilon))$ gradient evaluations on admissible diagonal Hessian sequences, attained by fixed-step gradient descent. Cautious momentum thus improves the dependence on κ from linear to square root, up to a logarithmic factor.

The two parts of Theorem 2.1 also show the coupling scale for this acceleration. Define the normalized coupling parameter

$$\theta = \rho\sqrt{\kappa}. \tag{2.6}$$

Part (i) guarantees accelerated convergence for $\theta \leq 373/2000$, and part (ii) provides a divergent admissible sequence for $\theta \geq 2/5$ when $\kappa \geq \kappa_0$, see Figure 1(b). Thus the coupling threshold for accelerated convergence under the fixed tuning (2.3) is $\Theta(\kappa^{-1/2})$, with the two constants differing by a factor of about 2.14. The exact thresholds for convergence and acceleration, and the behavior between these bounds, remain open. Appendix A.4 gives a more detailed discussion of these thresholds.

3 How coupling affects the dynamics

The diagonal case shows how the gate controls overshoot while retaining momentum. We then explain how coupling can change gate decisions and outline the arguments for convergence and divergence.

3.1 The diagonal case

When $\rho = 0$, we have $g_{t,i}(x_t) = d_{t,i}(x_{t,i} - x_i^\star)$ with $d_{t,i} > 0$, so the gradient and error have the same sign in each coordinate. The gate therefore accepts only updates directed toward $x_i^\star$, although an accepted step may cross it. After a crossing, any proposed update that would continue moving away is rejected. During these rejected updates, the coordinate remains fixed while the momentum buffer continues to update. Movement resumes when the proposed update points back toward $x_i^\star$. A coordinate that reaches $x_i^\star$ stays there.

Under the tuning (2.3) and with $v_0 = 0$, each crossing leaves an error magnitude less than half that at the start of the corresponding run of accepted updates. A small error after a crossing may lead to a long interval of rejected updates. However, the momentum update rule allows us to bound this waiting time in terms of the error and momentum at the start of the waiting interval.

Specifically, Theorem C.1 in Appendix C gives $|x_{t,i} - x_i^\star| \leq 2e^{-\gamma_\kappa t}|x_{0,i} - x_i^\star|$ for every coordinate and every integer $t \geq 0$, where $\gamma_\kappa = \Theta((\sqrt{\kappa}\log(2+\kappa))^{-1})$. Since the factor is uniform across coordinates, the same bound holds in every ℓ_p norm, $1 \leq p \leq \infty$, uniformly over dimension and admissible diagonal Hessian sequences.

3.2 COUPLING CAN REVERSE THE GRADIENT SIGN

Even arbitrarily small coupling can let a gradient coordinate have the opposite sign to its error. We give such an example. For $0 < \rho < 1$, consider the symmetric 2×2 Hessian with $H_{11} = \mu$, $H_{22} = L$, and $H_{12} = H_{21} = \rho\sqrt{\mu L}$. We can verify it is positive definite and has relative coupling norm ρ. For $x - x^\star = (-\chi, 1)$ with $0 < \chi < \theta$, we have $g_1 = \mu(\theta - \chi) > 0$, whereas $x_1 - x_1^\star < 0$. If the new momentum in coordinate 1 is positive, the gate accepts a step away from $x_1^\star$. Scaling the error and momentum by a common positive number shows that this effect can occur arbitrarily close to $x^\star$ for every $\rho \in (0, 1)$.

3.3 PROOF IDEAS

Coupling perturbs the scalar dynamics through the gradient:

$$g_{t,i}(x_t) = d_{t,i}\left(x_{t,i} - x_i^\star + \frac{(F_t(x_t - x^\star))_i}{d_{t,i}}\right).$$

Since F_t has zero diagonal, the last term depends only on the other coordinates and the current Hessian. The coupling term can vary at every iteration. It can reverse the gradient sign and change whether the gate accepts the coordinate update.

Section 4 proves convergence by combining a scalar contraction estimate under bounded perturbations with a piecewise quadratic Lyapunov function of the error and retained momentum. Under the coupling condition in Theorem 2.1(i), the decrease of this function over a block of consecutive iterations bounds the sum of the perturbation terms in the scalar estimates. Combining these bounds gives contraction uniformly over dimension. The block length counts all iterations, including those with rejected updates, and the contraction estimate holds for any momentum at the start of the block.

Section 5 first constructs a scalar trajectory whose error grows from one cycle to the next. The fourteen coordinates follow scaled copies of this trajectory, each starting at a different stage of the cycle. Coupling provides the gradients needed for accepted outward updates, and a finite sequence of diagonal Hessians generates the required momentum from zero. For sufficiently large κ, this gives an admissible Hessian sequence satisfying Theorem 2.1(ii). The sequence is fixed in advance and eventually periodic.

4 PROOF OF ACCELERATED CONVERGENCE

We prove Theorem 2.1(i) by combining two estimates. The first bounds the coupling terms by the decrease of a Lyapunov function. The second gives a scalar contraction bound under bounded perturbations. Together, they give uniform convergence over dimension and Hessians.

We translate $x^\star$ to zero, so x_t denotes the error, and absorb the step size into the momentum and Hessian. Set

$$\begin{gathered} r = 1 - \beta = \frac{1}{4\sqrt{\kappa}}, \qquad y_t = \eta v_t, \qquad K_t = \eta H_t = A_t + J_t, \\ A_t = \operatorname{diag}(a_{t,1}, \ldots, a_{t,d}), \qquad r^2 \le a_{t,i} \le M := \frac{1}{16}. \end{gathered} \tag{4.1}$$

Here $A_t = \eta D_t$, $J_t = \eta F_t$, and $0 < r \le 1/4$. The diagonal curvature bounds follow from $\eta\mu = r^2$ and $\eta L = M$. Scaling preserves the relative coupling bound: $\|A_t^{-1/2} J_t A_t^{-1/2}\|_{\mathrm{op}} \le \rho$. Writing $h_t = K_t x_t$ and $p_t = \beta y_t + h_t$, the update becomes

$$y_{t+1} = p_t, \qquad x_{t+1,i} = x_{t,i} - p_{t,i}\mathbf{1}\{p_{t,i}h_{t,i} > 0\}. \tag{4.2}$$

Since $\rho/r = 4\theta$, the convergence condition in Theorem 2.1(i) is equivalent to $\rho/r \le 373/500$.

4.1 A LYAPUNOV FUNCTION THAT CONTROLS COUPLING

We construct a scalar Lyapunov function to track both the coordinate error and retained momentum:

$$U(x, y) = \begin{cases} x^2 + sxy + by^2, & xy \ge 0, \\ \max\{x^2 - xy, by^2\}, & xy < 0. \end{cases} \tag{4.3}$$

Here $b, s > 0$ are universal constants specified in Appendix A.1. The function satisfies $U(x, 0) = x^2$ and $U(x, y) \geq \max\{x^2, by^2\}$. Thus a bound on U bounds both the error and momentum, and U reduces to the squared error when the momentum is zero.

Write $\Delta U = U(x^+, y^+) - U(x, y)$ for the difference after one scalar iteration. For gradient $h = ax$ with $0 < a \leq M$, Lemma A.1 gives $\Delta U \leq -ax^2 - ry^2$ for both accepted and rejected updates. For the perturbed gradient $h = ax + u$ with $r^2 \leq a \leq M$, Lemma A.3 gives

$$\Delta U \leq -c_0 a x^2 + Q(r)\frac{u^2}{a}. \tag{4.4}$$

Here $c_0 \in (0, 1)$ is universal and $Q(r) = O(r^{-2})$, with explicit choices in Appendix A.1. The bound holds for every state and every u, including perturbations that reverse the gradient sign.

For the coupled system, define $V_t = \sum_i U(x_{t,i}, y_{t,i})$, $B_t = \sum_i a_{t,i} x_{t,i}^2$, and $u_t = J_t x_t$. Summing (4.4) over coordinates gives

$$V_{t+1} - V_t \leq -c_0 B_t + Q(r) \sum_i \frac{u_{t,i}^2}{a_{t,i}}. \tag{4.5}$$

Further, the relative coupling bound controls the perturbation sum:

$$\sum_i \frac{u_{t,i}^2}{a_{t,i}} = \|A_t^{-1/2} J_t x_t\|_2^2 \leq \rho^2 \|A_t^{1/2} x_t\|_2^2 = \rho^2 B_t. \tag{4.6}$$

Together with $Q(r) = O(r^{-2})$, this bounds the perturbation term in (4.5) by a universal constant times $(\rho/r)^2 B_t$. And we can then prove the following lemma.

Lemma 4.1 (Lyapunov decrease)**.** *There is a universal constant $\delta > 0$ such that, if $\rho \leq (373/500)r$, every admissible iteration from every state satisfies $V_{t+1} \leq V_t - \delta B_t$.*

Proof. Combining (4.5) and (4.6) gives $V_{t+1} - V_t \leq -[c_0 - Q(r)\rho^2]B_t$. Under $\rho \leq (373/500)r$, (A.12) gives $c_0 - Q(r)\rho^2 \geq \delta > 0$. □

This estimate makes V_t nonincreasing, but it does not give contraction. We next control both error and momentum over a fixed number of iterations to obtain contraction of V_t.

4.2 Scalar contraction under bounded perturbations

We consider the scalar form of (4.2) with $h_t = a_t(x_t + e_t)$. The perturbation e_t may vary over time and reverse the gradient sign. The next lemma bounds the Lyapunov value after a fixed number of iterations.

Lemma 4.2 (Scalar contraction over a block)**.** *There is a universal constant $C_E > 0$ and, for each $0 < r \leq 1/4$, a positive integer $N(r)$ satisfying*

$$N(r) = O\left(\frac{1}{r}\log\frac{2}{r}\right) = O\left(\sqrt{\kappa}\log(8\sqrt{\kappa})\right). \tag{4.7}$$

For every $E \geq 0$, every initial state (x_0, y_0), and every sequence satisfying $r^2 \leq a_t \leq M$ and $|e_t| \leq E$ for $0 \leq t < N(r)$, the scalar iterates satisfy

$$U(x_{N(r)}, y_{N(r)}) \leq \frac{1}{4}U(x_0, y_0) + C_E E^2. \tag{4.8}$$

Both a_t and e_t can depend on the trajectory. The block length counts all iterations, including those with rejected updates.

Without perturbations, the bound reduces U to at most $1/4$ of its initial value. The term $C_E E^2$ bounds the additional contribution from perturbations.

The proof analyzes intervals during which the momentum has one sign. Bounds on movement and waiting times show that, within the block, both $|x_t|$ and $|y_t|$ fall below a threshold determined by the initial state and E. We then bound their subsequent growth up to iteration $N(r)$. This fixed endpoint allows us to sum the coordinate estimates at the same iteration. Appendix A.2 gives the proof and explicit choices of $N(r)$ and C_E. To use this estimate for the coupled system, we next bound the sum of the perturbation terms across coordinates.

4.3 Combining scalar contraction with Lyapunov decrease

Assume $\rho/r \le 373/500$. Fix a block starting at τ, set $N = N(r)$, and define $E_i = \max_{\tau \le t < \tau+N} |u_{t,i}/a_{t,i}|$. Since $h_{t,i} = a_{t,i}(x_{t,i} + u_{t,i}/a_{t,i})$, applying Lemma 4.2 to each coordinate and summing gives $V_{\tau+N} \le \frac{1}{4} V_\tau + C_E \sum_i E_i^2$.

The maxima defining E_i may occur at different iterations. We bound each E_i^2 by the sum of $(u_{t,i}/a_{t,i})^2$ over the block. Using $a_{t,i} \ge r^2$, (4.6), and Lemma 4.1, we obtain

$$\sum_i E_i^2 \le \sum_{t=\tau}^{\tau+N-1} \sum_i \frac{u_{t,i}^2}{a_{t,i}^2} \le \frac{\rho^2}{r^2} \sum_{t=\tau}^{\tau+N-1} B_t \le \frac{\rho^2}{\delta r^2}(V_\tau - V_{\tau+N}). \tag{4.9}$$

Thus the perturbation terms are bounded by a multiple of the Lyapunov decrease over the same block. Setting $A_\rho = C_E \rho^2/(\delta r^2)$, plugging in and rearranging give

$$V_{\tau+N} \le \frac{1}{4} V_\tau + A_\rho (V_\tau - V_{\tau+N}), \text{ thus } V_{\tau+N} \le q(A_\rho) V_\tau, \tag{4.10}$$

where $q(A) = \frac{A+1/4}{A+1} < 1$ for $A \ge 0$. Since $\rho/r \le 373/500$, we have $A_\rho \le A_* := (C_E/\delta)(373/500)^2$. The function q is increasing, so $q_* := q(A_*) < 1$ is a universal constant satisfying

$$V_{\tau+N(r)} \le q_* V_\tau. \tag{4.11}$$

This bound holds for every admissible Hessian sequence and every state at time τ, with the same q_* in every dimension.

4.4 Convergence at every iteration

For any integer $t \ge 0$, let $k = \lfloor t/N(r) \rfloor$. Applying (4.11) over the k complete blocks and using Lemma 4.1 on the remaining iterations gives $V_t \le V_{kN(r)} \le q_*^k V_0$. Zero initial momentum gives $V_0 = \|x_0\|_2^2$, and $U(x, y) \ge x^2$ implies $\|x_t\|_2^2 \le V_t$. As a result,

$$\|x_t\|_2 \le q_*^{\lfloor t/N(r) \rfloor/2} \|x_0\|_2 \qquad (t \ge 0). \tag{4.12}$$

Since $q_* \ge 1/4$, removing the floor introduces a factor $q_*^{-1/2} \le 2$. Because $q_* < 1$ is universal, combining this estimate with (4.7) and restoring $x^\star$ gives (2.4). Appendix A.3 gives the explicit constant C_*. For $0 < \varepsilon \le 1/2$, completing $O(\log(1/\varepsilon))$ blocks ensures that the relative error is at most ε at every subsequent iteration. Together with (4.7), this gives the bound (2.5) on gradient evaluations.

5 Proof of divergence

We first prescribe a trajectory with growing error and construct admissible Hessians that generate it. We then prepend a finite sequence of diagonal Hessians to reach its starting state from zero momentum. This will prove Theorem 2.1(ii).

We set $x^\star = 0$ and use the scaled variables and update in (4.1)–(4.2). The scaled Hessians will have diagonal entries in $[r^2, M]$ and relative coupling at most $(8/5)r = 2/(5\sqrt{\kappa})$. The construction applies to all sufficiently small positive r, and thus all sufficiently large κ.

5.1 A scalar trajectory with growing error

Recall that x denotes the scalar error, $y = \eta v$ the scaled momentum, and $p = \beta y + h$ the updated scaled momentum. The gate accepts an update exactly when $ph > 0$. Fix the constant $a_\circ \in (0, M)$ specified in Appendix B.1. For all sufficiently small positive r, Lemma B.1 provides an initial state with $x = 1$ and $y < 0$ for which the first fourteen iterations have the following prescribed gradients and gate decisions.

Iterations	Gradient	Behavior
3	$h = a_\circ x$	Rejected ($ph < 0$), x stays at 1.
5	$h = a_\circ x$	Accepted ($ph > 0$), x moves toward 0.
2	$h = r^2 x$	Accepted ($ph > 0$), x crosses 0 on the second iteration.
4	$h = r^3 y = r^2(x + ry - x)$	Accepted ($ph > 0$), x moves away from 0.

During the final four iterations, $x < 0 < y$, so $h = r^3 y > 0$ and $p = (\beta + r^3)y > 0$. The gate thus accepts each update away from 0. Here the gradient has the form $h = a(x + e)$ with $a = r^2$, and coordinate coupling will later provide the perturbation term $ae = r^2(ry - x)$.

