SHARP HETEROGENEITY BARRIERS FOR ADAM: BASIS-AWARE CONVERGENCE AND HIGH-DIMENSIONAL DIVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Convergence proofs for Adam on finite sums $f = n^{-1}\sum_i f_i$ often assume strong growth, $\sum_i \|\nabla f_i\|^2 \le D_1\|\nabla f\|^2$. The constant is at least n, with equality exactly when all component gradients agree, yet the known cyclic counterexamples have $D_1 = \Omega(n^2)$. We ask how much heterogeneity Adam tolerates at fixed memory $0 \le \beta_1 \le \beta_2 < 1$, and the answer depends on the dimension and on the coordinate system. In one dimension the threshold is $D_1 = 2n$: a sharp inequality for Adam's averaged normalized update gives convergence below it, for strongly convex averages with Lipschitz component gradients, and divergent instances exist at $D_1 = 2n + O_{\beta_1,\beta_2}(\log n)$. In higher dimensions, positive-definite quadratics with a common minimizer can make Adam diverge with D_1/n arbitrarily close to one, under cyclic order, IID minibatches and random reshuffling; in the random cases the probability of divergence tends to one for distant starting points. An outward divergent ray needs $D_1/n > 1 + 4\kappa/(\kappa+1)^2$ at condition number κ, and this is exact under cyclic order. Bounded conditioning gives a dimension-independent convergence region, which is nonempty exactly when $\kappa < 3 + 2\sqrt{2}$; for $\beta_1 = \beta_2 = 0$ this limit is sharp, because at every larger κ trajectories that keep turning can diverge as $D_1/n \downarrow 1$. The threshold $2n$ returns under coordinatewise growth, for separable averages and with a single shared second moment, also under arbitrary epoch orders and IID sampling, with non-asymptotic bounds for cyclic order. Since a quadratic average is separable in its eigenbasis, a change of basis alone can turn divergence into convergence. All results allow denominator regularization and bias correction, and interval arithmetic certifies cyclic and IID divergence at the default $(0.9, 0.999)$. In a small Transformer, function-preserving rotations consistently increase test loss, and spectral bounds on Adam's frozen updates separate two bases with identical aggregate growth.

1 INTRODUCTION

Adam (Kingma & Ba, 2015) divides a momentum estimate, coordinate by coordinate, by a running root mean square of past gradients. On a finite sum $f = n^{-1}\sum_{i=1}^n f_i$ processed one component at a time, this normalization reacts to how much the component gradients disagree, and the classical divergence examples for Adam are built on exactly this effect (Reddi et al., 2018). The disagreement is usually measured by the strong growth constant (Schmidt & Le Roux, 2013; Vaswani et al., 2019), the least D_1 such that

$$\sum_{i=1}^n \|\nabla f_i(x)\|^2 \le D_1\|\nabla f(x)\|^2 \qquad \text{for all } x.$$

By Cauchy–Schwarz, $D_1 \ge n$, with equality exactly when all component gradients coincide. Zhang et al. (2022) proved that unmodified Adam converges once β_2 is close enough to one relative to n. Zhang et al. (2026) constructed divergent cyclic instances with $D_1 = \Omega(n^2)$ for memory parameters in an n-dependent region, and asked in their Remark 7 whether counterexamples exist for $n < D_1 < 2n^2$.

We study the question at fixed memory, in the form it takes for any coordinatewise optimizer: which values of the normalized ratio $C := D_1/n \ge 1$ guarantee convergence? In one dimension the answer

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.

is $C < 2$, and the same threshold holds for separable averages. Without bounds on the dimension and the conditioning, however, no value above $C = 1$ is safe. Positive-definite quadratic sums with a common minimizer make Adam diverge with C arbitrarily close to one, far below the $D_1 = \Omega(n^2)$ of the known examples (Figure 1, Table 1).

Contributions.

- **Scalar threshold** (Section 3). For a stationary stream of scalar gradients with mean μ and second moment $C\mu^2$, we prove $\operatorname{sgn}(\mu)\,\mathbb{E}[m/\sqrt{v}] \geq 2/C - 1$ for every memory pair $0 \leq \beta_1 \leq \beta_2 < 1$ and show that the bound is sharp. It gives cyclic convergence in one dimension below $D_1 = 2n$ for strongly convex averages, and streams that nearly attain it make the actual iterates diverge at $D_1 = 2n + O_{\beta_1,\beta_2}(\log n)$.
- **Divergence near minimal heterogeneity** (Section 4). Adam diverges on positive-definite quadratic sums with a common minimizer and $D_1/n \downarrow 1$, under cyclic order, IID minibatches and fresh reshuffling. The iterates escape along outward rays, and such rays need growing conditioning: at condition number κ they require $C > 1 + 4\kappa/(\kappa+1)^2$, exactly so under cyclic order. Without memory, trajectories that keep turning avoid this cost at every $\kappa > 3 + 2\sqrt{2}$, the point where our bounded-conditioning convergence region closes.
- **Geometry-aware convergence** (Section 5). Coordinatewise growth, a separable average or a shared second moment brings back the threshold $2n$ in every dimension, also under arbitrary epoch orders and IID sampling, and cyclic runs obey non-asymptotic bounds. Since a quadratic average is separable in its eigenbasis, one problem can diverge in one orthonormal basis and converge in another.

All trajectory results concern the actual Adam iterates started from zero moments, not an averaged or continuous-time surrogate. We also certify finite cyclic and IID instances at the default $(\beta_1, \beta_2) = (0.9, 0.999)$ and measure basis sensitivity in a small Transformer (Section 6).

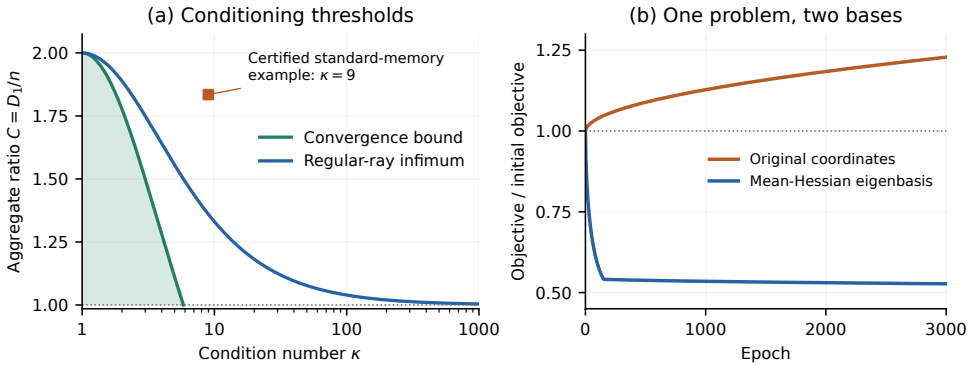


Figure 1: **(a)** Thresholds on $C = D_1/n$ against the condition number κ of the average Hessian. Cyclic Adam converges in the shaded region below the green curve $C = 8\kappa/(\kappa+1)^2$, for every $0 \leq \beta_1 \leq \beta_2 < 1$ and fixed $\epsilon > 0$ (Section 4). On or below the blue curve $C = 1 + 4\kappa/(\kappa+1)^2$ it has no regular outward ray, and attracting divergent rays exist arbitrarily close above the curve (Theorem 5). The marker is a certified divergent instance with $\kappa = 9$ at $(\beta_1, \beta_2) = (0.9, 0.999)$. **(b)** The same instance from $x_0 = 1000e$, run in two orthonormal bases (Appendix H.1). The relative objective grows in the original coordinates and decreases in an eigenbasis of the average Hessian (Corollary 1).