Let $z_j(r) = (X_j(r), Y_j(r))^\top$ denote the scalar state after j iterations. The same lemma gives $z_{14}(r) = -g(r)z_0(r)$ with $g(r) > 1$: one cycle reverses both signs and multiplies both magnitudes by $g(r)$. Simultaneous sign reversal and scaling preserve the gate decisions, so repeating the gradient pattern gives $z_{t+14}(r) = -g(r)z_t(r)$ for every integer $t \geq 0$.

5.2 A GROWING TRAJECTORY IN FOURTEEN DIMENSIONS

To realize the prescribed gradients through coordinate coupling, we use fourteen copies of the scalar trajectory, each starting at a different stage of the cycle. At every iteration, some coordinates thus have prescribed gradients of the form $a_\circ x$. Coupling with these coordinates will produce the gradients $r^3 y$ needed by the coordinates moving away from 0 as mentioned in the above section.

Set $\vartheta = \vartheta(r) = g(r)^{1/14} > 1$. We scale the copy starting at stage j by ϑ^{-j} so that the combined Euclidean error norm grows by ϑ at every iteration. The resulting reference error and scaled momentum vectors $\xi_t, \upsilon_t \in \mathbb{R}^{14}$ are given by

$$\xi_{t,j} = \vartheta^{-j} X_{t+j}(r), \qquad \upsilon_{t,j} = \vartheta^{-j} Y_{t+j}(r), \qquad 0 \leq j < 14. \tag{5.1}$$

Let $S(v_0, \ldots, v_{13}) = (v_1, \ldots, v_{13}, -v_0)$. Using $z_{t+14}(r) = -g(r)z_t(r)$ and $\vartheta^{14} = g(r)$, we obtain

$$\xi_{t+1} = \vartheta S \xi_t, \qquad \upsilon_{t+1} = \vartheta S \upsilon_t. \tag{5.2}$$

Since S preserves the Euclidean norm, $\|\xi_{t+1}\|_2 = \vartheta \|\xi_t\|_2$. Thus the reference error norm grows geometrically at every iteration.

5.3 CONSTRUCTING ADMISSIBLE HESSIANS

We now construct Hessians that make the method follow the above vector trajectory. The main challenge is to produce the gradients $r^3 \upsilon_{t,j}$, which require coordinate coupling as mentioned previously. The coupling must also preserve symmetry and the other prescribed gradients.

Let $\mathcal{H}_t$, $\mathcal{N}_t$, and $\mathcal{C}_t$ denote the coordinate index sets assigned gradients $a_\circ \xi_{t,j}$, $r^2 \xi_{t,j}$, and $r^3 \upsilon_{t,j}$ at iteration t, respectively. Fix an iteration and omit the time subscripts on these sets and the vectors. We write $\widehat{K}_t = \widehat{A}_t + \widehat{J}_t$, where $\widehat{A}_t$ is diagonal and $\widehat{J}_t$ is symmetric with zero diagonal. We use diagonal values r^2 on index set $\mathcal{N} \cup \mathcal{C}$ and add coupling only between $\mathcal{H}$ and $\mathcal{C}$.

This already gives the prescribed gradient on $\mathcal{N}$. On $\mathcal{C}$, coupling must provide the difference $r^3 \upsilon_\mathcal{C} - r^2 \xi_\mathcal{C} = r^2 w$, where $w = r\upsilon_\mathcal{C} - \xi_\mathcal{C}$. Set $R = \|\xi_\mathcal{H}\|_2^2$. Since $\mathcal{H}$ is nonempty and $\xi_i \neq 0, \forall i$ by Lemma B.1, we have $R > 0$. To map $\xi_\mathcal{H}$ to $r^2 w$ while preserving symmetry, choose

$$(\widehat{J}_t)_{\mathcal{C}\mathcal{H}} = r^2 \frac{w \xi_\mathcal{H}^\top}{R}, \qquad (\widehat{J}_t)_{\mathcal{H}\mathcal{C}} = (\widehat{J}_t)_{\mathcal{C}\mathcal{H}}^\top, \tag{5.3}$$

with all other blocks zero. The transpose block also contributes $r^2(w^\top \xi_\mathcal{C}/R)\xi_\mathcal{H}$ to the gradient on $\mathcal{H}$. Since this is a common scalar multiple of $\xi_\mathcal{H}$, we cancel it by setting the diagonal entries of $\widehat{A}_t$ on $\mathcal{H}$ to

$$\widehat{a} = a_\circ - r^2 \frac{w^\top \xi_\mathcal{C}}{R}. \tag{5.4}$$

The above constructed Hessian satisfies

$$(\widehat{K}_t \xi_t)_\mathcal{H} = a_\circ \xi_\mathcal{H}, \quad (\widehat{K}_t \xi_t)_\mathcal{N} = r^2 \xi_\mathcal{N}, \quad (\widehat{K}_t \xi_t)_\mathcal{C} = r^3 \upsilon_\mathcal{C}. \tag{5.5}$$

Together with the scalar gate decisions, these identities ensure that an iteration starting at (ξ_t, υ_t) reaches $(\xi_{t+1}, \upsilon_{t+1})$.

We next verify the above Hessians are admissible. Under (5.2), the coordinate groups shift with the vectors, and the common factors ϑ^2 cancel in (5.3) and (5.4). This gives $\widehat{K}_{t+1} = S\widehat{K}_t S^\top$, as proved in (B.10). Since $S^{14} = -I$, the Hessians are fourteen-periodic. Moreover, this transformation preserves the diagonal values up to reordering and preserves the relative coupling norm. It therefore suffices to check both bounds at $t = 0$.

At $t = 0$, the vectors ξ_0 and υ_0 have finite limits as $r \downarrow 0$, by Appendix B.1 and (5.1). Also, $R \geq 1$ because $0 \in \mathcal{H}_0$ and $\xi_{0,0} = 1$. Thus $w^\top \xi_{\mathcal{C}}/R$ remains bounded, and (5.4) gives $\widehat{a} \to a_\circ \in (0, M)$. Since $r^2 \to 0$, we obtain $r^2 \leq \widehat{a} \leq M$ for all sufficiently small positive r. All remaining diagonal entries equal r^2.

For the relative coupling bound, normalizing the block in (5.3) by the diagonal curvatures changes its factor r^2 to $r/\sqrt{\widehat{a}}$. Its rank-one form thus gives

$$\|\widehat{A}_t^{-1/2} \widehat{J}_t \widehat{A}_t^{-1/2}\|_{\mathrm{op}} = r\Gamma(r), \qquad \Gamma(r) = \frac{\|r\upsilon_{\mathcal{C}} - \xi_{\mathcal{C}}\|_2}{\sqrt{\widehat{a}}\,\|\xi_{\mathcal{H}}\|_2}. \tag{5.6}$$

Appendix B.2 shows that Γ extends continuously to $r = 0$ and proves $\Gamma(0) < 8/5$. Hence the relative coupling is less than $(8/5)r = 2/(5\sqrt{\kappa})$ for all sufficiently small positive r.

The relative coupling norm is less than one, so $\widehat{K}_t$ is positive definite. Rescaling by $1/\eta$ restores the diagonal interval $[\mu, L]$ and leaves the relative coupling norm unchanged. The resulting Hessians are therefore admissible whenever $2/(5\sqrt{\kappa}) \leq \rho < 1$. The remaining step is to reach the reference initial state from zero momentum.

5.4 Starting from zero momentum

Now we prepend admissible diagonal Hessians to reach (ξ_0, υ_0) constructed above from zero momentum. The identity $\upsilon_{0,j}/\xi_{0,j} = Y_j(r)/X_j(r)$ and Lemma B.1 ensure that each coordinate target satisfies the hypotheses of Lemma B.2. Applying that lemma to all fourteen coordinates gives a suitable initial point $x_0 \neq 0$ with $\upsilon_0 = 0$ and $(x_{100}, y_{100}) = (\xi_0, \upsilon_0)$.

For $t \geq 0$, set $H_{100+t} = \widehat{K}_t/\eta$. By (5.5) and the scalar gate decisions, induction gives $(x_{100+t}, y_{100+t}) = (\xi_t, \upsilon_t)$. Thus (5.2) yields

$$\|x_{100+t}\|_2 = \vartheta(r)^t \|x_{100}\|_2 \longrightarrow \infty, \qquad x_{100} \neq 0, \qquad t \in \mathbb{N}_0. \tag{5.7}$$

All construction steps hold simultaneously for sufficiently small positive r. Hence there is a universal $\kappa_0 < \infty$ such that the construction is valid whenever $\kappa \geq \kappa_0$ and $2/(5\sqrt{\kappa}) \leq \rho < 1$. The admissible Hessians are fixed in advance and fourteen-periodic after the first 100 iterations, and the corresponding quadratics have common minimizer zero. This proves Theorem 2.1(ii).

6 Conclusion

We characterize the robustness of raw cautious momentum to coordinate coupling on time-varying quadratics with a common minimizer. For the fixed tuning studied in this paper, the convergence and divergence bounds give the robustness scale $\Theta(\kappa^{-1/2})$ for accelerated convergence. The dimension-independent convergence guarantee improves over the optimal worst-case rate for linear methods with predetermined coefficients in this setting.

The divergence construction shows that the sign gate can repeatedly accept updates away from the minimizer. This motivates evaluating how momentum tuning and coordinate coupling affect the stability of practical cautious optimizers, particularly with normalized updates or nonquadratic objectives.

The exact thresholds for convergence and acceleration are still unresolved. Further questions are whether other tunings can retain acceleration under stronger coupling and whether a fixed Hessian can cause divergence at the same coupling scale under the present tuning.

AI USE STATEMENT

We used OpenAI GPT-5.6 during the research and writing of this paper. In the convergence analysis, we used it to explore several candidate choices for the coefficients b and s of the piecewise quadratic Lyapunov function in (4.3). The final coefficients and their verification are given in Appendix A.1. In the divergence analysis, we used GPT-5.6 to help construct the scalar gradient pattern and associated trajectory in Lemma B.1. This trajectory has increasing error magnitude over successive cycles and forms the basis of the fourteen-dimensional divergence construction in Section 5.

GPT-5.6 was also used to draft and edit parts of the paper, including the presentation of proofs. The authors verified all final mathematical claims, constructions, constants, and proofs and take responsibility for the full content of the paper.

REFERENCES

Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 908–916, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/arjevani16.html.

Yang Cao, Xiaoyu Li, and Zhao Song. Grams: Gradient descent with adaptive momentum scaling. In *ICLR 2025 First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2025. URL https://openreview.net/forum?id=GmKQnpQdsc.

Da Chang and Ganzhao Yuan. MGUP: A momentum-gradient alignment update policy for stochastic optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TDFSKAspoQ.

Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pp. 1376–1381, 2018. doi: 10.23919/ACC.2018.8430824.

Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02269-2. URL https://arxiv.org/abs/2307.11291.

Kanan Gupta, Jonathan W. Siegel, and Stephan Wojtowytsch. Nesterov acceleration despite very noisy gradients. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=kHXUb494SY.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJTutzbA-.

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://doi.org/10.1137/15M1009597.

Kaizhao Liang, Lizhang Chen, Bo Liu, and qiang liu. Cautious optimizers: Improving training with one line of code. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zBPZeRjfgu.

Michael Muehlebach and Michael I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22(73):1–50, 2021. URL http://jmlr.org/papers/v22/20-207.html.

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.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

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

Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam requires better rotation dependent assumptions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KD4wgunbhO.

A Details of the accelerated convergence proof

We prove the scalar estimates used in Section 4 to establish Theorem 2.1(i). We first derive the Lyapunov bounds in Lemma 4.1, then prove the scalar contraction bound in Lemma 4.2 for arbitrary initial momentum and bounded shifts of the gradient zero. Appendix A.3 gives explicit convergence constants, and Appendix A.4 derives coupling thresholds from both parts of Theorem 2.1.

Throughout this appendix, $0 < r \le 1/4$, $\beta = 1 - r$, and $M = 1/16$, as in (4.1). With the common minimizer shifted to zero, let x, y, and h denote the scalar error, scaled momentum, and scaled gradient. One scalar call is

$$p = \beta y + h, \qquad y^+ = p, \qquad x^+ = x - p\mathbf{1}\{ph > 0\}.$$

For U defined in (4.3), we write $\Delta U = U(x^+, y^+) - U(x, y)$.

A.1 Scalar Lyapunov estimates

Our goal is to prove the estimate (4.4). Summing this estimate over coordinates and applying the relative coupling bound will then give Lemma 4.1.

We first specify the coefficients of the Lyapunov function U in (4.3). Set

$$b = \frac{120}{23}, \qquad s = \frac{15}{23}. \tag{A.1}$$

These coefficients are chosen to satisfy

$$2 - s - (1 - s + b)M = 1, \qquad 2bM = s. \tag{A.2}$$

Expanding (A.3) in the proof of Lemma A.1 naturally leads to the expressions $2 - s - (1 - s + b)a$ for updates that do not cross zero and $2b\beta a - s$ for updates that do (see cases 1 and 2). To obtain the desired coefficient bounds uniformly over $0 < a \le M$, it suffices to require $2 - s - (1 - s + b)a \ge 1$ and $2ba \le s$. We therefore choose b and s by imposing equality at $a = M$, giving (A.2). The bound $2ba \le s$ is also used in the perturbation estimate of Lemma A.3.

The function U is continuous and homogeneous of degree two, with $U(x, 0) = x^2$ and $U(x, y) \ge \max\{x^2, by^2\}$. Before we proceed, we further write down ΔU explicitly for reader's reference as it will be frequently used in this section. Recall that $p = \beta y + h$. The gate determines whether the new position is $x - p$ or x, while the signs of the new position and momentum determine which branch of U applies. Plugging into (4.3) gives

$$\Delta U = -U(x, y) + \begin{cases} x^2 + (s-2)xp + (1-s+b)p^2, & ph > 0, \quad (x-p)p \ge 0, \\ \max\{x^2 - 3xp + 2p^2, bp^2\}, & ph > 0, \quad (x-p)p < 0, \\ x^2 + sxp + bp^2, & ph \le 0, \quad xp \ge 0, \\ \max\{x^2 - xp, bp^2\}, & ph \le 0, \quad xp < 0. \end{cases} \tag{A.3}$$

The first two cases correspond to accepted updates, and the last two correspond to rejected updates. Now we are ready to present lemmas and proofs.

Lemma A.1 (Decrease of U without perturbations)**.** *For $0 < a \le M$ and $h = ax$, the scalar update satisfies $\Delta U \le -ax^2 - ry^2$.*

Proof. If $x = 0$, then $h = 0$, so the position remains at zero and the momentum becomes βy. Hence $\Delta U = -b(1 - \beta^2)y^2 \le -ry^2$, where we used $b(1 - \beta^2) = br(2 - r) \ge r$. By symmetry under simultaneous sign reversal, it remains to consider $x > 0$. Then $h = ax > 0$, so the update is accepted exactly when $p > 0$.

Case 1: $y \ge 0$. Here $p = \beta y + ax > 0$, so the update is accepted. If $p \le x$, the position does not cross zero. Substituting $p = \beta y + ax$ into the first case of (A.3), and using $\beta y \le (1 - a)x$ together with (A.2), gives

$$\Delta U \le -ax^2 + \big\{\beta[-1 + (1 - s + 2b)a] - s\big\}xy - b(1 - \beta^2)y^2 \le -ax^2 - ry^2. \tag{A.4}$$

The last inequality uses $(1 - s + 2b)a \le 31/46 < 1$ and $b(1 - \beta^2) \ge r$.

If $p > x$, the position crosses zero. The second case of (A.3) gives $U(x^+, y^+) = bp^2$, since the other expression in the maximum is at most $2p^2$ and $b > 2$. The bounds $ba^2 \le 1 - a$ and $2b\beta a \le s$ then give $\Delta U \le -ax^2 - b(1-\beta^2)y^2 \le -ax^2 - ry^2$.