Two mechanisms. In one dimension the threshold comes from Adam's resemblance to sign descent (Balles & Hennig, 2018; Kunstner et al., 2023). When the component gradients arrive in long constant blocks, the normalized update follows the sign of the current gradient, so its average is close to the mean sign. A stream with mean one and second moment C can be positive on a fraction of phases as small as $1/C$, and its mean sign can then approach $2/C - 1$, which is negative exactly when $C > 2$. Theorem 1 shows that no memory pair with $\beta_1 \leq \beta_2$ does worse. In higher dimensions a different effect takes over. The growth constant weights each coordinate by its gradient energy, but at $\epsilon = 0$ coordinatewise normalization removes the scale of every coordinate, so a coordinate

Table 1: Results for fixed $0 \le \beta_1 \le \beta_2 < 1$ and $D_0 = 0$. Convergence results assume Lipschitz component gradients, a strongly convex average and diminishing steps; divergence results also allow constant steps. All rows allow either bias convention and every $\epsilon \ge 0$, except that the bounded-conditioning row takes $\epsilon > 0$ and the non-ray row takes $\beta_1 = \beta_2 = 0$. The convergence rows extend to arbitrary epoch orders and IID sampling (Theorem 8). $D_\square$ is defined in Section 5.

Setting	Heterogeneity	Outcome
One dimension	$D_1 < 2n$	converges (Thm. 2)
	$D_1 = 2n + O_{\beta_1,\beta_2}(\log n)$	diverges (Thm. 3)
General dimension	$D_1/n \downarrow 1$, $\kappa \to \infty$	diverges: cyclic, IID, reshuffling (Thm. 4)
Condition number κ	$C \le 1 + 4\kappa/(\kappa+1)^2$	no outward ray of U_0; sharp for cyclic (Thm. 5)
	$C < 8\kappa/(\kappa+1)^2$	converges (Thm. 21)
	$\kappa > 3 + 2\sqrt{2}$, $C \downarrow 1$	diverges, not along a ray (Thm. 6)
Coordinatewise growth	$D_\square < 2n$	converges in every dimension (Thm. 7)
Separable average	$D_1 < 2n$	converges; $2n$ sharp (Thm. 7)
Shared second moment	$D_1 < 2n$	converges in every basis (Section 5)

with tiny gradients still moves by a normalized step of order one. Many such coordinates with strongly disagreeing gradients hardly change D_1, yet together they can dominate the update. Our ray constructions use weak curvature to keep these coordinates large while their gradients stay small, and this is where conditioning enters.

2 Setup

We consider $f(x) = n^{-1}\sum_{i=1}^n f_i(x)$ on $\mathbb{R}^p$ under the growth condition

$$\sum_{i=1}^n \|\nabla f_i(x)\|^2 \le D_1\|\nabla f(x)\|^2 + D_0 \qquad \text{for every } x. \tag{1}$$

With $D_0 = 0$ this is the strong growth condition used in interpolation analyses (Vaswani et al., 2019); a positive D_0 allows convergence to a neighborhood instead of the minimizer. We call the least such D_1 *optimal* and write $C = D_1/n$. For strongly convex f, Cauchy–Schwarz gives $D_1 \ge n$ even when $D_0 > 0$. For quadratic components $f_i(x) = \frac{1}{2}x^\top H_i x$ we write $A = n^{-1}\sum_i H_i$ for the average Hessian and $\kappa = \mathrm{cond}_2(A)$ for its condition number.

Adam. We write $q = \beta_1$ and $b = \beta_2$. For a scalar input stream (d_t) the moment estimates are

$$m_t = q\, m_{t-1} + (1-q)\, d_t, \qquad v_t = b\, v_{t-1} + (1-b)\, d_t^2, \tag{2}$$

and throughout we assume that the memory pair is *admissible*:

$$0 \le q \le b < 1. \tag{3}$$

This covers the default $(0.9, 0.999)$; for $q > b$ the stationary bound of Section 3 can fail (Appendix A.4). The *cyclic protocol* starts from x_0 with $m_0 = v_0 = 0$. At sample $t \ge 1$ it evaluates $G_t = \nabla f_{i_t}(x_{t-1})$ with $i_t = 1 + ((t-1) \bmod n)$, applies (2) to each coordinate of G_t, and sets $x_t = x_{t-1} - \gamma_t\, \hat m_t \oslash (\sqrt{\hat v_t} + \epsilon\mathbf{1})$, where $\epsilon \ge 0$ is fixed, $(\hat m_t, \hat v_t)$ are the raw or the bias-corrected moments, and $0/0 := 0$. Every sample in epoch k uses the step $\gamma_t = \eta_k/n$, so an epoch moves the iterate by $O(\eta_k)$. Our convergence results use *diminishing* scales ($\eta_k > 0$, $\eta_k \to 0$, $\sum_k \eta_k = \infty$), such as $\eta_0(k+1)^{-\alpha}$ with $0 < \alpha \le 1$. Our divergence results hold for every *nonsummable* schedule, that is, every positive nonincreasing (η_k) with $\sum_k \eta_k = \infty$, constant schedules included. Under random sampling, an *IID block* averages s component gradients drawn independently and uniformly and takes one step η_k; a *reshuffling block* visits all components in a fresh uniformly random order, again with sample step η_k/n.

Frozen fields and rays. Freeze the iterate at x and cycle through the component gradients forever. The moments become periodic, with stationary values $(m_i^*(x), v_i^*(x))$, and their average normalized

update is the *frozen field*

$$U_\epsilon(x) = \frac{1}{n}\sum_{i=1}^{n} m_i^*(x) \oslash \big(\sqrt{v_i^*(x)} + \epsilon \mathbf{1}\big), \qquad U = U_0.$$

For random protocols U is the stationary expected block direction, and for periodic streams $\mathbb{E}$ is the uniform average over phases. A direction e of a quadratic finite sum is a *regular outward ray* if $U(e) = -ae$ for some $a > 0$ and every coordinate of every $H_i e$ is nonzero; the frozen step $-U(e)$ then points away from the minimizer. The ray is *attracting* if the induced angular dynamics is linearly stable at e (Appendix C.1). At $\epsilon = 0$ quadratic problems are homogeneous: the trajectory from rx_0 with scales η_k is r times the trajectory from x_0 with scales η_k/r. Divergence from all sufficiently distant points on a ray therefore means that, from a fixed starting point, shrinking the learning rate does not restore convergence, and at large radius a fixed $\epsilon > 0$ plays no role.

3 The sharp scalar barrier

For a stationary scalar stream with $\mu = \mathbb{E}d_t \neq 0$ we set

$$C = \frac{\mathbb{E}d_t^2}{\mu^2}, \qquad J = \operatorname{sgn}(\mu)\,\mathbb{E}\frac{m_t}{\sqrt{v_t}}.$$

Positive J means that on average Adam steps against the mean gradient, that is, downhill.

Theorem 1 (Sharp stationary drift inequality)**.** *Let (d_t) be strictly stationary with finite second moment, let (m_t, v_t) be the stationary solution of (2) for an admissible pair, and assume that $v_t > 0$ almost surely and $\mathbb{E}|\log v_t| < \infty$. Then*

$$J \geq \frac{2}{C} - 1. \tag{4}$$

For every admissible pair and every $C \geq 1$, this bound is the exact infimum over periodic streams, and streams of period n come within $O_{C,\bar b}(\log n/n)$ of it, uniformly over $0 \leq q \leq b \leq \bar b < 1$. For every finite $\epsilon \geq 0$, $C < 2$ still implies $\operatorname{sgn}(\mu)\,\mathbb{E}[m_t/(\sqrt{v_t} + \epsilon)] > 0$.

Proof idea. Normalize $\mu = 1$. Two moment bounds carry the argument. First, the exponential moving average r of d with parameter b satisfies $r^2 \leq v$ by Jensen, so $1 \leq \mathbb{E}|r| \leq a := \mathbb{E}\sqrt{v} \leq \sqrt{C}$. Second, $\mathbb{E}[m^2/v] \leq 1$. To prove it, let u be the moving average of d^2 with parameter q, so that $m^2 \leq u$, and compare u with v: with $\vartheta = (b-q)/(1-b) \geq 0$, eliminating d_t^2 gives $u_t - v_t = \vartheta\big(v_t - (1-q)\sum_{j\geq 1} q^{j-1} v_{t-j}\big)$. Stationarity and Jensen give $\mathbb{E}[v_{t-j}/v_t] \geq \exp(\mathbb{E}\log v_{t-j} - \mathbb{E}\log v_t) = 1$, hence $\mathbb{E}[(u-v)/v] \leq 0$ and

$$\mathbb{E}[m^2/v] \leq \mathbb{E}[u/v] \leq 1. \tag{5}$$

Admissibility enters only through $\vartheta \geq 0$. Now $y = m/\sqrt{v}$ and $z = \sqrt{v}$ satisfy $\mathbb{E}[yz] = 1$, $\mathbb{E}z = a$, $\mathbb{E}z^2 = C$ and $\mathbb{E}y^2 \leq 1$, and the Cauchy–Schwarz inequality for their covariance gives

$$1 \leq Ja + \sqrt{(1-J^2)(C-a^2)}. \tag{6}$$

If $Ja \geq 1$, then $J \geq 1/\sqrt{C} \geq 2/C - 1$. Otherwise, solving (6) gives $J \geq [a - \sqrt{(C-1)(C-a^2)}]/C$, which increases with a and equals $2/C - 1$ at $a = 1$. Two long constant blocks, on which $m/\sqrt{v} \approx \operatorname{sgn}(d)$, attain the bound (Appendix A.1). □

At a point of a scalar problem with $f'(x) \neq 0$, the ratios $d_i = f_i'(x)/f'(x)$ form a period with mean one and, when $D_0 = 0$, second moment at most D_1/n. So $D_1 < 2n$ makes the frozen field point toward the minimizer. The next theorem shows that the actual iterates follow it.

Theorem 2 (Cyclic convergence below $2n$)**.** *Let f be one-dimensional and strongly convex, let every f_i' be globally Lipschitz, and let (1) hold with $D_1 < 2n$ and $D_0 \geq 0$. For every admissible pair, cyclic Adam with diminishing steps has a bounded trajectory and*

$$\limsup_{t\to\infty} |f'(x_t)| \leq R := \sqrt{\frac{D_0}{2n - D_1}}. \tag{7}$$

In particular, $D_0 = 0$ gives convergence to the unique minimizer.

The components need not be convex; only the average must be. Under vanishing steps the moments track their frozen values, and trapping turns the frozen margin into convergence (Lemmas 4 and 5).

Theorem 3 (Divergence just above $2n$)**.** *Fix an admissible pair, $\lambda > 0$ and $\epsilon \geq 0$. For every sufficiently large even n there is a one-dimensional quadratic finite sum with average exactly $f(x) = \lambda x^2/2$ and optimal*

$$D_1 = 2n + O_{q,b}(\log n) > 2n, \tag{8}$$

on which cyclic Adam satisfies $x_t \to +\infty$ from every sufficiently large positive initialization, for every nonsummable schedule and either bias convention.

The instance realizes the extremal case of Theorem 1 with components $f_i(x) = \lambda d_i x^2/2$. The first $n/2 - r_n$ coefficients are equal and positive, their common value fixed by the mean, and the remaining $n/2 + r_n$ equal $-1/n$. Over a period the input signs average to $-2r_n/n$, and only $O_{q,b}(\log n)$ phases near the two block boundaries remember the other block. An integer $r_n = O_{q,b}(\log n)$ large enough to absorb these transients gives $J_n \leq -r_n/(2n) < 0$ and $0 < D_1 - 2n \leq 8r_n + 7$ (Appendix A.3). A scalar problem has no transverse direction to control, so moment tracking alone turns this negative drift into divergence of the actual iterates. Some components here are concave. For $b > 0$, Section 5.2 lifts this instance to positive-definite components with average Hessian I and the same D_1.

Finite n and the role of β_2. At a fixed period the threshold is larger. For $n \geq 2$ and $b > 0$ we give an explicit $C_{\mathrm{fin}}(n, b) > 2$, which tends to 2 as $n \to \infty$ and to infinity as $b \uparrow 1$; for example, $C_{\mathrm{fin}}(1000, 0.999) \approx 49.15$ (Appendix A.2). This quantifies the large-β_2 regime of Zhang et al. (2022): when the period is fixed, a long second-moment memory tolerates much more heterogeneity.