Case 2: $y < 0$ and $p \ge 0$. Write $y = -Y$, where $Y > 0$. Since $p = ax - \beta Y \ge 0$,

$$Y \le \frac{ax}{\beta} \le \frac{x}{12}, \qquad 0 \le p \le ax < x.$$

Thus the position does not cross zero, and $(x^+, y^+) = (x - p, p)$. This formula also holds at $p = 0$, when the gate rejects the update but the displacement is zero.

Since $0 \le p \le ax$, we have $p^2 \le axp$. Applying this bound to the expression for $U(x - p, p)$ in (A.3) gives

$$U(x-p,p) \le x^2 - \big[2 - s - (1 - s + b)a\big]xp \le x^2 - xp,$$

where the last inequality follows from (A.2) and $a \le M$. Before the update, $U(x, -Y) \ge x^2 + xY$. The decrease therefore satisfies

$$-\Delta U = U(x,-Y) - U(x-p,p) \ge xY + xp = ax^2 + rxY \ge ax^2 + rY^2.$$

Here the second equality uses $p = ax - \beta Y$ and $\beta = 1 - r$, while the last inequality uses $Y \le x$.

Case 3: $p < 0$. The update is rejected. Write $y = -Y$ and $p = -Z$, so

$$0 < Z = \beta Y - ax < Y, \qquad Y - Z = rY + ax.$$

The last case of (A.3) gives $U(x^+, y^+) = \max\{x^2 + xZ, bZ^2\}$. It suffices to show that each expression in the maximum is at most $U(x, -Y) - ax^2 - rY^2$.

For the first expression, if $Y \le x$, comparison with $U(x, -Y) \ge x^2 + xY$ gives

$$U(x,-Y) - (x^2 + xZ) \ge x(Y - Z) = ax^2 + rxY \ge ax^2 + rY^2.$$

If $Y > x$, then $x^2 + xZ \le 2Y^2$. Comparison with $U(x, -Y) \ge bY^2$ therefore gives a decrease of at least $(b-2)Y^2$, which is grater than $ax^2 + rY^2$ because $a + r \le 5/16 < b - 2$.

For the second expression, if $Y \ge x/b$, then

$$U(x,-Y) - bZ^2 \ge b(Y^2 - Z^2) = b(rY + ax)(Y + Z) \ge ax^2 + rY^2.$$

If $Y < x/b$, then $bZ^2 < x^2/b$. Comparison with $U(x, -Y) \ge x^2$ gives a decrease greater than $x^2/2$, whereas $ax^2 + rY^2 \le (a + r)x^2 \le 5x^2/16$. Here we used $b > 2$ and $a + r \le 5/16$.

Both expressions satisfy the required bound. This completes the proof. □

We next bound the change in U for an arbitrary gradient h. Set $J = 1 + \frac{9}{4(b-2)} = 503/296$.

Lemma A.2 (Change in U for any gradient)**.** *For every real h, the scalar update satisfies*

$$\Delta U \le -\frac{3b}{4} ry^2 + Jx^2 + \frac{b}{r} h^2. \tag{A.5}$$

Proof. We first consider a rejected update with $xp > 0$. By symmetry, take $x, p > 0$. Rejection implies $h \le 0$, so $y > 0$ and $0 < p \le \beta y$. Thus the position remains fixed, while the momentum decreases without changing sign. By (A.3), $\Delta U = sx(p - y) + b(p^2 - y^2) \le -b(1 - \beta^2)y^2$. Since $1 - \beta^2 \ge r$, this implies (A.5).

For all remaining updates, we show that $U(x^+, y^+) \le bp^2 + Jx^2$. Together with $U(x, y) \ge by^2$, this will reduce the proof to bounding $p^2 - y^2$.

For an accepted update with $(x - p)p \ge 0$, we have $xp \ge p^2 \ge 0$. Since $s < 1$, the first case of (A.3) gives $U(x^+, y^+) \le bp^2 + x^2 - xp \le bp^2 + Jx^2$. For an accepted update with $(x - p)p < 0$, the new value is $\max\{x^2 - 3xp + 2p^2, bp^2\}$. For a rejected update with $xp \le 0$, it is $\max\{x^2 - xp, bp^2\}$; this formula also holds at $xp = 0$. To bound the first expression in each maximum, we complete the square in p:

$$x^2 - 3xp + 2p^2 \le bp^2 + \left(1 + \frac{9}{4(b-2)}\right) x^2 = bp^2 + Jx^2,$$

$$x^2 - xp \le bp^2 + \left(1 + \frac{1}{4b}\right) x^2 \le bp^2 + Jx^2.$$

The other expression in each maximum is bp^2, so the claimed bound holds in both cases.

It remains to estimate the change in squared momentum. Using $p = \beta y + h$ and $2\beta yh \le ry^2 + \beta^2 h^2/r$, we obtain

$$p^2 - y^2 \le -\beta r y^2 + \frac{r+\beta^2}{r} h^2 \le -\frac{3}{4} r y^2 + \frac{1}{r} h^2.$$

The last inequality uses $\beta \ge 3/4$ and $r + \beta^2 = 1 - r + r^2 \le 1$. Combining this estimate with $U(x^+, y^+) \le bp^2 + Jx^2$ and $U(x, y) \ge by^2$ proves (A.5). □

We now combine Lemmas A.1 and A.2 to bound the effect of a perturbation u for the gradient $h = ax + u$. Set $\ell = 2 - s = 31/23$, and $c_0 = 51/52$.

Lemma A.3 (Effect of a gradient perturbation)**.** *For $r^2 \le a \le M$, every real u, and $h = ax + u$, the scalar update satisfies* (4.4)*, namely*

$$\Delta U \le -c_0 a x^2 + Q(r) \frac{u^2}{a}, \qquad Q(r) = \frac{J}{r^2} + 1.$$

Proof. Case 1: $xh \le 0$. In this case, $|u| = |h - ax| = a|x| + |h|$. We can therefore bound both positive terms in Lemma A.2 using u^2/a. Discarding its negative momentum term gives

$$\Delta U + ax^2 \le (J + a)x^2 + \frac{b}{r} h^2 \le \max\{J/a + 1, ba/r\} \frac{u^2}{a} \le Q(r) \frac{u^2}{a}. \tag{A.6}$$

The last inequality uses $r^2 \le a \le M$ and $bMr \le b/64 < J$. Since $c_0 < 1$, this proves (4.4), including the cases $x = 0$ and $h = 0$.

Case 2: $xh > 0$. By symmetry, assume $x, h > 0$, and let $p_0 = \beta y + ax$ be the candidate without perturbation. Both the unperturbed and perturbed updates are accepted exactly when their candidate is positive. Thus (A.3), together with $b > 2$, gives their new Lyapunov values as $\Phi(p_0)$ and $\Phi(p_0 + u)$, where

$$\Phi(p) = \begin{cases} \max\{x^2 - xp, bp^2\}, & p \le 0, \\ x^2 + (s-2)xp + (1-s+b)p^2, & 0 \le p \le x, \\ bp^2, & p \ge x. \end{cases} \tag{A.7}$$

Lemma A.1 controls $\Phi(p_0) - U(x, y)$. It remains to bound the additional change $\Phi(p_0 + u) - \Phi(p_0)$.

The function Φ is continuous and piecewise quadratic. On $0 < p < x$, its derivative lies between $-\ell x$ and $2bp$; outside this interval, the derivative is either $-x$ or $2bp$. Since $\ell > 1$, we have $|\Phi'(p)| \le \max\{\ell x, 2b|p|\}$ wherever the derivative exists. Using $|p| \le \beta|y| + ax + |p - p_0|$ and $2ba \le s < \ell$, we obtain $|\Phi'(p)| \le \ell x + 2b\beta|y| + 2b|p - p_0|$. Integrating between p_0 and $p_0 + u$ gives, for either sign of u,

$$\Phi(p_0 + u) - \Phi(p_0) \le \ell|xu| + 2b\beta|yu| + bu^2. \tag{A.8}$$

We now combine this estimate with the decrease $-ax^2 - ry^2$ from Lemma A.1. The bound $2b\beta|yu| \le ry^2 + \frac{b^2\beta^2}{r} u^2$ uses the negative ry^2 term to control the term involving both y and u. Consequently,

$$\Delta U \le -ax^2 + \ell|xu| + \left(\frac{b^2\beta^2}{r} + b \right) u^2. \tag{A.9}$$

For the remaining term involving x and u, Young's inequality gives $\ell|xu| \le \frac{a}{52} x^2 + 13\ell^2 \frac{u^2}{a}$. Plugging in this bound gives the required decrease $-c_0 a x^2$, since $c_0 = 51/52$.

To finish, we check the coefficient of u^2/a. Since $a \le 1/16$, it suffices to prove

$$\frac{b^2(1-r)^2}{16r} + \frac{b}{16} + 13\ell^2 \le Q(r). \tag{A.10}$$

Let

$$f(r) = Q(r) - \frac{b^2(1-r)^2}{16r} - \frac{b}{16}.$$

Differentiation gives $r^3 f'(r) \le -2J + b^2/64 < 0$ for $0 < r \le 1/4$, so f is decreasing on this interval. Substituting the constants from (A.1) at $r = 1/4$ yields

$$f(r) \ge f(1/4) > 13\ell^2. \tag{A.11}$$

This proves (A.10) and thus (4.4). □

Summing Lemma A.3 over coordinates and applying the relative coupling bound gives (4.5)–(4.6). Since $Q(r)\rho^2 = (J + r^2)(\rho/r)^2$, the bounds $r^2 \le M$ and $\rho/r \le 373/500$ give

$$c_0 - Q(r)\rho^2 \ge \delta, \text{ where } \delta := c_0 - (J + M)\left(\frac{373}{500}\right)^2 > 0. \tag{A.12}$$

Here the last inequality can be proved by directly plugging in the values of c_0, J, and M. This proves Lemma 4.1.

A.2 Scalar contraction with bounded perturbations

We now prove Lemma 4.2, which bounds the scalar Lyapunov value after a fixed block of calls. Here a *block* consists of $N(r)$ consecutive calls, indexed by $t = 0, \dots, N(r) - 1$. It starts from (x_0, y_0) and ends at $(x_{N(r)}, y_{N(r)})$. The length $N(r)$ depends only on r and will be specified in (A.23). A *momentum phase* is a maximal sequence of consecutive calls for which the new momentum $p_t = y_{t+1}$ has the same nonzero sign. A block may contain several phases and may begin or end within a phase. The argument allows arbitrary incoming momentum and bounded, time-varying shifts of the gradient zero. We first establish the bound for the auxiliary function

$$W_4(x, y) = \begin{cases} x^2 + xy + 4y^2, & xy \ge 0, \\ \max\{x^2 - xy, 4y^2\}, & xy < 0. \end{cases} \tag{A.13}$$

This function is continuous and satisfies

$$\max\{x^2, 4y^2\} \le W_4(x, y) \le \frac{9}{2}(x^2 + y^2), \qquad W_4(x, 0) = x^2. \tag{A.14}$$

The lower bound follows directly from the two branches, and the upper bound uses $|xy| \le (x^2+y^2)/2$. The branches agree when $xy = 0$, which gives continuity. Comparing the coefficients with those of U also gives

$$W_4 \le \frac{23}{15}U, \qquad U \le \frac{30}{23}W_4. \tag{A.15}$$

These comparisons will transfer the following bound to the Lyapunov function used in the coupled analysis.

Lemma A.4 (Block contraction for W_4)**.** *Let $N(r)$ be the block length specified in* (A.23) *below. From any initial state, let the scalar update evolve under the actual gate with $h_t = a_t(x_t + e_t)$, $r^2 \le a_t \le M$, and $|e_t| \le E$ for $0 \le t < N(r)$. Then*

$$W_4(x_{N(r)}, y_{N(r)}) \le \frac{1}{8}W_4(x_0, y_0) + 2048E^2. \tag{A.16}$$

Both a_t and e_t may depend on the trajectory.

For the proof, we first bound the position and momentum during each phase and obtain a bound valid from any initial state. We then estimate the number of calls needed to reach the box $|x_t| \le B$, $|y_t| \le B$, with B chosen later. Finally, we control the subsequent evolution up to the fixed endpoint $N(r)$, where the contraction estimate is required.

A.2.1 Bounds during and between momentum phases

We bound the position and momentum during each full phase and derive a contraction estimate for the starting positions of successive phases.

We omit the time index. Let (x, y) denote the position and incoming momentum. Recall $h = a(x+e)$, $p = \beta y + h$, $x^+ = x - p\mathbf{1}\{ph > 0\}$, and $y^+ = p$, where $\beta = 1 - r$, $r^2 \le a \le M = 1/16$, and $|e| \le E$.

A positive momentum phase is a maximal sequence of consecutive calls with $p > 0$. It is full if its first call has incoming momentum $y \le 0$. The initial phase may instead begin with $y_0 > 0$. We handle this initial partial phase in the next part. The bounds within a full phase apply after its first update and before the first subsequent call with nonpositive candidate. The definition and bounds of negative phases follow by symmetry under simultaneous sign reversal.

Let A be the position before the first update of a full positive phase, and put $X = |A|$. Since $y \le 0 < p = \beta y + h$, the first gradient is positive. Thus $A > -E$, and the first update is accepted. During the rest of the phase, calls with $h > 0$ decrease the position, while calls with $h \le 0$ are rejected and satisfy $0 < p \le \beta y$. Hence x is nonincreasing. In particular, $x \le A \le X$, so every positive gradient in the phase is at most $M(X + E)$.

We first relate the movement of the position to the largest momentum attained. Fix a finite prefix of the phase, and let P be its largest output momentum. Every accepted call starts at $x > -E$ and subtracts at most P. Rejected calls leave the position fixed. Therefore $-E - P \le x \le A$ throughout the prefix, and the total accepted displacement is at most $A + E + P \le X + E + P$.

To bound P, we now relate momentum growth to this displacement. On the first call, $y \le 0$ gives $0 < p \le h \le M(X + E)$, and hence $p^2 \le 2M(X + E)p$. On every subsequent accepted call,

$$p^2 - y^2 = 2hp - h^2 - (1 - \beta^2)y^2 \le 2M(X + E)p. \tag{A.17}$$

Rejected calls do not increase the squared momentum. Summing these bounds from the first update to a call where P is attained gives

$$P^2 \le 2M(X + E)\big(\text{total accepted displacement}\big) \le 2M(X + E)(X + E + P).$$

Solving this quadratic inequality and using $M + \sqrt{M^2 + 2M} < 1/2$ gives $P \le (X + E)/2$. As a result,

$$|y| \le \frac{X + E}{2}, \qquad |x| \le \max\{X, 3E\}. \tag{A.18}$$

For the position bound above, we used $-E-P \le x \le X$ and $E+P \le (X+3E)/2 \le \max\{X, 3E\}$. Because the prefix was arbitrary, these estimates hold throughout the phase, even if it does not end within the block.

We next show how these bounds control the start of the next phase. Suppose the positive phase ends, and let P now denote its largest momentum before the terminating call. At that call, $y > 0$ and $p \le 0$ imply $h < 0$. Since $h = a(x + e)$, the position before the call satisfies $x < E$. Combining this with the preceding lower bound gives $-E - P \le x < E$.

We have already shown that $|x| \le E + P$ immediately before the terminating call. This same x is the starting position of the next full phase, whenever one occurs. Indeed, if $p < 0$, the terminating call itself is the first update of the next phase. If $p = 0$, the position remains at x until the next nonzero candidate, which starts a full phase of either sign. Combining this observation with $P \le (X + E)/2$ gives

$$X_{\text{next}} = |x| \le E + P \le X/2 + 3E/2. \tag{A.19}$$

Thus the starting magnitude decreases whenever $X > 3E$. Once a phase starts with $X \le 3E$, every subsequent phase also starts within this bound.

A.2.2 A UNIFORM BOUND FROM ANY INITIAL STATE

We now consider arbitrary incoming momentum and show that, from any initial state (x_0, y_0),

$$\max\{|x_t|, |y_t|\} \le |x_0| + |y_0| + 3E, \qquad 0 \le t \le N(r), \tag{A.20}$$

provided $r^2 \le a_t \le M$ and $|e_t| \le E$ for every $0 \le t < N(r)$.

Write $X_0 = |x_0|$, $Y_0 = |y_0|$, and $G = X_0 + Y_0 + 3E$.

If $y_0 = 0$, the position stays fixed until the first nonzero candidate, which starts a full phase at position magnitude X_0.

If $y_0 \ne 0$, assume $y_0 > 0$ by symmetry and consider the initial positive phase. When $x_0 < -E$, every call with positive candidate is rejected: the position remains fixed and the momentum decreases. Any subsequent full phase starts at position magnitude X_0.

When $x_0 \geq -E$, put $C = x_0 + E \geq 0$. Let P be the largest momentum over a finite prefix of the initial phase, including the initial value Y_0. Throughout this prefix, x is nonincreasing, and each accepted update starts at $x > -E$ and moves by at most P. Hence $-E - P \leq x \leq x_0$, so the total accepted displacement is at most $C + P$.

The same summation as in (A.17), now retaining the initial term Y_0^2, gives $P^2 \leq Y_0^2 + 2MC(C+P)$. Solving this quadratic inequality and again using $M + \sqrt{M^2 + 2M} < 1/2$, we obtain

$$P \leq MC + \sqrt{(M^2 + 2M)C^2 + Y_0^2} \leq Y_0 + C/2 \leq Y_0 + (X_0 + E)/2.$$

In particular, $E + P \leq X_0/2 + Y_0 + 3E/2 \leq G$. Together with $X_0 \leq G$ and the position interval above, this gives $|x| \leq G$ and $|y| \leq P \leq G$. Since the prefix was arbitrary, these bounds hold throughout the initial phase.

In every case, the state stays within G until a full phase begins. If one begins, its initial position magnitude is at most G, as $G \geq 3E$, the map $X \mapsto X/2 + 3E/2$ preserves $[0, G]$. Thus (A.19) and (A.18) bound all later positions by G and momenta by $(G+E)/2 \leq G$. Intervening zero-momentum intervals leave the position fixed. This proves (A.20).

A.2.3 Time to reach the target box or start a new phase

To obtain contraction at the fixed endpoint, we first seek a state with sufficiently small position and momentum. Once such a state is reached, (A.20) can be restarted there to control all remaining calls. We therefore introduce a target box that is small enough for the final Lyapunov estimate, while allowing for the perturbation bound E.

Set

$$R = \max\{\sqrt{W_4(x_0, y_0)}, 128E\}, \qquad B = R/32, \qquad S = 2R = 64B. \tag{A.21}$$

If $R = 0$, the state remains zero and the lemma is immediate. Otherwise, our target is the box $|x|, |y| \leq B$. The choice $R \geq 128E$ ensures that $E \leq B/4$. Thus, if the box is reached, restarting (A.20) bounds both the position and momentum magnitudes at every later call by $2B + 3E \leq 11R/128$. The final part of the proof will use this bound to obtain the required contraction estimate.

We also need a uniform bound before the box is reached. By (A.14), $|x_0| \leq R$ and $|y_0| \leq R/2$. Hence (A.20) gives

$$|x_t|, |y_t| \leq R + R/2 + 3R/128 < S \tag{A.22}$$

for every state index $0 \leq t \leq N(r)$.

We now bound the number of calls needed either to enter the box or to start a subsequent full phase. If the box is not reached, another full phase must start within a bounded number of calls. In the next part, we use (A.19) to show that only a bounded number of full phases are needed before the first update of a phase enters the box. Combining this phase count with the time between successive phases then bounds the total number of calls needed to reach the box.

For states with nonzero momentum, it suffices to consider $y > 0$ by symmetry. Start from a state outside the box with $y > 0$, and stop at the first call for which (x^+, p) lies in the box or $p \leq 0$. By the stopping rule, every call before stopping has $p > 0$. Hence each accepted call decreases x by p, while each rejected call leaves x unchanged. Thus x is nonincreasing and can pass through the right, middle, and left regions below only in that order, possibly skipping a region.

Within each region, call an update internal if its input and output positions both lie in that region, its output momentum is positive, and its output state remains outside the target box. We restart the time index at entry into each region.

Next, we bound the number of calls spent in each region before leaving it or stopping, and then combine these bounds to control the time until the state enters the box or the momentum becomes nonpositive.

Right region: $x > B$. We show that the positive gradient forces the position to leave this region within a bounded number of calls. Since $x > B$, $|e| \leq E$, and $a \geq r^2$, we have

$$h = a(x + e) \geq r^2(B - E) \geq \frac{3}{4} r^2 B,$$

where the last inequality uses $E \le B/4$. Moreover, $y > 0$ gives $p = \beta y + h > 0$, so every call is accepted.

We first bound the momentum after j internal calls. Using local indices, let $y_0 > 0$ denote the momentum upon entry into the region. Unrolling the recurrence, discarding the positive initial contribution, and using $1 - \beta = r$, we obtain

$$y_j = \beta^j y_0 + \sum_{k=0}^{j-1} \beta^{j-1-k} h_k \ge \frac{3}{4} r^2 B \sum_{k=0}^{j-1} \beta^k = \frac{3}{4} rB(1 - \beta^j).$$

To make this lower bound independent of j, set $n_0 := \lceil 1/r \rceil$. For $j \ge n_0$, $\beta^j = (1-r)^j \le e^{-rj} \le e^{-1} < 1/2$, and hence $y_j \ge 3rB/8$. Thus, from the n_0th internal call onward, each call decreases the position by at least $3rB/8$.

Now suppose that $n_A := \lceil 256/r \rceil$ calls were internal. The definitions of n_A and n_0 give $n_A - n_0 + 1 \ge 255/r$, so at least this many calls satisfy the preceding displacement bound. Since every call is accepted, the total decrease in position is

$$\sum_{j=1}^{n_A} y_j \ge (n_A - n_0 + 1)\frac{3rB}{8} \ge \frac{255}{r}\frac{3rB}{8} > 64B = S.$$

However, if all these calls were internal, both the starting and ending positions would lie in (B, S), so their difference would be less than S. This contradiction shows that fewer than n_A calls can be internal to the right region.

Middle region: $-B \le x \le B$. An internal call ends with $|x^+| \le B$ and $p > 0$. Since its output must remain outside the target box, we must have $p > B$. We show that the momentum cannot remain this large while the position stays in the region after a certain number of calls.

Each accepted internal call decreases x by $p > B$. Since x is nonincreasing and the region has width $2B$, there are at most two such calls. Moreover, $p > 0$ on every internal call, so $h > 0$ is equivalent to acceptance. Thus at most two internal calls can contribute a positive gradient to the momentum recurrence. On each such call,

$$0 < h = a(x + e) \le M(B + E) \le \frac{5B}{64},$$

where the last inequality uses $M = 1/16$ and $E \le B/4$.

Each call multiplies the existing momentum by $\beta < 1$ and then adds the gradient. Nonpositive gradients only decrease the momentum further. There are at most two positive gradients during internal calls, each bounded by $5B/64$. Their combined contribution is therefore at most $10B/64$, even if we ignore its decay on later calls.

More precisely, let $y_0 \le S$ be the momentum upon entry and y_j the momentum after j internal calls. Then

$$y_j = \beta^j y_0 + \sum_{k=0}^{j-1} \beta^{j-1-k} h_k \le \beta^j S + \frac{10B}{64}.$$

Here we discard the nonpositive gradient terms and bound the weight of each positive term by one.

The gradient contribution is already less than $B/2$. Thus, to bring the momentum below B, it suffices to reduce the term $\beta^j S$ to at most $B/2$. Since $S = 64B$, this requires $\beta^j \le 1/128$. Using $\beta^j \le e^{-rj}$, we therefore choose $n_M := \lceil r^{-1} \log 128 \rceil$, which gives

$$\beta^{n_M} S \le e^{-r n_M} S \le \frac{S}{128} = \frac{B}{2}.$$

If n_M calls were internal, we would consequently have

$$y_{n_M} \le \frac{B}{2} + \frac{10B}{64} < B.$$

This contradicts the requirement that every internal call in the middle region end with momentum greater than B. Thus fewer than n_M calls can be internal to this region.

Left region: $x < -B$**.** As in the right region, the curvature and perturbation bounds give $h \le -r^2(B-E) < 0$. Thus every call with $p > 0$ is rejected, and the position stays fixed outside the box. Our goal is therefore to bound the time until $p \le 0$.

Using local indices and $y_0 \le S$, the same geometric-sum calculation as in the right region gives

$$y_j \le \beta^j S - r(B-E)(1-\beta^j).$$

The contribution from the entering momentum decays, while the negative gradients continue to reduce the momentum. The above upper bound becomes nonpositive when

$$\beta^j \left(1 + \frac{S}{r(B-E)}\right) \le 1.$$

Since $S = 64B$ and $B - E \ge 3B/4$, we have

$$\frac{S}{r(B-E)} \le \frac{256}{3r} < \frac{128}{r}.$$

It therefore suffices to make $\beta^j \le (1 + 128/r)^{-1}$. We choose $n_H := \lceil r^{-1} \log(1 + 128/r) \rceil$, so that

$$\beta^{n_H} \le e^{-r n_H} \le \frac{1}{1 + 128/r}.$$

If all n_H calls had positive output momentum, the preceding estimate would give $y_{n_H} \le 0$, a contradiction. Thus a call with $p \le 0$ occurs within n_H calls.

Since x is nonincreasing before stopping, each region is visited at most once. Besides its internal calls, each visited region contributes at most one call that leaves the region or meets a stopping condition. Combining the three bounds, within $n_A + n_M + n_H + 3$ calls we either reach the box or obtain $p \le 0$.

If the box has not been reached, the condition $p \le 0$ ends the current positive phase and allows us to continue with a new full phase. Indeed, if $p < 0$, this call changes the momentum from $y > 0$ to $y^+ = p < 0$. Recall that a full negative phase begins with a call whose incoming momentum is nonnegative and whose output momentum is negative. Thus the current call is already the first update of a full negative phase. If $p = 0$, then $x^+ = x$ and $y^+ = 0$. The new momentum already satisfies the target bound, so remaining outside the box requires $|x| > B$. On the next call, $y = 0$ and $h \ne 0$, so $p = h \ne 0$. A call with zero incoming momentum and nonzero output momentum is, by definition, the first call of a full phase.

Thus at most one additional call is needed to pass from $p \le 0$ to the start of a new full phase. Including the three possible exit or stopping calls, we define

$$n_A = \lceil 256/r \rceil, \qquad n_M = \left\lceil \frac{\log 128}{r} \right\rceil, \qquad n_H = \left\lceil \frac{\log(1 + 128/r)}{r} \right\rceil, \tag{A.23}$$
$$D = n_A + n_M + n_H + 4, \qquad N(r) = 10D.$$

The same argument applies to negative momentum by simultaneous sign reversal.

Consequently, from any state of the block outside the target box with nonzero momentum, the trajectory either enters the box or completes the first update of a subsequent full phase within D calls, provided at least D calls remain in the block.

We next combine this time bound with (A.19) to show that the first update of the ninth full phase enters the box, unless an earlier call has already done so. This gives a first visit within $9D$ calls and motivates the choice $N(r) = 10D$.

A.2.4 Reaching the target box after finitely many phases

We now combine the phase-transition bound with the preceding time bound to show that the target box is reached within $9D$ calls.

If the initial state is in the box, the first visit occurs at time zero. Otherwise, if $y_0 \ne 0$, the preceding time bound gives a full phase within D calls unless the box is reached first. If $y_0 = 0$, then

$|x_0| > B > E$, so the first gradient is nonzero. The first call therefore has $p = h \neq 0$ and already starts a full phase.

Let X_j denote the position magnitude before the first update of the jth full phase. By (A.22), $X_1 \leq S = 2R$. Iterating (A.19) gives

$$X_j \leq 2^{-(j-1)} X_1 + 3E\big(1 - 2^{-(j-1)}\big) \leq 2^{-(j-1)} 2R + 3E.$$

Thus each transition halves the contribution from the first starting position, while the accumulated perturbation term remains at most $3E$. Since $2R = 64B$ and $3E \leq 3B/4$, eight transitions reduce the first term to $B/4$. Consequently, the ninth full phase satisfies $X_9 \leq \frac{2R}{256} + 3E \leq \frac{2R}{256} + \frac{3R}{128} = B$.

We next check that its first update enters the box. Suppose this phase is positive as the negative case follows by symmetry. By the definition of a full positive phase, its first call has $y \leq 0 < p = \beta y + h$. Hence $h > 0$, the call is accepted, and $p \leq h$. Since $h = a(x + e) > 0$ and $|x| = X_9 \leq B$, the position before this update satisfies $-E < x \leq B$. Moreover, $0 < p \leq h = a(x + e) \leq M(B + E) \leq \frac{5B}{64} < B$. Since $E + p \leq B/4 + 5B/64 < B$, the updated state satisfies

$$-B < -E - p < x^+ = x - p \leq B, \qquad 0 < y^+ = p < B.$$

Both coordinates therefore lie within the target bounds.

It remains to translate this phase count into a call count. The first full phase starts within D calls. After its first update, the momentum is nonzero, so if the box has not been reached, the time bound applies again and gives the next full phase within another D calls. Inductively, the first update of the jth full phase occurs by jD calls unless the box is reached earlier. For $j \leq 9$, the choice $N(r) = 10D$ ensures that at least D calls remain whenever we apply the time bound. Since the ninth full phase enters the box on its first update, the first visit occurs by $9D \leq N(r)$.

A.2.5 Contraction at the end of the block

We have proved that the box is reached before the block ends. The contraction estimate, however, is required at the fixed time $N(r)$. We therefore control the position and momentum from the first visit until that endpoint, and then convert these bounds into a Lyapunov estimate.

Let $\tau \leq 9D$ be the first visit to the box. Since (A.20) holds from any initial state, we may restart it at (x_τ, y_τ) with the same perturbation bound E. This gives

$$|x_t|, |y_t| \leq |x_\tau| + |y_\tau| + 3E \leq 2B + 3E \leq \frac{11B}{4}, \qquad \tau \leq t \leq N(r).$$

Using the upper bound on W_4 in (A.14), we obtain

$$W_4(x_{N(r)}, y_{N(r)}) \leq \frac{9}{2}\big(x_{N(r)}^2 + y_{N(r)}^2\big) \leq 9(11B/4)^2 < \frac{1}{8}R^2 \leq \frac{1}{8}W_4(x_0, y_0) + 2048E^2.$$

Here we used $B = R/32$ and $R^2 \leq W_4(x_0, y_0) + 128^2 E^2$. This proves Lemma A.4.

To obtain the estimate for U, we use (A.15). With $N = N(r)$, this gives

$$U(x_N, y_N) \leq \frac{30}{23} W_4(x_N, y_N) \leq \frac{30}{23}\left[\frac{1}{8}W_4(x_0, y_0) + 2048E^2\right] \leq \frac{1}{4}U(x_0, y_0) + \frac{61440}{23}E^2.$$

Thus the contraction estimate in Lemma 4.2 holds with $C_E = 61440/23$, for arbitrary incoming momentum.