4 High-dimensional divergence and the cost of conditioning

In one dimension the sign of the mean gradient decides where the iterates go. In higher dimensions coordinatewise normalization can push them outward along a ray even when D_1/n is close to one.

Theorem 4 (Divergence with $D_1/n \downarrow 1$)**.** *Fix an admissible pair, $\epsilon \geq 0$, either bias convention and a nonsummable schedule. For each protocol (cyclic order, IID minibatches of any fixed size, or fresh uniform reshuffling) there are positive-definite quadratic finite sums with a common minimizer, fixed n, dimension $p \to \infty$ and optimal $D_1/n \downarrow 1$, each with a ray e, such that Adam started from zero moments at $x_0 = r_0 e$ behaves as follows. Under cyclic order, $\|x_t\| \to \infty$ for all sufficiently large r_0; under random sampling, the probability that $\|x_t\| \to \infty$ tends to one as $r_0 \to \infty$, uniformly for initial directions near e. The average Hessian is nonseparable, and its condition number grows with p.*

Construction. The construction makes the mechanism of Section 1 exact. The parameter has M tail coordinates and one head coordinate e_0, and the ray is $e = c\mathbf{1}_M - e_0$. Each component is a rank-two perturbation $H_i = A + d_i v^\top + v d_i^\top$ of a fixed average A with $\sum_i d_i = 0$, $v^\top e = 1$ and $d_i \perp e$, so the average Hessian is exactly A and $H_i e = Ae + d_i$. The tail is divided into n phase groups. Along the ray, group r receives the r-th cyclic shift of a scalar period a/M with negative stationary drift, and the head receives a constant positive input. This gives $U(e) = -e$, so every tail coordinate moves outward by a normalized step of order one, although the tail carries only a fraction $O(1/M)$ of the gradient energy and $D_1/n = 1 + O(1/M)$ (Appendix E). For large M all transverse modes of the angular dynamics are stable, and nearby directions are attracted to the ray. Under random sampling the scalar stream carries its positive mean on rare large inputs, while its frequent negative inputs keep the normalized drift negative (Appendix E.3).

Because $n^{-1} \sum_i \|\nabla f_i - \nabla f\|^2 \leq (C-1)\|\nabla f\|^2$, the relative gradient variance in these examples is arbitrarily small. At the other end, $D_1 = n$ with $D_0 = 0$ forces all component gradients to coincide, and Theorem 7 applies. Without bounds on the dimension and the conditioning, $D_1/n = 1$ is therefore the only value that guarantees convergence.

Matched aggregate statistics. Take a cyclic instance of Theorem 4 with $C < 2$ and replace its components by $\tilde{H}_i = a_i A$, where the positive scalars a_i have mean one and mean square C. The new problem has the same n, p, A and optimal D_1, but its coordinatewise constant is $nC < 2n$, so cyclic Adam converges from every start (Theorem 7, Corollary 4). Two problems that agree in n, p, A and D_1 can therefore behave oppositely under identical settings; Table 2 contains a certified pair.

The cost of conditioning. The condition number of A in Theorem 4 grows without bound. Theorem 5 shows that outward rays cannot avoid this and gives the exact trade-off under cyclic order.

Theorem 5 (Sharp conditioning threshold for divergent rays)**.** *Fix an admissible pair and $\kappa > 1$. Among positive-definite quadratic finite sums with a common minimizer and $\mathrm{cond}_2(A) = \kappa$, with n and p free, the infimum of $C = D_1/n$ over instances with a regular outward ray is*

$$C_{\mathrm{ray}}(\kappa) = 1 + \frac{4\kappa}{(\kappa+1)^2}, \tag{9}$$

and every such instance has $C > C_{\mathrm{ray}}(\kappa)$. Conversely, for every $\bar{C} > C_{\mathrm{ray}}(\kappa)$ there is a finite instance with $C < \min(\bar{C}, 2)$ and an attracting regular outward ray. For every nonsummable schedule, fixed $\epsilon \geq 0$ and bias convention, cyclic Adam from zero moments at every sufficiently distant point on this ray satisfies $\|x_t\| \to \infty$.

Proof idea. Let $g = Ae$, $E_i = H_i - A$ and $P = \{j : e_j g_j > 0\}$. For $j \in P$ a step against g_j moves coordinate j toward the minimizer, so an outward ray needs the normalized drift in coordinate j to oppose its mean g_j. Theorem 1 then forces $n^{-1}\sum_i (H_i e)_j^2 > 2g_j^2$, that is, $n^{-1}\sum_i (E_i e)_j^2 > g_j^2$. Summing over P and using $0 < e^\top g \leq \sum_{j\in P} e_j g_j \leq \|e\|\|g_P\|$ gives

$$C - 1 \;\geq\; \frac{n^{-1}\sum_i \|E_i e\|^2}{\|g\|^2} \;>\; \frac{\|g_P\|^2}{\|g\|^2} \;\geq\; \frac{(e^\top A e)^2}{\|e\|^2\|Ae\|^2} \;\geq\; \frac{4\kappa}{(\kappa+1)^2},$$

where the first inequality evaluates the growth operator at e and the last is the sharp bound on the angle between e and Ae. The matching construction orients a two-dimensional head–tail block of A to attain this angle bound and reuses the rank-two perturbations above. Repeating each component in long constant blocks drives the angular Jacobian to $-I$, making the ray attracting (Appendix F). □

Equivalently, a regular outward ray of cyclic Adam with $C = 1+\delta < 2$ needs $\kappa > (1+\sqrt{1-\delta})^2/\delta = 4/\delta - 2 + O(\delta)$. Near minimal heterogeneity, divergence along rays thus requires a condition number of order $1/\delta$. The same lower bound holds for every outward ray under reshuffling and IID sampling, and minibatches of size s raise it to $C > 1 + 4s\kappa/(\kappa+1)^2$ (Appendix F.1). Such rays also need dimension $p \gtrsim 1/(C-1)$ (Proposition 7); the construction of Theorem 4 attains both orders.

From frozen rays to trajectories. Over an epoch the moments track their frozen values, and a joint estimate of radius and angle in $h_k = \eta_k/r_k$ keeps the iterates in the attracting cone while r_k grows like $a\sum_{j<k}\eta_j$. The cyclic transfer (Theorem 17) applies from $r_0 \geq K_*(\eta_0 + \epsilon)$, and the stochastic version gives escape probability at least $1 - \exp(-c_* r_0/\eta_0)$ (Theorem 19).

A dimension-independent convergence region. Bounded conditioning also gives convergence, with neither separability nor convex components. Suppose every component gradient is globally Lipschitz, f is C^2 with $mI \preceq \nabla^2 f \preceq LI$ everywhere, $\kappa = L/m$, and (1) holds with $D_0 = 0$ and $D_1/n < 8\kappa/(\kappa+1)^2$. Then cyclic Adam with diminishing steps converges to the minimizer for every admissible pair and fixed $\epsilon > 0$, and also at $\epsilon = 0$ when $q = b$ (Theorem 21); for $D_0 > 0$, Corollary 6 bounds the residual neighborhood. The proof pairs the frozen field with the cubic potential $\sum_j |x_j - x_{*,j}|^3$ plus an ϵ-dependent quadratic. The region does not depend on the dimension and is nonempty exactly when $\kappa < 3 + 2\sqrt{2}$; at $\kappa = 1$ both thresholds equal 2 (Figure 1a).

Sharpness beyond rays. Once $\kappa \geq 3 + 2\sqrt{2}$ the region is empty, and without memory this limit is sharp, because trajectories that keep turning avoid the conditioning cost of Theorem 5.

Theorem 6 (Non-ray divergence at fixed conditioning)**.** *Let $\beta_1 = \beta_2 = 0$, $\epsilon \geq 0$, $\kappa > 3 + 2\sqrt{2}$ and $\delta > 0$. There are a positive-definite quadratic finite sum with a common minimizer, $\mathrm{cond}_2(A) = \kappa$ and $1 < C < 1 + \delta$, a vector $\bar{x}$ and a constant R_0 such that, for every nonsummable schedule, cyclic Adam from $x_0 = R\bar{x}$ with $R \geq R_0(\sup_k \eta_k + \epsilon)$ satisfies $\|x_t\| \to \infty$, while $x_t/\|x_t\|$ has no limit.*

So without memory a dimension-independent margin above $C = 1$ exists for every $\kappa < 3 + 2\sqrt{2}$ and for no $\kappa > 3 + 2\sqrt{2}$ (Corollary 5). The construction replaces the fixed escape direction by an expanding path that keeps turning. Two-dimensional sign clocks decide when the path turns,

and finite gradient profiles implement the turns through their medians while adding arbitrarily little covariance. A positive-definite embedding and a bounded-error tracking argument then make the actual iterates diverge (Appendix F.5).

5 Convergence from coordinate geometry

Aggregate growth controls the total gradient energy but says nothing about how it is spread over coordinates. Controlling the spread directly, or through a separable average, restores the threshold $2n$ in every dimension. Let $D_\square$ be the least constant such that

$$\sum_{i=1}^{n} \big(\partial_j f_i(x)\big)^2 \le D_\square \big(\partial_j f(x)\big)^2 \qquad \text{for every } x \text{ and every coordinate } j. \tag{10}$$

Summing over j gives (1) with $D_0 = 0$ and $D_1 \le D_\square$.

Theorem 7 (Geometry-aware convergence). *Let every ∇f_i be globally Lipschitz and let f be strongly convex. Suppose either (i) (10) holds with $D_\square < 2n$, or (ii) $f(x) = \sum_j F_j(x_j)$ and (1) holds with $D_0 = 0$, $D_1 < 2n$. For every admissible pair, cyclic Adam with diminishing steps converges to the unique minimizer in every fixed dimension. In both cases the threshold $2n$ is asymptotically sharp.*

5.1 Coordinatewise growth

Part (i) applies the scalar margin to each coordinate separately (Appendix B.3). Coordinates with large gradients keep second moments bounded away from zero, so their updates follow the frozen ones, and coordinates with small gradients carry little weight. Since $D_\square = D_1$ in one dimension, Theorem 3 shows that the constant 2 is sharp when components may be nonconvex. The divergent families of Section 4 have D_1/n close to one but $D_\square = \infty$.

5.2 Aggregate growth for separable averages

With only aggregate growth, a separable average offers a second route, because the magnitude-weighted gradient $g \odot |g|$ is then itself a gradient. For a stationary vector stream G write $\mu = \mathbb{E}G \neq 0$ and $C = \mathbb{E}\|G\|_2^2 / \|\mu\|_2^2$.

Proposition 1 (Magnitude-weighted pairings). *Assume the hypotheses of Theorem 1 in every coordinate with $\mu_j \neq 0$, and let $U_\epsilon = \mathbb{E}[m \oslash (\sqrt{v} + \epsilon \mathbf{1})]$. Then*

$$(\mu \odot |\mu|)^\top U_0 \;\ge\; \Big(\frac{2}{C} - 1\Big)\|\mu\|_2^2, \tag{11}$$

with a constant that is sharp already in one dimension, and for every finite $\epsilon \ge 0$,

$$\big\langle \mu \odot (|\mu| + \epsilon\mathbf{1}),\, U_\epsilon \big\rangle \;\ge\; \|\mu\|_2^2 - \tfrac{1}{2}\,\mathbb{E}\|G\|_2^2. \tag{12}$$

So aggregate $C < 2$ gives a positive pairing even when some coordinate ratios exceed two. Theorem 7(ii) needs only the average to be separable; the components may be coupled. With $g = \nabla f$ the proof uses the coercive potential

$$W_\epsilon(x) = \sum_{j=1}^{p} \int_{x_{*,j}}^{x_j} F_j'(s)\,|F_j'(s)|\,ds + \epsilon\big(f(x) - f(x_*)\big), \qquad \nabla W_\epsilon = g \odot |g| + \epsilon g.$$

By (12) its pairing with the frozen field is at least $\frac{1}{2}(2 - D_1/n)\|g\|^2$, and tracking and trapping transfer this margin to the iterates. With $D_0 > 0$ the iterates approach a bounded residual set (Theorem 11).

Sharpness. Theorem 3 already shows that $2n$ cannot be improved in (ii). For $b > 0$ the same holds with positive-definite components whose average Hessian is I. Let $g_i \in \mathbb{R}^n$ be the cyclic shifts of a mean-one scalar period, $Q_i = g_i g_i^\top / n$ and $K = n^{-1}\sum_i Q_i$. The lift $H_i = I - K + Q_i$ has average exactly I and satisfies $H_i \mathbf{1} = g_i$, so along the ray $\mathbf{1}$ each coordinate sees a shifted copy of the scalar period and inherits its drift. Lifting the scalar instance of Theorem 3 keeps $D_1 = 2n + O_{q,b}(\log n)$ and makes every H_i positive definite with condition number below 52; cyclic Adam diverges on the lift for every admissible pair with $b > 0$ (Theorem 18).

Table 2: Certified finite instances. Directed-interval arithmetic at two precisions verifies the stationary drift, positive definiteness, the optimal D_1 and the linear stability of all transverse modes, so the transfer theorem applies from zero moments at distant points on the ray. The twins share n, p, A, D_1, initialization and schedule. Values are rounded; appendices give exact values or enclosures.