To complete the proof of Lemma 4.2, it remains to verify the claimed bound on the block length. Each of the three ceilings in (A.23) is at most its argument plus one. Together with the four additional calls in D, these contribute at most 70 to $N(r) = 10D$. Thus $N(r) \leq \frac{10[256 + \log 128 + \log(1 + 128/r)] + 70r}{r} \leq \frac{2680 + 10\log(1/r)}{r}$. For the second inequality, we used $r \leq 1/4$. Thus $N(r) \leq \frac{2048}{r}\log\frac{2}{r}$. This proves (4.7). Plugging in $r = 1/(4\sqrt{\kappa})$ gives the explicit bound used in Section 4.4: $N(r) \leq 8192\sqrt{\kappa}\log(8\sqrt{\kappa})$.

A.3 Explicit constants for accelerated convergence

Here we verify that the resulting constants in Section 4 are uniform over admissible trajectories and compute the constant C_* in (2.4).

Assume $\rho/r \le 373/500$ and use the scaled variables from (4.1), with $x^\star$ translated to zero. On a block of $N = N(r)$ calls starting at τ, the perturbation in coordinate i is

$$e_{t,i} = \frac{u_{t,i}}{a_{t,i}}, \qquad E_i = \max_{\tau \le t < \tau+N} |e_{t,i}|.$$

Lemma 4.2 applies to each coordinate with the momentum inherited at time τ. To control the perturbation terms, we bound each E_i^2 by the sum of $e_{t,i}^2$ over the block. Using (4.6) and summing Lemma 4.1 over the same calls then gives (4.9). This argument allows the coordinate maxima to occur at different times and the Hessians to depend on previous states. Summing the scalar contraction bounds and applying (4.9) therefore gives (4.10).

The coefficient A_ρ in that estimate depends on the coupling through ρ/r. Recall that C_E from Appendix A.2.5 and δ from (A.12) are universal constants. Under $\rho/r \le 373/500$, the coefficient A_ρ in (4.10) therefore satisfies

$$A_\rho = \frac{C_E}{\delta}\left(\frac{\rho}{r}\right)^2 \le \frac{C_E}{\delta}\left(\frac{373}{500}\right)^2 = A_*. \tag{A.24}$$

Since $q(A) = (A + 1/4)/(A + 1)$ is increasing on $A \ge 0$ and takes values in $[1/4, 1)$, the same contraction factor $q_* := q(A_*) < 1$ applies to every block.

Recall that $V_t = \sum_i U(x_{t,i}, y_{t,i})$. The estimates allow arbitrary momentum at the start of each block, so they apply successively along the actual trajectory. Iterating (4.10) and using monotonicity of V_t on the remaining calls gives $V_t \le q_*^{\lfloor t/N(r) \rfloor} V_0$. For zero initial momentum, $V_0 = \|x_0\|_2^2$, while $\|x_t\|_2^2 \le V_t$. Removing the floor using $\lfloor t/N(r) \rfloor \ge t/N(r) - 1$ contributes a factor $q_*^{-1/2} \le 2$ after taking square roots. Moreover, $-\log q_* \ge 1 - q_* = \frac{3}{4(1+A_*)}$. Combining these estimates with $N(r) \le 8192\sqrt{\kappa}\log(8\sqrt{\kappa})$ from Appendix A.2 gives

$$\|x_t\|_2 \le 2\exp\left[-\frac{3t}{8(1+A_*)N(r)}\right]\|x_0\|_2 \le 2\exp\left[-\frac{t}{C_*\sqrt{\kappa}\log(8\sqrt{\kappa})}\right]\|x_0\|_2, \tag{A.25}$$

where

$$C_* = \frac{8 \times 8192}{3}(1 + A_*) \approx 1.1363 \times 10^{11}. \tag{A.26}$$

Restoring $x^\star$ proves (2.4) for every $\kappa \ge 1$ and $\rho \le 373/(2000\sqrt{\kappa})$, uniformly over finite dimensions, admissible Hessian sequences, initial positions with zero momentum, and integer times $t \ge 0$. Momentum is retained across block boundaries throughout the argument. For any integer $k \ge 0$ with $q_*^{k/2} \le \varepsilon$, (4.12) gives relative error at most ε for every $t \ge kN(r)$. Consequently,

$$T_\varepsilon \le N(r)\left\lceil \frac{2\log(1/\varepsilon)}{\log(1/q_*)} \right\rceil, \qquad 0 < \varepsilon \le \tfrac{1}{2}. \tag{A.27}$$

Together with (4.7), this proves (2.5).

A.4 Coupling thresholds for acceleration and convergence

This section makes precise the coupling threshold discussed in Section 2. Under the fixed tuning (2.3), we first consider the uniform rate (2.4), then the requirement that every admissible trajectory converges.

Let $\mathcal{R}_{\kappa,\rho}(t)$ denote the worst relative error at call t: the supremum of

$$\frac{\|x_t - x^\star\|_2}{\|x_0 - x^\star\|_2}$$

over all nonzero initial errors with zero momentum, all finite dimensions, and all admissible Hessian sequences under (2.3). To measure the coupling tolerance for the rate (2.4), fix C_* at the value in (A.26), set

$$\tau_\kappa^* = C_*\sqrt{\kappa}\log(8\sqrt{\kappa}),$$

and define

$$\rho_{\mathrm{acc}}^{(C_*)}(\kappa) = \sup\left\{\rho \in [0,1) : \mathcal{R}_{\kappa,\rho}(t) \le 2e^{-t/\tau_\kappa^*} \text{ for all } t \in \mathbb{N}_0\right\}. \tag{A.28}$$

The superscript records that the multiplier C_* is fixed as ρ varies.

As ρ increases, the admissible class includes all trajectories allowed at smaller coupling bounds. Thus $\mathcal{R}_{\kappa,\rho}(t)$ is nondecreasing in ρ, and a failure of the rate guarantee persists for every larger coupling bound.

The two parts of Theorem 2.1 now bound this radius. The convergence estimate (2.4) holds whenever $\rho \le 373/(2000\sqrt{\kappa})$, giving the lower bound. For $\kappa \ge \kappa_0$, the divergent trajectory at $\rho = 2/(5\sqrt{\kappa})$ violates the exponential estimate. That trajectory remains admissible at every larger coupling bound, giving the upper bound. Therefore, for every $\kappa \ge \kappa_0$,

$$\frac{373}{2000} \le \sqrt{\kappa}\,\rho_{\mathrm{acc}}^{(C_*)}(\kappa) \le \frac{2}{5}. \tag{A.29}$$

In particular, $\rho_{\mathrm{acc}}^{(C_*)}(\kappa) = \Theta(\kappa^{-1/2})$.

Let $\rho_{\mathrm{conv}}(\kappa)$ denote the supremum of $\rho \in [0,1)$ for which every admissible trajectory, in every finite dimension and with zero initial momentum, satisfies $x_t \to x^\star$. This requires convergence for each trajectory without prescribing a rate.

The rate guarantee in (A.28) implies convergence, so $\rho_{\mathrm{acc}}^{(C_*)}(\kappa) \le \rho_{\mathrm{conv}}(\kappa)$. The divergent trajectory in Theorem 2.1(ii) gives the same upper bound for both radii. Together with (A.29), this yields, for every $\kappa \ge \kappa_0$,

$$\frac{373}{2000} \le \sqrt{\kappa}\,\rho_{\mathrm{acc}}^{(C_*)}(\kappa) \le \sqrt{\kappa}\,\rho_{\mathrm{conv}}(\kappa) \le \frac{2}{5}.$$

Thus both radii have order $\kappa^{-1/2}$.

B Details of the divergence construction

We provide the details of the construction in Section 5, which proves Theorem 2.1(ii). Our goal is to construct an admissible Hessian sequence, fixed in advance, for which cautious momentum diverges from a suitable initial point with zero momentum. We first prescribe a growing trajectory, then construct Hessians that generate it and a preparation sequence that reaches its starting state.

Appendix B.1 constructs the fourteen-call scalar pattern. This pattern produces growth and every prescribed update agrees with the actual gate. Some prescribed gradients have the opposite sign to the scalar error, so realizing them requires coordinate coupling. Appendix B.2 constructs symmetric Hessians that generate fourteen scaled copies of this trajectory at different stages of the cycle, and verifies their admissibility and periodicity. The resulting reference trajectory starts with nonzero momentum. To satisfy the initialization required by the theorem, Appendix B.3 constructs exactly 100 diagonal calls that reach its starting state from zero momentum. Finally, Appendix B.4 combines these constructions, verifies that the entire Hessian sequence is fixed in advance, and proves geometric growth at every call after preparation.

Throughout this section, $r = 1/(4\sqrt{\kappa})$, $\beta = 1 - r$, $a_\circ = 31063/500000$, and $M = 1/16$. All parts of the construction hold on one common sufficiently small interval of positive r, giving the universal cutoff κ_0 in Theorem 2.1(ii).

B.1 Constructing the growing scalar trajectory

Our goal is to construct a scalar trajectory that returns to $-g(r)$ times its initial state after fourteen calls, where $g(r) > 1$. The cycle first builds momentum while moving toward zero, then continues to accept updates after crossing zero. We will choose the initial momentum so that these outward updates produce a larger error at the end of the cycle.

The construction must satisfy three requirements: the state must grow over a cycle, every prescribed update must agree with the actual gate, and the intermediate states must satisfy the bounds needed for the later coupling and preparation arguments. We use $a_\circ = 31063/500000 < M$ as an explicit

curvature for which these requirements hold. The strict gap between $a_\circ$ and M also leaves room for the diagonal correction in Appendix B.2.

Recall that x denotes the scalar error and $y = \eta v$ the scaled momentum. We construct the following cycle. The first three calls use $h = a_\circ x$ and retain the position while the negative momentum moves toward zero. The next five calls use the same gradient and are accepted, moving the position toward zero. Two accepted calls with $h = r^2 x$ then bring the position across zero. The final four calls use $h = r^3 y$, which keeps the gradient aligned with the momentum and allows the position to move farther from zero. We will verify these signs along the constructed trajectory.

Under these prescribed gate decisions, the four corresponding maps on $(x, y)^\top$ are

$$\begin{aligned} \mathsf{B}_r &= \begin{pmatrix} 1 & 0 \\ a_\circ & \beta \end{pmatrix}, \qquad \mathsf{A}_r = \begin{pmatrix} 1 - a_\circ & -\beta \\ a_\circ & \beta \end{pmatrix}, \\ \mathsf{R}_r &= \begin{pmatrix} 1 - r^2 & -\beta \\ r^2 & \beta \end{pmatrix}, \qquad \mathsf{L}_r = \begin{pmatrix} 1 & -q \\ 0 & q \end{pmatrix}, \qquad q = \beta + r^3. \end{aligned} \tag{B.1}$$

The maps B_r, A_r, R_r, and L_r describe the four parts of the cycle, respectively. Their composition is

$$\mathsf{P}_r = \mathsf{L}_r^4 \mathsf{R}_r^2 \mathsf{A}_r^5 \mathsf{B}_r^3. \tag{B.2}$$

For an initial state $z_0(r)$, write $z_j(r) = (X_j(r), Y_j(r))^\top$ for the states obtained by applying these maps successively. Thus $z_j(r)$ is the state before call j for $0 \le j < 14$, and $z_{14}(r) = \mathsf{P}_r z_0(r)$. The following lemma establishes growth and bounds the momentum-to-position ratios needed for preparation from zero momentum.

Lemma B.1 (A growing scalar trajectory)**.** *For all sufficiently small positive r, there are continuous choices $h(r) > 0$ and $g(r) > 1.002$ such that*

$$\mathsf{P}_r z_0(r) = -g(r) z_0(r), \qquad z_0(r) = (1, -h(r))^\top. \tag{B.3}$$

Starting from $z_0(r)$, all fourteen prescribed gate decisions agree with the actual update. Moreover, for every $0 \le j < 14$,

$$X_j(r) \neq 0, \qquad \frac{1}{50} < \left| \frac{Y_j(r)}{X_j(r)} \right| < 3.$$

Proof. We first analyze the prescribed matrix product at $r = 0$, where the last six maps coincide. This simplifies the calculation of the initial momentum and the intermediate states. We then continue the construction to positive r and verify the actual gate decisions. The low-gradient calls require a separate check because their gradients vanish at $r = 0$.

Since $\mathsf{R}_0 = \mathsf{L}_0$, the limiting cycle is

$$\mathsf{P}_0 = \mathsf{L}_0^6 \mathsf{A}_0^5 \mathsf{B}_0^3 = \begin{pmatrix} 1 & -6 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 - a_\circ & -1 \\ a_\circ & 1 \end{pmatrix}^5 \begin{pmatrix} 1 & 0 \\ 3a_\circ & 1 \end{pmatrix}. \tag{B.4}$$

To obtain growth with sign reversal, we seek an eigenvalue less than -1. Each factor above has determinant one, so it suffices to show that the trace of P_0 is less than -2. Plugging in the chosen value of $a_\circ$ and evaluating the trace gives

$$\det \mathsf{P}_0 = 1, \qquad \operatorname{tr} \mathsf{P}_0 < -2. \tag{B.5}$$

Hence P_0 has two distinct real negative eigenvalues, one of which is less than -1.

We now determine an initial momentum corresponding to this growing direction. Normalizing the initial position to one, we seek an eigenvector of the form $(1, -h)^\top$. Eliminating the eigenvalue from the two component equations gives

$$f_{\mathrm{neg}}(h) := (\mathsf{P}_0)_{12} h^2 + \big[(\mathsf{P}_0)_{22} - (\mathsf{P}_0)_{11}\big] h - (\mathsf{P}_0)_{21} = 0.$$

Thus the initial momentum is determined by the cycle.

To verify the intermediate gate decisions and the bounds needed later, we locate this root accurately enough to bound the entire trajectory. We use a simple bisection to show that there is a unique root h_0 satisfying

$$0.22197 < h_0 < 0.22198. \tag{B.6}$$

Indeed, f_{neg} is negative at the left endpoint and positive at the right endpoint. Its derivative is affine and positive at both endpoints, so it is positive throughout the interval.

Set $g_0 = -(\mathsf{P}_0)_{11} + (\mathsf{P}_0)_{12}h_0$. The root equation gives the eigenvector identity, and evaluating this affine expression at the endpoints of (B.6) gives

$$\mathsf{P}_0(1, -h_0)^\top = -g_0(1, -h_0)^\top, \qquad 1.004 < g_0 < 1.0041. \tag{B.7}$$

In particular, the gain exceeds one by a fixed positive amount. This margin will allow us to retain the weaker bound $g(r) > 1.002$ for all sufficiently small positive r.

It remains to verify that the intermediate states follow the prescribed gate decisions. Starting from $z_0 = (1, -h_0)^\top$, apply three B_0 maps, five A_0 maps, and six L_0 maps, and write $z_j = (X_j, Y_j)^\top$ for the resulting states. Each component is affine in h_0, so evaluating it at the endpoints of (B.6) bounds its value throughout that interval. This gives Table 1. We use the table to check the gate signs and then derive the preparation and coupling bounds.

Table 1: Bounds on the limiting scalar trajectory. Each row gives the state before call j. An interval contains the corresponding component.

Call j	X_j	Y_j
0	1	$[-0.2220, -0.2219]$
1	1	$[-0.1599, -0.1598]$
2	1	$[-0.0978, -0.0977]$
3	1	$[-0.0357, -0.0355]$
4	$[0.9734, 0.9735]$	$[0.0265, 0.0266]$
5	$[0.8864, 0.8865]$	$[0.0870, 0.0871]$
6	$[0.7443, 0.7444]$	$[0.1420, 0.1421]$
7	$[0.5560, 0.5561]$	$[0.1883, 0.1884]$
8	$[0.3331, 0.3333]$	$[0.2228, 0.2229]$
9	$[0.1102, 0.1104]$	$[0.2228, 0.2229]$
10	$[-0.1126, -0.1125]$	$[0.2228, 0.2229]$
11	$[-0.3355, -0.3353]$	$[0.2228, 0.2229]$
12	$[-0.5584, -0.5582]$	$[0.2228, 0.2229]$
13	$[-0.7812, -0.7811]$	$[0.2228, 0.2229]$

For calls 0–7, the position before the call is positive, so the gradient $a_\circ X_j$ is positive. The new momentum on call j is Y_{j+1}. It is negative on calls 0–2 and positive on calls 3–7, giving the prescribed three rejected updates followed by five accepted updates. The table also shows that $X_j, Y_j > 0$ for $j = 8, 9$, whereas $X_j < 0 < Y_j$ for $j = 10, \ldots, 13$. Thus the position crosses zero on the second low-curvature call, while the momentum remains positive. These signs will verify the remaining gate decisions for positive r.