Family (App.)	(q, b)	n	p	κ	$C = D_1/n$	Outcome
Twins (E.4)	$(.9, .999)$	20 000	400 001	50 000	1.00133919	div. / conv.
Identity lift (D.2)	$(.9, .999)$	80 000	80	1	2.380135	diverges
Fixed κ (F.4)	$(.9, .999)$	75 000	4 901	9	1.834814	diverges
IID (E.3)	$(.9, .999)$	32 768	491 521	$< 7.88 \cdot 10^8$	< 1.92654	div. w.p. $\to 1$

Corollary 1 (One problem, opposite outcomes in two bases)**.** *Fix an admissible pair, $\kappa > 1$, finite $\epsilon \geq 0$, a bias convention and a nonincreasing diminishing schedule. There are positive-definite quadratic finite sums with a common minimizer, $C < 2$ and $\mathrm{cond}_2(A) = \kappa$, and a sufficiently distant common initial point, such that zero-moment cyclic Adam diverges in the original basis and converges in an eigenbasis of A. Every component spectrum and the optimal D_1 are preserved.*

For the proof, take a divergent instance of Theorem 5 with $C < 2$ and rotate to an eigenbasis of A. The average becomes separable, and Theorem 7(ii) applies.

5.3 A SHARED SECOND MOMENT

If the coordinatewise second moment is replaced by one scalar $s_t = bs_{t-1} + (1-b)\|G_t\|^2$ and the update by $x_t = x_{t-1} - \gamma_t m_t/(\sqrt{s_t} + \epsilon)$, Adam becomes rotation equivariant. Its stationary pairing then satisfies $\langle \mu/\|\mu\|, \mathbb{E}[m/\sqrt{s}]\rangle \geq 2/C - 1$ with a sharp constant, and $D_1 \leq C_0 n < 2n$ gives cyclic convergence for every strongly convex average in every dimension (Theorems 20 and 12). The failures of Section 4 thus come from normalizing each coordinate separately.

5.4 SAMPLING ORDERS AND RATES

Theorem 8 (Other sampling orders and rates)**.** *Let $D_0 = 0$ and fix an admissible pair and $\epsilon \geq 0$. Assume the hypotheses of Theorem 2, of Theorem 7, of the shared-moment guarantee, or of Theorem 21, in the last case with the growth bound replaced by (112) if $\epsilon = 0$.*

(i) If also $\eta_{k+1}/\eta_k \to 1$, Adam converges under arbitrary, possibly history-dependent, epoch permutations, including random reshuffling.
(ii) With IID component sampling and sample steps $\gamma_t = a(t+1)^{-\alpha}$, $a > 0$, $0 < \alpha \leq 1$, Adam converges almost surely from every initial point.
(iii) Cyclic runs with $\eta_k \propto (k+1)^{-1/2}$ satisfy $\min_{k<K}(f(x_{kn}) - f_*) = O(\log K/\sqrt{K})$ in all four settings if $\epsilon > 0$, and in the first three if $\epsilon = 0$.

Appendix B proves (i)–(iii) without assuming that the iterates stay bounded (Theorems 13–16).

6 CERTIFIED INSTANCES AND EXPERIMENTS

Certificates. Each certificate in Table 2 covers every transverse mode, including 4 900 at $\kappa = 9$ and 491 520 for IID sampling. Figure 1b runs the $\kappa = 9$ instance in two bases; it moves outward in the original coordinates and toward the minimizer in an eigenbasis of A. After 1000 epochs its final-to-initial objective ratio is 1.128, against 0.101 for its twin with equal n, p, A and D_1 (Appendix H.1). From $x_0 = 1000e$ with $\eta_k = (k+1)^{-1/2}$ and the Transformer's global-norm clipping at one, its radius after 3000 epochs is 1108.09, as without clipping; with clipping and decoupled decay $\lambda = 0.01$ at constant step 0.1, the radius falls to 102.2 by epoch 6000, near $a/\lambda = 100$ (Appendices H.1 and C.2).

Basis sensitivity in a Transformer. We reparametrize a two-layer character-level Transformer with 65,280 parameters as $\theta = \theta_0 + Qz$, with orthogonal Q and $z_0 = 0$, so that both bases start from the

same function. Both arms train on Tiny Shakespeare (Karpathy, 2015) with AdamW (Loshchilov & Hutter, 2019) for 1000 updates and share batches, clipping, physical weight decay and a learning rate tuned on separate development seeds. The global rotation composes a permutation with orthogonal 64×64 blocks (Zhang et al., 2025b; Xie et al., 2025; Singh, 2026). It increases test NLL in all six pairs, by 0.0427 nats per character on average (paired 95% CI $[0.0286, 0.0568]$), a 4.4% rise in perplexity, and an independent six-pair cohort gives 0.0420, again positive in every pair (Figure 2). Rotations confined to single tensors also raise NLL in every pair, by 0.0150 on average, while momentum SGD and a shared-second-moment control agree across the two bases to roundoff.

Coordinate statistics at trained states. At each endpoint, the finite-period estimates of Section 3, evaluated on coordinate moments and gradient spectra of one minibatch bank, bound the frozen work $\mu^\top \mathbb{E}[m \oslash (\sqrt{v} + \epsilon \mathbf{1})]$ within 3% (Proposition 8). The aggregate ratio is identical in the two bases, but the work bounds are disjoint in all twelve same-state pairs (Appendix H.3), so coordinatewise quantities detect a basis dependence that aggregate growth misses. Even IID batches 7–21 times larger, which bring the aggregate ratio below 2, leave the median coordinate ratio at 3.0–6.4.

7 Related work

Convergence of Adam. Earlier work studies online counterexamples (Reddi et al., 2018; Heilman & Mohanty, 2026), oscillating effective step sizes (Chen et al., 2019), memory-dependent convergence of RMSprop and Adam (Zou et al., 2019; Shi et al., 2021; Zhang et al., 2022; 2026), modified updates such as ADOPT (Taniguchi et al., 2024), and convergence under bounded gradients or noise, relaxed smoothness, or growth conditions (Défossez et al., 2022; Li et al., 2023; Wang et al., 2023; Zhang et al., 2025a; Jin et al., 2025). ODE analyses treat joint small-step/long-memory limits (Barakat & Bianchi, 2021) and fixed-memory averaged IID dynamics (Dereich et al., 2025b); there, zeros of the averaged field can differ from critical points of the objective even on strongly convex quadratics (Dereich & Jentzen, 2024), and data symmetry decides convergence in a two-point model (Dereich et al., 2025a). We instead carry protocol-specific stationary averages over to the actual iterates.

Divergence. Bock & Weiß (2019) exhibit constant-step limit cycles on a strongly convex quadratic, Toint (2025) a fixed-step full-batch divergence, and Dereich et al. (2024) nonconvergence for non-vanishing steps. Wang & Klabjan (2022, Theorem 3.1) prove that Adam can fail to approach the optimal value in expectation on strongly convex stochastic quadratics whose components have different minimizers. Our random constructions have positive-definite components, a common minimizer and vanishing relative gradient variance, and still diverge with probability tending to one. Per-sample implicit bias (Baek et al., 2026), order effects (Sweeney, 2026) and loss spikes (Molybog et al., 2023) are related phenomena.

Geometry and rotation dependence. Adam's sensitivity to rotations and its relation to Transformer Hessians have been studied empirically and theoretically (Zhang et al., 2024; 2025b; Xie et al., 2025; Singh, 2026). VectorAdam restores equivariance within coordinate groups (Ling et al., 2022), and SOAP runs Adam in the eigenbasis of a changing preconditioner (Vyas et al., 2025). Coordinatewise strong growth is the zero-noise case of coordinatewise affine noise (Zhang et al., 2025a), and related assumptions appear for sign methods and adaptive batch sizes (Safaryan & Richtárik, 2021; Lau et al., 2025; 2026). We add divergence as $C \downarrow 1$, exact ray thresholds, the sharp conditioning limit of dimension-independent memoryless convergence, and opposite outcomes in two bases.

8 Conclusion

Below $D_1 = 2n$, Adam converges in one dimension, for separable averages and with a shared second moment, under cyclic order, arbitrary epoch orders and IID sampling. In general dimension coordinatewise normalization amplifies small but disagreeing gradients, and Adam can diverge as $D_1/n \downarrow 1$. Divergent rays pay an exact price in conditioning, and without memory, trajectories that keep turning avoid it beyond $\kappa = 3 + 2\sqrt{2}$. Coordinatewise growth restores convergence, and a change of basis alone can reverse the outcome. Heterogeneity assumptions for Adam should therefore encode conditioning or coordinate geometry.

AI USE STATEMENT

Under the author's direction, generative AI tools assisted with selected theoretical refinements and proofs, including the scalar logarithmic-overhead construction and residual bounds. They also assisted with hypothesis refinement and experimental design, implementation of numerical methods, synthetic tests and data-processing code, analysis of recorded results, and manuscript, figure and reference preparation. The author takes responsibility for the final content and accompanying materials.

REPRODUCIBILITY STATEMENT

Complete proofs appear in Appendices A–H, with a roadmap in Table 3. The supplement provides the finite-construction inputs and certificate commands (`CERTIFICATE_CLAIM_INDEX.md`), trajectory scripts and records (`MANUSCRIPT_MAP.md`), and the Transformer code, data splits, seeds, coordinate statistics and per-seed results (Appendices H.2–H.3).

REFERENCES

B. Baek, M. Song, and C. Yun. Implicit bias of per-sample Adam on separable data: Departure from the full-batch regime. In *ICLR*, 2026.

L. Balles and P. Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *ICML*, pp. 404–413, 2018.

A. Barakat and P. Bianchi. Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1):244–274, 2021.

S. Bock and M. G. Weiß. Non-convergence and limit cycles in the Adam optimizer. In *ICANN 2019: Deep Learning*, volume 11728 of *LNCS*, pp. 232–243, 2019.

X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *ICLR*, 2019.

A. Défossez, L. Bottou, F. Bach, and N. Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

S. Dereich and A. Jentzen. Convergence rates for the Adam optimizer. *arXiv:2407.21078*, 2024.

S. Dereich, R. Graeber, and A. Jentzen. Non-convergence of Adam and other adaptive stochastic gradient descent optimization methods for non-vanishing learning rates. *arXiv:2407.08100*, 2024.

S. Dereich, T. Do, A. Jentzen, and P. von Wurstemberger. Adam symmetry theorem: Characterization of the convergence of the stochastic Adam optimizer. *arXiv:2511.06675*, 2025a.

S. Dereich, A. Jentzen, and S. Kassing. ODE approximation for the Adam algorithm: General and overparametrized setting. *arXiv:2511.04622*, 2025b.

S. Heilman and S. Mohanty. On the convergence of Adam, revisited. *arXiv:2607.03519*, 2026.

R. Jin, X. Li, Y. Yu, and B. Wang. A comprehensive framework for analyzing the convergence of Adam: Bridging the gap with SGD. In *ICML*, 2025.

A. Karpathy. char-rnn: Multi-layer recurrent neural networks for character-level language models. Tiny Shakespeare corpus, https://github.com/karpathy/char-rnn, 2015.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

F. Kunstner, J. Chen, J. W. Lavington, and M. Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *ICLR*, 2023.

T. T.-K. Lau, W. Li, C. Xu, H. Liu, and M. Kolar. Adaptive batch size schedules for distributed training of language models with data and model parallelism. In *CPAL*, volume 280 of *PMLR*, pp. 278–304, 2025.

T. T.-K. Lau, H. Liu, and M. Kolar. AdAdaGrad: Adaptive batch size schemes for adaptive gradient methods. *Statistical Learning and Data Science*, pp. 100020, 2026. doi: 10.1016/j.slads.2026.100020.

H. Li, A. Rakhlin, and A. Jadbabaie. Convergence of Adam under relaxed assumptions. In *NeurIPS*, 2023.

S. Z. Ling, N. Sharp, and A. Jacobson. VectorAdam for rotation equivariant geometry optimization. In *NeurIPS*, 2022.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

I. Molybog, P. Albert, M. Chen, Z. DeVito, D. Esiobu, N. Goyal, P. S. Koura, S. Narang, A. Poulton, R. Silva, B. Tang, D. Liskovich, P. Xu, Y. Zhang, M. Kambadur, S. Roller, and S. Zhang. A theory on Adam instability in large-scale machine learning. *arXiv:2304.09871*, 2023.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *ICLR*, 2018.

M. Safaryan and P. Richtárik. Stochastic sign descent methods: New algorithms and better theory. In *ICML*, 2021.

M. Schmidt and N. Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv:1308.6370*, 2013.

N. Shi, D. Li, M. Hong, and R. Sun. RMSprop converges with proper hyper-parameter. In *ICLR*, 2021.

D. Singh. The loss does not see the basis, but Adam does. *arXiv:2608.05136*, 2026.

J. Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise. *arXiv:2606.29554*, 2026.

S. Taniguchi, K. Harada, G. Minegishi, Y. Oshima, S. C. Jeong, G. Nagahara, T. Iiyama, M. Suzuki, Y. Iwasawa, and Y. Matsuo. ADOPT: Modified Adam can converge with any β_2 with the optimal rate. In *NeurIPS*, 2024.