We next record the bounds needed to prepare this trajectory from zero momentum. The preparation argument in Appendix B.3 requires the absolute momentum-to-position ratios to lie in a common interval bounded away from zero. Every position in the table is nonzero, and direct comparison of the interval endpoints gives

$$\frac{1}{50} < \left|\frac{Y_j}{X_j}\right| < 3, \qquad 0 \le j < 14. \tag{B.8}$$

The constants $1/50$ and 3 provide one common range for all fourteen states, with strict inequalities that persist under sufficiently small changes in r.

For the coupling construction, we need a different comparison. The coordinates assigned gradient $a_\circ x$ will supply the correction needed to produce the gradients $r^3 y$ on other coordinates. The required coupling is controlled by the squared positions in these two groups. We therefore define

$$S_H = \sum_{j=0}^{7} X_j^2, \qquad S_C = \sum_{j=10}^{13} X_j^2.$$

Appendix B.2 will show that it suffices to have $S_C/(a_\circ S_H) < (8/5)^2$. The coefficient $8/5$ corresponds to the coupling bound in the theorem, since $(8/5)r = 2/(5\sqrt{\kappa})$.

Squaring the interval endpoints in Table 1 and summing gives $S_H > 6.596$ and $S_C < 1.048$. Direct evaluation then gives $\left(\frac{8}{5}\right)^2 a_\circ \cdot 6.596 - 1.048 > 0$. Combining these bounds gives

$$S_C/(a_\circ S_H) < (8/5)^2. \tag{B.9}$$

This will be used later in the coupling argument

Verification for positive r. We now transfer the limiting construction to the parameters used by the algorithm. The entries of P_r are polynomials in r, and $-g_0$ is a simple real eigenvalue of P_0. Its eigenvalue and eigenvector therefore admit continuous real branches near zero. Since the first component of the limiting eigenvector is one, we may normalize these branches as

$$\mathsf{P}_r(1, -h(r))^\top = -g(r)(1, -h(r))^\top, \qquad h(0) = h_0, \quad g(0) = g_0.$$

The bounds on h_0 and g_0 imply $h(r) > 0$ and $g(r) > 1.002$ for sufficiently small positive r.

The intermediate states obtained from the prescribed maps are also continuous in r. Hence all position and momentum signs in Table 1, together with the ratio bounds in (B.8), persist on a sufficiently small positive interval. In particular, the first eight calls retain the prescribed three rejected and five accepted updates.

The remaining calls require a direct check because their gradients vanish at $r = 0$. On calls 8–9, the position and incoming momentum are positive. For $r > 0$, both the gradient r^2x and the candidate $\beta y + r^2 x$ are therefore positive, so both calls are accepted. On calls 10–13, we have $y > 0$ and $q = \beta + r^3 > 0$. The assigned gradient $h = r^3 y$ gives candidate $p = qy$, and hence $ph = qr^3y^2 > 0$. These four calls are also accepted.

Thus all constructed maps agree with the actual update. The eigenvalue identity establishes growth over the cycle, while the continued profile bounds give the nonzero positions and the required ratio estimates. This proves the lemma. □

We can now repeat the cycle indefinitely. Multiplying the state and prescribed gradient by the same nonzero real factor multiplies the gate product by its square, preserving every gate decision. As a result,

$$z_{14n+j}(r) = (-g(r))^n z_j(r), \qquad n \in \mathbb{N}_0, \quad 0 \le j < 14.$$

Since $g(r) > 1$ and the finitely many positions $X_j(r)$ are nonzero, the scalar error diverges in magnitude. The next subsection constructs admissible coupled Hessians that generate these prescribed gradients.

B.2 Constructing admissible periodic Hessians

We now construct admissible Hessians that generate the growing trajectory from Appendix B.1. The prescribed gradients $a_\circ x$ and $r^2 x$ can be produced by positive diagonal curvature. During the final four calls, however, the gradient $r^3 y$ has the opposite sign to the scalar error and requires coordinate coupling. We therefore use fourteen coordinates at different stages of the scalar cycle, so that coordinates with gradient $a_\circ x$ are available at every call to supply the required correction.

Our goal is to construct symmetric scaled Hessians that produce all prescribed gradients while satisfying the diagonal and relative coupling bounds. We first recall the growing vector trajectory, then verify the Hessian construction. The way the coordinates advance through the scalar cycle will also imply periodicity and allow us to check admissibility at a single reference time.

Fix a sufficiently small positive r, use the scalar trajectory extended by $z_{t+14}(r) = -g(r)z_t(r)$, and set $\vartheta = g(r)^{1/14} > 1$. As in (5.1), the reference position and scaled momentum vectors have coordinates

$$\xi_{t,j} = \vartheta^{-j} X_{t+j}(r), \qquad \upsilon_{t,j} = \vartheta^{-j} Y_{t+j}(r), \qquad 0 \le j < 14.$$

Thus each coordinate follows one copy of the scalar trajectory, starting at a different stage.

The weights ϑ^{-j} ensure that advancing one call multiplies the vector norm by the same factor ϑ. Indeed, for $j < 13$, $\xi_{t+1,j} = \vartheta \xi_{t,j+1}$. For the last coordinate, $X_{t+14} = -g(r)X_t$ and $\vartheta^{14} = g(r)$ give $\xi_{t+1,13} = -\vartheta \xi_{t,0}$. The same relations hold for momentum. Consequently, with

$$S(v_0, \dots, v_{13}) = (v_1, \dots, v_{13}, -v_0),$$

we obtain

$$\xi_{t+1} = \vartheta S \xi_t, \qquad v_{t+1} = \vartheta S v_t,$$

proving (5.2). Since S is orthogonal, the reference position norm grows by ϑ at every call. We also have $S^{14} = -I$, which will imply periodicity of the Hessians.

Producing the prescribed gradients. At time t, coordinate j is at stage $(t + j) \bmod 14$ of the scalar cycle. Let $\mathcal{H}_t$, $\mathcal{N}_t$, and $\mathcal{C}_t$ contain the coordinates at stages 0–7, 8–9, and 10–13, respectively. These sets contain eight, two, and four coordinates, and their prescribed gradients are $a_\circ \xi_{t,j}$, $r^2 \xi_{t,j}$, and $r^3 v_{t,j}$, respectively.

At a fixed time, suppress the time subscripts and write the scaled Hessian as $\widehat{K}_t = \widehat{A}_t + \widehat{J}_t$, where $\widehat{A}_t$ is diagonal. We use diagonal curvature r^2 on $\mathcal{N} \cup \mathcal{C}$ and add coupling only between $\mathcal{H}$ and $\mathcal{C}$. The gradient on $\mathcal{N}$ is then already correct. On $\mathcal{C}$, coupling must supply the difference between the desired gradient and the diagonal contribution:

$$r^3 v_{\mathcal{C}} - r^2 \xi_{\mathcal{C}} = r^2 w, \qquad w = r v_{\mathcal{C}} - \xi_{\mathcal{C}}.$$

Set $R = \|\xi_{\mathcal{H}}\|_2^2$. All reference positions are nonzero by Lemma B.1, so $R > 0$.

The rank-one block

$$(\widehat{J}_t)_{\mathcal{C}\mathcal{H}} = r^2 \frac{w \xi_{\mathcal{H}}^\top}{R}$$

maps $\xi_{\mathcal{H}}$ to the required correction $r^2 w$. We include its transpose as the $\mathcal{H}\mathcal{C}$ block and set all other blocks of $\widehat{J}_t$ to zero. This makes $\widehat{J}_t$ symmetric with zero diagonal.

The transpose block also contributes

$$r^2 \frac{w^\top \xi_{\mathcal{C}}}{R} \xi_{\mathcal{H}}$$

to the gradient on $\mathcal{H}$. Since this is a common scalar multiple of $\xi_{\mathcal{H}}$, we cancel it by choosing the diagonal curvature on $\mathcal{H}$ to be

$$\widehat{a} = a_\circ - r^2 \frac{w^\top \xi_{\mathcal{C}}}{R}.$$

These are the choices in (5.4) and (5.3). They give gradient $r^3 v_{\mathcal{C}}$ on $\mathcal{C}$, restore gradient $a_\circ \xi_{\mathcal{H}}$ on $\mathcal{H}$, and leave gradient $r^2 \xi_{\mathcal{N}}$ on $\mathcal{N}$. This proves (5.5). Together with the scalar gate decisions, these identities ensure that a call from (ξ_t, v_t) reaches (ξ_{t+1}, v_{t+1}).

Periodicity and reduction to one reference time. Before checking admissibility, we show that the same bounds will hold at every call. Under (5.2), both reference vectors are multiplied by ϑ and transformed by S. The three coordinate sets shift accordingly. Thus R and $w^\top \xi_{\mathcal{C}}$ are both multiplied by ϑ^2, leaving $\widehat{a}$ unchanged.

In the coupling block, the numerator $w \xi_{\mathcal{H}}^\top$ and the denominator R also acquire the same factor ϑ^2. After cancellation, the remaining change is conjugation by S. Both $\widehat{A}_t$ and $\widehat{J}_t$ therefore transform by this conjugation, giving

$$\widehat{K}_{t+1} = S \widehat{K}_t S^\top, \qquad \widehat{K}_{t+14} = \widehat{K}_t. \tag{B.10}$$

The second identity follows from $S^{14} = -I$.

Conjugation by a signed permutation preserves the diagonal values up to reordering and preserves the relative operator norm. Consequently, once admissibility is established at reference time $t = 0$, the same bounds hold for every $t \geq 0$.

Checking admissibility. We first verify the diagonal bounds at $t = 0$. Since $0 \in \mathcal{H}_0$ and $\xi_{0,0} = 1$, we have $R \geq 1$. As $r \downarrow 0$, the scalar profile and the weights ϑ^{-j} have finite limits. Hence $w^\top \xi_{\mathcal{C}}/R$ remains bounded, and

$$\widehat{a} = a_\circ - r^2 \frac{w^\top \xi_{\mathcal{C}}}{R} \longrightarrow a_\circ \in (0, M).$$

Together with $r^2 \to 0$, this gives $r^2 \leq \widehat{a} \leq M$ for sufficiently small positive r. All remaining diagonal entries equal r^2, so the entire diagonal lies in $[r^2, M]$.

It remains to bound the relative coupling. The diagonal values on $\mathcal{C}$ and $\mathcal{H}$ are r^2 and $\widehat{a}$, respectively. Normalizing the coupling block by their inverse square roots therefore gives

$$T = \frac{r}{\sqrt{\widehat{a}}} \frac{w \xi_{\mathcal{H}}^\top}{R}.$$

The full normalized matrix has off-diagonal blocks T and $T^\top$ and is zero elsewhere. Its square has diagonal blocks $T^\top T$ and $TT^\top$, so its operator norm equals $\|T\|_{\mathrm{op}}$. Using the norm of this rank-one block, we obtain

$$\|\widehat{A}_t^{-1/2} \widehat{J}_t \widehat{A}_t^{-1/2}\|_{\mathrm{op}} = r\Gamma(r), \qquad \Gamma(r) = \frac{\|r v_{\mathcal{C}} - \xi_{\mathcal{C}}\|_2}{\sqrt{\widehat{a}}\, \|\xi_{\mathcal{H}}\|_2}.$$

This proves (5.6). The conjugation identity in (B.10) also shows that this norm, and hence $\Gamma(r)$, is independent of time.

To bound $\Gamma(r)$, we use the limiting position bounds from Appendix B.1. Its denominator has a positive limit because $\widehat{a} \to a_\circ > 0$ and $R \geq 1$. Thus Γ extends continuously to $r = 0$. At reference time zero, the coordinates in $\mathcal{H}_0$ and $\mathcal{C}_0$ have indices 0–7 and 10–13, respectively. Moreover, $w \to -\xi_{\mathcal{C}}$, and the squared weights ϑ^{-2j} tend to $g_0^{-j/7}$.

Since $g_0 > 1$, these limiting weights decrease with j. The weights in the numerator are therefore at most $g_0^{-10/7}$, whereas those in the denominator are at least g_0^{-1}. Writing X_j, S_H, and S_C for the limiting quantities from Appendix B.1, we obtain

$$\Gamma(0)^2 = \frac{\sum_{j=10}^{13} g_0^{-j/7} X_j^2}{a_\circ \sum_{j=0}^{7} g_0^{-j/7} X_j^2} \leq g_0^{-3/7} \frac{S_C}{a_\circ S_H} < \left(\frac{8}{5}\right)^2. \tag{B.11}$$

The last inequality uses $g_0^{-3/7} < 1$ and the bound (B.9).

By continuity, $\Gamma(r) < 8/5$ for all sufficiently small positive r. Consequently, the relative coupling satisfies

$$\rho_{\mathrm{actual}}(r) = r\Gamma(r) < \frac{8}{5} r = \frac{2}{5\sqrt{\kappa}}. \tag{B.12}$$

This bound holds at every reference time. For sufficiently small r, it is also less than one, so

$$\widehat{K}_t = \widehat{A}_t^{1/2} \big(I + \widehat{A}_t^{-1/2} \widehat{J}_t \widehat{A}_t^{-1/2}\big) \widehat{A}_t^{1/2} \succ 0.$$

Finally, rescaling by $1/\eta = 16L$ gives diagonal entries in $[\mu, L]$, since $r^2/\eta = \mu$ and $M/\eta = L$. The rescaling leaves the relative coupling norm unchanged. Thus the constructed Hessians satisfy (2.1) whenever $2/(5\sqrt{\kappa}) \leq \rho < 1$.

We have obtained a 14-periodic admissible Hessian sequence that generates the growing reference trajectory.

B.3 Reaching the required initial state from zero momentum

The reference trajectory constructed in the preceding subsection starts with nonzero momentum. To satisfy the initialization in Theorem 2.1, we must reach this state through the original recurrence, starting from zero momentum.

We construct this preparation using diagonal Hessians, so the coordinates evolve independently. The main requirement is that every coordinate reaches its target after the same number of calls. The following lemma provides a common length of 100 calls for all target ratios arising in Lemma B.1.

Lemma B.2 (Uniform diagonal preparation)**.** *For all sufficiently small positive r, every scalar target (x, y) with $x \neq 0$ and $1/50 \leq |y/x| \leq 3$ is reachable in exactly 100 calls from a suitable initial position and zero momentum, using predetermined scalar curvatures in $[r^2, M]$. One sufficiently small interval of positive r works for all such targets.*

Proof. The scalar recurrence with gradient ax is homogeneous: multiplying the initial position by a nonzero real factor multiplies every subsequent position and momentum by that factor, while preserving the gate decisions. It therefore suffices to reach a nonzero state with the prescribed ratio y/x. Scaling the initial position then matches both target components.

Write $Q = |y/x| \in [1/50, 3]$. Starting from $(1, 0)$, we construct the ratios $-Q$ and Q in 100 calls, using different call orders and curvatures. For $-Q$, we begin with six calls with gradient $h = Mx$ to obtain a negative position and positive momentum. We then append 94 rejected calls that reduce the momentum while holding the position fixed. For Q, we first use 94 calls with gradient $h = r^2 x$, then choose a constant curvature $a \in [r^2, M]$ for the final six calls to obtain the desired ratio.

We first analyze the six-call sequence at $r = 0$. At this limit, the initial 94 calls in the second construction leave $(1, 0)$ unchanged, so the six calls have the same limiting initial state in both constructions. For constant curvature $a \in [0, M]$, consider the recurrence obtained by accepting each update:

$$\bar{y}_{j+1} = \bar{y}_j + a\bar{x}_j, \qquad \bar{x}_{j+1} = \bar{x}_j - \bar{y}_{j+1}, \qquad (\bar{x}_0, \bar{y}_0) = (1, 0).$$

Using $a \leq M = 1/16$, induction gives

$$\begin{aligned} &0 \leq \bar{y}_j \leq ja, \\ &1 \geq \bar{x}_j = 1 - \sum_{k=1}^{j} \bar{y}_k \geq 1 - \frac{j(j+1)}{2} a \geq \frac{1}{16}, \qquad 1 \leq j \leq 5. \end{aligned} \tag{B.13}$$

Thus all six starting positions $\bar{x}_0, \ldots, \bar{x}_5$ are positive. For $a > 0$, each gradient and each new momentum are positive, so all six calls are accepted, even if the sixth update makes $\bar{x}_6$ negative.

At $a = M$, direct evaluation gives

$$\bar{x}_6 < 0 < \bar{y}_6, \qquad 4 < \frac{\bar{y}_6}{-\bar{x}_6} < \frac{9}{2}. \tag{B.14}$$

These bounds provide the starting state and ratio for the 94 rejected calls in the opposite-sign case. In the matching-sign case, the negative final position at $a = M$ provides an endpoint for the continuity argument as a varies. The uniform positive bound in (B.13) will be used below to verify acceptance of the six calls for sufficiently small positive r in both constructions.

Targets with opposite signs. In this case, we begin with the six calls constructed above with $a = M$ to obtain a negative position and positive momentum. We then append 94 calls that adjust the momentum while holding the position fixed.

Starting from $(1, 0)$, make six calls with curvature M. The strict signs established above persist for sufficiently small positive r, so all six calls are accepted and their final state satisfies $x_6(r) < 0 < y_6(r)$. Set $Q_0(r) = \frac{y_6(r)}{-x_6(r)}$. By (B.14), $Q_0(0) \in (4, 9/2)$.

We use the remaining 94 calls to reduce this ratio to the desired value Q. If the position is held at $x_6(r)$ and the curvature is fixed at a_h, the normalized momentum $Q_j := y_{6+j}/(-x_6(r))$ satisfies

$$Q_{j+1} = \beta Q_j - a_h.$$

Solving this recurrence gives $Q_{94} = \beta^{94} Q_0(r) - a_h \sum_{j=0}^{93} \beta^j$. We therefore choose

$$a_h(r, Q) = \frac{\beta^{94} Q_0(r) - Q}{\sum_{j=0}^{93} \beta^j}, \tag{B.15}$$

so that the prescribed sequence ends with $Q_{94} = Q$.

We now check that this curvature is admissible and that every prescribed hold agrees with the gate. At $r = 0$,

$$a_h(0, Q) = \frac{Q_0(0) - Q}{94}.$$

Uniformly over $Q \in [1/50, 3]$, we have

$$0 < \frac{Q_0(0) - 3}{94} \le a_h(0, Q) \le \frac{Q_0(0) - 1/50}{94} < M.$$

The last inequality follows from $Q_0(0) < 9/2$ and $(9/2 - 1/50)/94 < M$. Thus the choice of 94 calls leaves a strict margin between the required curvature and both endpoints 0 and M. Since $r^2 \to 0$, continuity on the compact target interval gives $r^2 < a_h(r, Q) < M$ for all sufficiently small positive r, uniformly in Q.

For the gate decisions, the recurrence defining Q_j ends at $Q_{94} = Q > 0$. Since $\beta > 0$ and $a_h > 0$, reading the recurrence backwards,

$$Q_j = \frac{Q_{j+1} + a_h}{\beta},$$

shows that every Q_j is positive. Consequently, each candidate momentum is positive, whereas its gradient $a_h x_6(r)$ is negative. Every one of these 94 position updates is therefore rejected, as prescribed.

After 100 calls, the position is still $x_6(r) \neq 0$ and the momentum-to-position ratio is $-Q$. Scaling the initial position gives the desired target with opposite signs.

Targets with matching signs. In this case, we place the 94 additional calls before the six-call sequence. Using the smallest allowed curvature during this prefix keeps its final state close to $(1, 0)$, so the last six calls can adjust the target ratio in the same setting as the limiting calculation.

Starting from $(1, 0)$, consider 94 accepted calls with curvature r^2. As $r \downarrow 0$, every state in this finite prefix converges to $(1, 0)$. Hence all positions remain positive for sufficiently small positive r. The gradients are then positive, and the momentum recurrence makes every new momentum positive. Thus all 94 calls agree with the actual gate.

For the final six calls, choose a common curvature $a \in [r^2, M]$. The accepted recurrence for these calls is continuous in (r, a). At $r = 0$, the entering state is $(1, 0)$, and (B.13) bounds the first five output positions away from zero uniformly over $a \in [0, M]$. By uniform continuity, all six pre-call positions remain positive for sufficiently small positive r, uniformly over $a \in [r^2, M]$. The entering momentum is positive, so each gradient and each new momentum are positive as well. All six calls are therefore accepted.

Write the final state as $(x_{100}(r, a), y_{100}(r, a))$. It depends continuously on a, and $y_{100}(r, a) > 0$ throughout the allowed curvature interval. We choose a by comparing its two endpoints.

At $a = r^2$, all 100 calls use curvature r^2, so the final state tends to $(1, 0)$ as $r \downarrow 0$. For sufficiently small positive r,

$$x_{100}(r, r^2) > 0, \qquad 0 < \frac{y_{100}(r, r^2)}{x_{100}(r, r^2)} < \frac{1}{50}.$$

At $a = M$, the final state tends to the six-call state in (B.14). Hence

$$x_{100}(r, M) < 0 < y_{100}(r, M)$$

for sufficiently small positive r.

For a prescribed $Q \in [1/50, 3]$, consider the continuous function

$$a \longmapsto y_{100}(r, a) - Q x_{100}(r, a).$$

It is negative at $a = r^2$, because the endpoint ratio is below $1/50 \le Q$. It is positive at $a = M$, because the final position is negative and the final momentum is positive. The intermediate value theorem therefore gives a curvature $a \in (r^2, M)$ such that $y_{100}(r, a) = Q x_{100}(r, a)$. Since the final momentum is positive, this identity also gives $x_{100}(r, a) > 0$. Thus the final state has the desired ratio Q. Scaling the initial position matches the prescribed target magnitude and sign.

All smallness conditions used above are uniform over $Q \in [1/50, 3]$: the rejected-update construction has uniform curvature margins, and the accepted-update construction uses uniform positivity bounds and endpoint conditions. Both constructions use exactly 100 calls with curvatures in $[r^2, M]$. The curvatures and initial position depend only on r and the prescribed target and are chosen before the algorithm is executed. This proves the lemma. □

B.4 Completing the divergence proof

We now combine the diagonal preparation with the periodic Hessians to obtain a sequence, fixed in advance, that produces geometric growth from zero momentum. Fix $r > 0$ sufficiently small that all preceding constructions apply.

By (5.1) and Lemma B.1, every coordinate of (ξ_0, υ_0) has nonzero position and satisfies

$$\frac{1}{50} < \left|\frac{\upsilon_{0,j}}{\xi_{0,j}}\right| = \left|\frac{Y_j(r)}{X_j(r)}\right| < 3.$$

Applying Lemma B.2 to these fourteen targets in parallel gives a nonzero initial point x_0, zero initial momentum $y_0 = \eta v_0 = 0$, and 100 diagonal scaled Hessians with entries in $[r^2, M]$ such that

$$(x_{100}, y_{100}) = (\xi_0, \upsilon_0).$$

For subsequent calls, prescribe $K_{100+t} = \widehat{K}_t$ for every $t \geq 0$, and set $H_t = K_t/\eta$ throughout. Both the preparation curvatures and the subsequent matrices are determined by the tuning and the fixed reference before the method runs. They define full quadratic-gradient maps $z \mapsto H_t z$ with common minimizer zero. The preparation has zero coupling, while the subsequent Hessians satisfy (B.12). Together with the diagonal bounds, this makes the complete sequence admissible whenever $2/(5\sqrt{\kappa}) \leq \rho < 1$. By (B.10), it is 14-periodic after the first 100 calls.

It remains to verify that the actual iterates follow the reference. By (5.5) and the verified scalar gate decisions, a call starting at (ξ_t, υ_t) reaches $(\xi_{t+1}, \upsilon_{t+1})$. Induction from the prepared state therefore gives

$$(x_{100+t}, y_{100+t}) = (\xi_t, \upsilon_t), \qquad t \in \mathbb{N}_0.$$

Since the signed shift S preserves the Euclidean norm, (5.2) yields

$$\|x_{100+t}\|_2 = \vartheta(r)^t \|x_{100}\|_2, \qquad t \in \mathbb{N}_0.$$

Here $x_{100} = \xi_0 \neq 0$ and $\vartheta(r)^{14} = g(r) > 1.002$, so $\|x_t\|_2 \to \infty$.

Finally, the scalar construction, Hessian bounds, and uniform preparation hold simultaneously on a common universal interval of sufficiently small positive r. Since $r = 1/(4\sqrt{\kappa})$, there is a universal finite κ_0 such that the complete construction works for every $\kappa \geq \kappa_0$. This proves part (ii) of Theorem 2.1.

C Diagonal convergence and comparison with linear methods

We study the diagonal case to explain the acceleration mechanism described in Section 3 and establish a comparison with linear methods. When $\rho = 0$, the coordinates evolve independently, and each gradient coordinate has the same sign as the corresponding error. The cautious gate therefore accepts a proposed update precisely when it points toward the coordinate minimizer. This setting allows us to analyze the effect of the gate directly, while still permitting arbitrary changes in the diagonal curvatures.

We first prove the coordinatewise convergence bound in Theorem C.1. The proof follows successive runs of accepted updates and the intervals of rejected updates that may follow them. We bound the error after each crossing of the minimizer and relate the subsequent waiting time to the error reduction already achieved. Combining these estimates gives a convergence bound at every iteration, including calls with rejected updates.

To assess this rate, we then consider linear methods whose coefficients are chosen before observing gradient responses. This class includes gradient descent, heavy-ball momentum, and Nesterov acceleration with predetermined parameter schedules. We define the class precisely and prove in Theorem C.3 that its optimal worst-case rate under the same diagonal oracle is attained by fixed-step gradient descent. Comparing this rate with the cautious convergence bound establishes the improvement obtained by cautious momentum over this class. We conclude by explaining why the restriction on the coefficients matters and why the lower bound does not extend to arbitrary adaptive methods.

C.1 Accelerated convergence for diagonal Hessians

For diagonal Hessians, the coordinates evolve independently. We can therefore prove the bound described in Section 3 by following a single coordinate through successive crossings of its minimizer. The main issue is that a small error after a crossing may be followed by many rejected updates. We relate this waiting time to the error reduction already achieved, which allows us to count every gradient call in the convergence rate.

Theorem C.1 (Accelerated convergence for diagonal Hessians)**.** *For every diagonal Hessian sequence with entries in $[\mu, L]$, cautious momentum under the tuning* (2.3) *and with zero initial momentum satisfies*

$$|x_{t,i} - x_i^\star| \le 2e^{-\gamma_\kappa t}|x_{0,i} - x_i^\star|, \qquad \gamma_\kappa = \frac{\log 2}{4\sqrt{\kappa}[6 + \log(2 + 4\sqrt{\kappa})]}, \tag{C.1}$$

for every coordinate i and every integer $t \ge 0$. The same factor holds in every ℓ_p norm, $1 \le p \le \infty$, uniformly in dimension and in the diagonal Hessian sequence. The bound counts all gradient calls, including those with rejected updates.

We first reduce the proof to a scalar recurrence. Translate the coordinate minimizer to zero, write x_t for the scalar error, and set $y_t = \eta v_t$. As in Section 4, let

$$r = \frac{1}{4\sqrt{\kappa}}, \qquad \beta = 1 - r, \qquad M = \frac{1}{16}.$$

The scaled curvature satisfies $r^2 \le a_t \le M$, and the scalar update is

$$p_t = \beta y_t + a_t x_t, \qquad y_{t+1} = p_t, \qquad x_{t+1} = x_t - p_t \mathbf{1}\{x_t p_t > 0\}. \tag{C.2}$$

Here the gate uses $x_t p_t$ because $a_t > 0$. The recurrence is invariant under simultaneous sign reversal and scaling. A coordinate that reaches zero stays there. If $x_t \ne 0$ but $p_t = 0$, the position is unchanged, the momentum becomes zero, and the next call is accepted.

We describe the trajectory through runs of accepted updates. We call an update accepted if $x_t p_t > 0$. An *accepted run* is a maximal sequence of consecutive accepted calls. Let σ denote its first call. If $\sigma = 0$, the incoming momentum is $y_0 = 0$. If $\sigma > 0$, the preceding call was rejected, so

$$x_\sigma = x_{\sigma-1}, \qquad y_\sigma = p_{\sigma-1}, \qquad x_{\sigma-1} p_{\sigma-1} \le 0.$$

Consequently, $x_\sigma y_\sigma = x_{\sigma-1} p_{\sigma-1} \le 0$. Since an accepted call starts at a nonzero position, the incoming momentum is therefore either zero or opposite in sign to the position.

We next explain why a run ends when the position first reaches or crosses zero. By symmetry, consider an accepted call with $x_t > 0$ and $p_t > 0$. The next candidate satisfies

$$p_{t+1} = \beta p_t + a_{t+1}(x_t - p_t) = (\beta - a_{t+1}) p_t + a_{t+1} x_t > 0,$$

where we used $\beta \ge 3/4 > M \ge a_{t+1}$. Thus the next call is accepted exactly when $x_{t+1} > 0$. Accepted updates therefore continue while the position remains positive, and the first update that reaches or crosses zero is followed by a rejected call. The same conclusion holds for a run starting at a negative position by simultaneous sign reversal. The following lemma shows that every such run reaches or crosses zero within a bounded number of calls.

Lemma C.2 (Bounds for accepted runs and rejected updates)**.** *Consider an accepted run starting at error magnitude $A > 0$, with incoming momentum zero or opposite in sign to the position. The run reaches or crosses zero in fewer than $5/r$ calls. If it crosses zero to error magnitude $B > 0$, its final momentum magnitude V satisfies*

$$B < V < \frac{A}{2}. \tag{C.3}$$

The number of rejected calls before the next accepted run, including a possible zero-candidate call, is at most

$$1 + \frac{1}{r} \log\left(1 + \frac{V}{rB}\right). \tag{C.4}$$

C.1.1 Error reduction and duration of accepted updates

We first bound the error after a crossing, then show that a crossing must occur within the claimed time. By symmetry, suppose the run starts at $x = A > 0$ with incoming momentum $y \le 0$. Its first accepted displacement satisfies $0 < u_1 = \beta y + aA \le MA < A$. Thus the first update leaves the position positive. While the position remains positive, the momentum recurrence gives positive candidates, so the following updates are also accepted.

Use local indices for this run, and suppose its nth update crosses to $x_n = -B < 0$. Let $u_1, \dots, u_n$ be the accepted displacements. Before the crossing, $0 < x_{j-1} \le A$, and

$$u_j = \beta u_{j-1} + a_{j-1} x_{j-1}, \qquad j \ge 2.$$

Expanding the change in squared momentum and discarding the nonpositive terms gives

$$u_j^2 - u_{j-1}^2 = 2a_{j-1}x_{j-1}u_j - a_{j-1}^2 x_{j-1}^2 - (1-\beta^2)u_{j-1}^2 \le 2MAu_j.$$

The first displacement also satisfies $u_1^2 \le 2MAu_1$. Summing these inequalities and using $\sum_{j=1}^n u_j = A + B$ therefore gives $V^2 = u_n^2 \le 2MA(A+B)$. Since $B = V - x_{n-1} < V$, we obtain $(V/A)^2 \le 2M(1 + V/A)$. The positive root of this quadratic is $M + \sqrt{M^2 + 2M} = (1 + \sqrt{33})/16 < 1/2$. This proves (C.3). If an update lands exactly at zero, the coordinate remains there and needs no further analysis.

We next bound the duration of the run. After its first update, the position and momentum are positive. Write $R = y/x \ge 0$. As long as the next position remains positive, its momentum-to-position ratio is

$$R^+ = \frac{\beta R + a}{1 - \beta R - a}.$$

This expression is increasing in both R and a on the domain $1 - \beta R - a > 0$. A crossing occurs when $\beta R + a \ge 1$. Consequently, the remaining time to reach or cross zero is no longer than for the recurrence with constant curvature $a = r^2$ and initial ratio $R = 0$.

To bound this comparison trajectory, normalize its initial state to $(\widehat{x}_0, \widehat{y}_0) = (1, 0)$. Its accepted recurrence satisfies

$$\widehat{x}_{n+2} = (2 - r - r^2)\widehat{x}_{n+1} - \beta \widehat{x}_n.$$

The characteristic roots are $\sqrt{\beta} e^{\pm i\phi}$, where $0 < \phi < \pi/2$ and

$$\cos\phi = \frac{2 - r - r^2}{2\sqrt{1-r}}, \qquad \sin^2\phi = \frac{r^2(3+r)}{4}, \qquad \phi \ge \frac{\sqrt{3}}{2} r. \tag{C.5}$$

The last bound follows from $\phi \ge \sin\phi$. The position therefore has the form $\widehat{x}_n = \beta^{n/2}\big[\cos(n\phi) + c\sin(n\phi)\big]$ for a real constant c determined by the initial state. The bracket starts at one and has its first zero at an angle in $(0, \pi)$. It is negative for an angle interval of length π after that zero. Since the sampling increment ϕ is less than π, a nonpositive sampled value occurs within $\lceil \pi/\phi \rceil$ calls.

Including the first accepted update of the original run, we conclude that

$$T_{\text{active}} \le 1 + \lceil \pi/\phi \rceil \le 2 + \frac{2\pi}{\sqrt{3}\, r} < \frac{5}{r},$$

where the last inequality uses $r \le 1/4$. This proves the accepted-run estimates in Lemma C.2.

C.1.2 Rejected updates and convergence at every iteration

We now bound the waiting time after a crossing. By symmetry, write the state immediately after that crossing as $x = -B$ and $y = V > 0$. The gradient is negative, so a candidate is accepted as soon as it becomes negative. Until then, the position remains fixed and the momentum satisfies, with local indices,

$$y_k \le \beta^k V - r^2 B \sum_{j=0}^{k-1} \beta^j = \beta^k V - rB(1 - \beta^k).$$

Thus the initial momentum decays while the negative gradients move the candidate toward acceptance.

Set

$$k_* = \left\lceil \frac{1}{r} \log\left(1 + \frac{V}{rB}\right) \right\rceil.$$

Since $\beta^k \le e^{-rk}$, the above upper bound is nonpositive at $k = k_*$. If movement has not already resumed, the momentum at that call must be zero, and the next call is accepted. Hence there are at most

$$k_* \le 1 + \frac{1}{r} \log\left(1 + \frac{V}{rB}\right)$$

rejected calls. This proves (C.4) and completes the proof of Lemma C.2.

The waiting-time bound grows when B is small. However, a small B also means that the preceding run has already achieved a large error reduction. We now use this relation to obtain a bound in terms of the total number of calls.

Assume $x_0 \neq 0$. Let σ_j be the time just before the first update of the jth accepted run, and let $A_j = |x_{\sigma_j}|$. The initial run starts at $\sigma_0 = 0$. For every run followed by a nonzero crossing, let V_j be its final momentum magnitude. Lemma C.2 gives

$$A_{j+1} < \frac{A_j}{2}, \qquad V_j < \frac{A_j}{2}.$$

Define

$$\ell_j = \log(A_j/A_{j+1}) \ge \log 2.$$

This quantity measures the logarithmic error reduction achieved by the run. Since $A_j/A_{j+1} = e^{\ell_j}$, we have

$$\log\left(1 + \frac{V_j}{rA_{j+1}}\right) \le \log\left(1 + \frac{e^{\ell_j}}{2r}\right) \le \log\left(1 + \frac{1}{2r}\right) + \ell_j.$$

Combining the accepted-run and waiting-time bounds, and using $1 \le 1/r$, yields

$$\sigma_{j+1} - \sigma_j \le \frac{1}{r}\left[6 + \log\left(1 + \frac{1}{2r}\right) + \ell_j\right].$$

Because every ℓ_j is at least $\log 2$, the fixed terms can also be bounded by a constant times ℓ_j. Specifically, set

$$D_r = 1 + \frac{6 + \log(1 + 1/(2r))}{\log 2} = \frac{6 + \log(2 + 1/r)}{\log 2}.$$

Then $\sigma_{j+1} - \sigma_j \le \frac{D_r}{r}\ell_j$. Summing over successive runs gives

$$\sigma_j \le \frac{D_r}{r} \log(A_0/A_j), \qquad A_j \le A_0 e^{-r\sigma_j/D_r}.$$

Thus the error at the start of each accepted run decays exponentially in the elapsed number of calls.

It remains to control the calls between these starting times. During the accepted part of run j, before the crossing, $|x_t| \le A_j$ and $t \le \sigma_j + 5/r$. Therefore

$$|x_t| \le A_0 e^{-r\sigma_j/D_r} \le e^{5/D_r} A_0 e^{-rt/D_r}.$$

After the crossing and before the next run, the position is fixed at magnitude A_{j+1}. Since $t \le \sigma_{j+1}$, we instead have

$$|x_t| = A_{j+1} \le A_0 e^{-r\sigma_{j+1}/D_r} \le A_0 e^{-rt/D_r}.$$

These two estimates cover every call. If a run lands exactly at zero, the estimate for its preceding accepted updates still applies, and all later errors vanish.

Finally, $D_r \log 2 = 6 + \log(2 + 1/r) > 5$, so $e^{5/D_r} \le 2$. Moreover, substituting $r = 1/(4\sqrt{\kappa})$ gives

$$\frac{r}{D_r} = \frac{r \log 2}{6 + \log(2 + 1/r)} = \gamma_\kappa.$$

We have therefore proved the scalar bound

$$|x_t| \le 2e^{-\gamma_\kappa t}|x_0|$$

at every integer time, including rejected calls. Restoring the coordinate minimizer gives (C.1). Since the factor is the same for every coordinate, taking any ℓ_p norm proves Theorem C.1.

C.2 THE OPTIMAL RATE FOR LINEAR METHODS WITH PREDETERMINED COEFFICIENTS

Theorem C.1 shows that cautious momentum retains accelerated convergence under arbitrary changes in diagonal curvature. We now compare this rate with methods that combine gradient responses linearly using coefficients chosen in advance. This class contains several standard momentum methods, so the comparison goes beyond fixed-step gradient descent. Our goal is to determine its optimal worst-case error after a prescribed number of calls.

Fix a finite dimension d and a horizon $T \in \mathbb{N}_0$, and recall that $\kappa = L/\mu$. On call t, the method chooses a query point z_t and receives

$$g_t(z_t) = H_t(z_t - x^\star),$$

where H_t is diagonal with entries in $[\mu, L]$. Let $\mathcal{L}_T$ consist of methods whose query points and final output have the form

$$z_t = x_0 + \sum_{j=0}^{t-1} C_{t,j} g_j(z_j), \qquad x_T = x_0 + \sum_{j=0}^{T-1} B_{T,j} g_j(z_j). \tag{C.6}$$

The first relation applies for $0 \le t < T$. Here z_t is the point queried on call t, whereas x_T is the point returned after all T calls.

The coefficient matrices $C_{t,j}, B_{T,j} \in \mathbb{R}^{d\times d}$ are chosen before observing the gradient responses. For a fixed random seed, they may depend only on d, μ, L, T and any supplied distribution of the Hessians. The seed ω is drawn independently of the Hessians. The coefficients do not depend on the unknown minimizer or the observed gradient responses. They may vary with time and need not commute. Thus the queries and output depend on earlier gradients through a linear rule whose coefficients are fixed in advance. This form is also translation equivariant: shifting both x_0 and $x^\star$ by the same vector shifts every query and the output by that vector, while leaving the gradient responses unchanged.

With their standard initialization, gradient descent, Polyak's heavy-ball method, and Nesterov's accelerated gradient method belong to this class when their step sizes and momentum parameters are predetermined. Unrolling their recurrences expresses each query and output in the form (C.6). The class also allows fixed matrix preconditioning, linear extrapolation, and predetermined linear restarts.

For a randomized method and a fixed Hessian sequence, we measure the relative terminal error by

$$\frac{\left(\mathbb{E}_\omega \|x_T - x^\star\|_2^2\right)^{1/2}}{\|x_0 - x^\star\|_2}, \qquad x_0 \neq x^\star.$$

The expectation is over the algorithm's seed. For a deterministic method, this reduces to the relative Euclidean error. We take the worst case over predetermined diagonal Hessian sequences and nonzero initial errors. The minimax factor is the infimum of this worst-case factor over methods in $\mathcal{L}_T$.

The next theorem determines this factor exactly. For the lower bound, it suffices to consider scalar Hessians $H_t = \Lambda_t I$, with independently chosen curvatures at the two endpoints μ and L. The probabilities below are chosen so that the error produced by gradient descent is orthogonal, in expectation, to the difference between its error and the error of any other method in the class.

Theorem C.3 (Exact optimal rate for linear methods)**.** *Assume $0 < \mu < L$ and set*

$$q_\kappa = \frac{\kappa - 1}{\kappa + 1} = \frac{L - \mu}{L + \mu}.$$

Let $H_t = \Lambda_t I$, where the curvatures are independent and satisfy

$$\Pr(\Lambda_t = \mu) = \frac{L}{L+\mu}, \qquad \Pr(\Lambda_t = L) = \frac{\mu}{L+\mu}. \tag{C.7}$$

Then every method in $\mathcal{L}_T$ satisfies

$$\mathbb{E}_{\Lambda,\omega} \|x_T - x^\star\|_2^2 \ge q_\kappa^{2T} \|x_0 - x^\star\|_2^2,$$

where the expectation includes the curvatures and, for a randomized method, its independent seed.

Gradient descent with step size $2/(L+\mu)$ attains equality under this law and satisfies

$$\|x_T - x^\star\|_2 \le q_\kappa^T \|x_0 - x^\star\|_2$$

on every diagonal Hessian sequence with entries in $[\mu, L]$. Consequently, the deterministic minimax norm factor and the randomized minimax root-mean-square factor are both exactly q_κ^T. For the randomized lower bound, the Hessian sequence can be fixed independently of the algorithm's seed.

Proof. We first prove the lower bound. Condition on the algorithm's seed, so that all coefficient matrices are fixed, and write $e_0 = x_0 - x^\star$. Under (C.7), each gradient has the form

$$g_t(z_t) = \Lambda_t(z_t - x^\star).$$

The query error $z_t - x^\star$ depends only on the earlier curvatures $\Lambda_0, \dots, \Lambda_{t-1}$. Multiplication by Λ_t therefore introduces one new variable.

Induction in (C.6) shows that the output error is a vector-valued polynomial

$$x_T - x^\star = P_T(\Lambda_0, \dots, \Lambda_{T-1})$$

in which each variable has degree at most one. Its constant coefficient is e_0: every gradient term contains at least one curvature variable, while the initial error contributes e_0. These properties hold for arbitrary coefficient matrices.

We compare this polynomial with the error produced by gradient descent. Set

$$h = \frac{2}{L+\mu}, \qquad Z_t = 1 - h\Lambda_t.$$

Gradient descent multiplies its error by Z_t on each call, so its terminal error polynomial is

$$P_T^* = e_0 \prod_{t=0}^{T-1} Z_t.$$

Under (C.7), the two possible values of Z_t are q_κ and $-q_\kappa$. The chosen probabilities give

$$\mathbb{E}_\Lambda[Z_t\Lambda_t] = 0, \qquad \mathbb{E}_\Lambda Z_t^2 = q_\kappa^2.$$

The first identity is the cancellation needed for the lower bound.

Both P_T and P_T^* have constant coefficient e_0. Their difference is therefore a sum of terms

$$c_S \prod_{s\in S} \Lambda_s, \qquad \varnothing \neq S \subseteq \{0, \dots, T-1\},$$

where each c_S is a fixed vector after conditioning on the seed. For each such term, independence gives

$$\mathbb{E}_\Lambda \left\langle P_T^*, c_S \prod_{s\in S} \Lambda_s \right\rangle = \langle e_0, c_S\rangle \prod_{s\in S} \mathbb{E}_\Lambda[Z_s\Lambda_s] \prod_{\substack{0\le s<T \\ s\notin S}} \mathbb{E}_\Lambda Z_s = 0.$$

Indeed, S is nonempty, so the first product contains a zero factor. Thus P_T^* is orthogonal in expectation to $P_T - P_T^*$.

Expanding the squared norm and using this orthogonality yields

$$\mathbb{E}_\Lambda\|P_T\|_2^2 = q_\kappa^{2T}\|e_0\|_2^2 + \mathbb{E}_\Lambda\|P_T - P_T^*\|_2^2. \tag{C.8}$$

Here we used independence and $\mathbb{E}_\Lambda Z_t^2 = q_\kappa^2$ to evaluate $\mathbb{E}_\Lambda\|P_T^*\|_2^2$. Discarding the second nonnegative term and averaging over the algorithm's seed proves the stated lower bound.

This distributional bound also gives a predetermined Hessian sequence. The law (C.7) has finitely many possible curvature sequences, each independent of the seed. So at least one of them must satisfy $\mathbb{E}_\omega\|x_T - x^\star\|_2^2 \ge q_\kappa^{2T}\|e_0\|_2^2$. This gives the randomized lower bound on a sequence fixed independently of the realized seed. For a deterministic method, the same argument gives the corresponding norm lower bound.

It remains to prove the matching upper bound. For gradient descent with step $h = 2/(L+\mu)$, every admissible diagonal Hessian satisfies $\|I - hH_t\|_{\mathrm{op}} \le \max_{\lambda\in[\mu,L]} |1 - h\lambda| = q_\kappa$. Applying this bound on successive calls gives $\|x_T - x^\star\|_2 \le q_\kappa^T\|e_0\|_2$. Under (C.7), each scalar multiplier has magnitude exactly q_κ, so gradient descent also attains equality in the expected squared-error bound. This proves both minimax statements. □

Remark C.4. To interpret the comparison in terms of gradient calls, note that, for $\kappa > 1$,

$$\frac{2}{\kappa+1} \leq -\log q_\kappa = \log\left(1 + \frac{2}{\kappa-1}\right) \leq \frac{2}{\kappa-1}.$$

Hence $-\log q_\kappa = \Theta(\kappa^{-1})$ for $\kappa \geq 2$. For $0 < \varepsilon \leq 1/2$, the optimal worst-case number of calls for this linear class is therefore

$$\Theta(\kappa \log(1/\varepsilon)).$$

In the same diagonal setting, Theorem C.1 gives cautious momentum the bound

$$O\big(\sqrt{\kappa}\log(2+\kappa)\log(1/\varepsilon)\big).$$

Cautious momentum lies outside $\mathcal{L}_T$ because its update gate depends on the observed gradient and momentum.