P. L. Toint. Divergence of the ADAM algorithm with fixed-stepsize: A (very) simple example. In *Mathematical Optimization for Machine Learning*, pp. 195–198. De Gruyter, 2025. doi: 10.1515/9783111376776-013.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *AISTATS*, 2019.

N. Vyas, D. Morwani, R. Zhao, I. Shapira, D. Brandfonbrener, L. Janson, and S. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *ICLR*, 2025.

B. Wang, J. Fu, H. Zhang, N. Zheng, and W. Chen. Closing the gap between the upper bound and lower bound of Adam's iteration complexity. In *NeurIPS*, 2023.

R. Wang and D. Klabjan. Divergence results and convergence of a variance reduced version of ADAM. *arXiv:2210.05607*, 2022.

S. Xie, M. A. Mohamadi, and Z. Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *ICLR*, 2025.

Q. Zhang, Y. Zhou, and S. Zou. Convergence guarantees for RMSProp and Adam in generalized-smooth non-convex optimization with affine noise variance. *Transactions on Machine Learning Research*, 2025a.

T. H. Zhang, L. Maes, A. Milligan, A. Jolicoeur-Martineau, I. Mitliagkas, D. Scieur, S. Lacoste-Julien, and C. Guille-Escuret. Understanding Adam requires better rotation dependent assumptions. In *NeurIPS*, 2025b.

Y. Zhang, C. Chen, N. Shi, R. Sun, and Z.-Q. Luo. Adam can converge without any modification on update rules. In *NeurIPS*, 2022.

Y. Zhang, C. Chen, T. Ding, Z. Li, R. Sun, and Z.-Q. Luo. Why transformers need Adam: A Hessian perspective. In *NeurIPS*, 2024.

Y. Zhang, B. Li, C. Chen, Z.-Q. Luo, and R. Sun. Adam converges without any modification on update rules. *arXiv:2603.02092*, 2026.

F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of Adam and RMSProp. In *CVPR*, pp. 11127–11135, 2019.

GUIDE TO THE PROOFS AND EXPERIMENTS

Notation. We write $C = D_1/n$ and $B = D_0/n$ for the aggregate growth constants and $D_\square$ for the coordinatewise one. In cyclic proofs, x_k denotes an epoch start and the sample step is η_k/n; random blocks are defined in Appendix E.3. For quadratics, $A = n^{-1}\sum_i H_i$, $\kappa = \mathrm{cond}_2(A)$; at a ray, $r_k = \ell(x_k)$, $\ell(e) = 1$, and $x_k/r_k = e + z_k$.

Table 3: Proof roadmap.

Result	Proof path	Page
Scalar and finite-period bounds	Moment energy → centered pairing → fixed-memory sharpness	12
Convergence, residuals and rates	Pairing and trapping; residual neighborhoods; non-asymptotic bounds; arbitrary orders and IID sampling, also at $\epsilon = 0$	16
Divergence transfer	Transverse stability → moment burn-in → joint first exit; clipping and weight decay	27
Identity-average lifts	Cyclic lift of the scalar profile → signed common shift and Fourier bounds → transfer; default-memory certificate	29
Near-minimum growth	Rank-two lift; full stability; cyclic and stochastic transfer; matched twin	31
Conditioning and dimension	Spectral-angle and dimension lower bounds → matching block construction → transfer; non-ray clocks and medians at fixed κ	38
Geometry and bounded conditioning	Two bases; shared second moment; conditioned potential; zero-stream limits	47
Experiments and certificates	Finite inputs, trajectories, Transformer pairs and spectral work bounds	50

A SCALAR AND FINITE-PERIOD INEQUALITIES

A.1 THE STATIONARY INEQUALITY AND FIXED-MEMORY SHARPNESS

Theorem 9 (Sharp stationary drift inequality, full version of Theorem 1)**.** *Let $(d_t)_{t\in\mathbb{Z}}$ be strictly stationary, with $\mu = \mathbb{E}d_t \neq 0$ and $\mathbb{E}d_t^2 < \infty$. Take the stationary solutions $m_t = qm_{t-1} + (1-q)d_t$ and $v_t = bv_{t-1} + (1-b)d_t^2$, where $0 \le q \le b < 1$. Assume $v_t > 0$ almost surely and $\mathbb{E}|\log v_t| < \infty$. With*

$$C = \frac{\mathbb{E}d_t^2}{\mu^2}, \qquad J = \mathrm{sgn}(\mu)\,\mathbb{E}\frac{m_t}{\sqrt{v_t}}, \qquad \textit{one has} \qquad J \ge \frac{2}{C} - 1.$$

For every fixed admissible (q, b) and every $C \ge 1$, this is the exact infimum over nonzero finite periodic streams with varying periods, viewed with a uniformly random phase.

Proof. Normalize $\mu = 1$ by scaling and changing the sign of the stream. Introduce $r = P_b d$ and $u = P_q(d^2)$, with P_b, P_q the stationary exponential filters. Jensen gives $r^2 \le v$, $m^2 \le u$, and hence $a := \mathbb{E}\sqrt{v} \ge \mathbb{E}|r| \ge 1$. The second ingredient is

$$\mathbb{E}(m^2/v) \le \mathbb{E}(u/v) \le 1. \tag{13}$$

Indeed, with $\Delta = u - v$ and $\vartheta = (b-q)/(1-b)$, elimination of d_t^2 and iteration of the stable recursion give

$$\Delta_t = q\Delta_{t-1} + \vartheta(v_t - v_{t-1}) = \vartheta\left(v_t - (1-q)\sum_{j\ge1} q^{j-1}v_{t-j}\right). \tag{14}$$

For each lag, stationarity and Jensen imply $\mathbb{E}(v_{t-j}/v_t) \ge \exp\{\mathbb{E}\log v_{t-j} - \mathbb{E}\log v_t\} = 1$. For a periodic stream, the same fact follows from AM–GM since the product of these ratios over a period

is one. If $b = q$, then $u = v$. If $b > q$, nonnegativity of u/v in (14) bounds $\sum_{j\ge1} q^{j-1}v_{t-j}/v_t$ by $(1+\vartheta)/[\vartheta(1-q)]$. Thus its nonnegative terms may be integrated termwise, yielding $\mathbb{E}(\Delta/v) \le 0$ and (13).

Set $y = m/\sqrt{v}$ and $z = \sqrt{v}$. Then $\mathbb{E}(yz) = 1$, $\mathbb{E}z = a$, $\mathbb{E}z^2 = C$ and $\mathbb{E}y^2 \le 1$. Centered Cauchy–Schwarz gives

$$1 \le Ja + \sqrt{(1-J^2)(C-a^2)}. \tag{15}$$

If $Ja \ge 1$, then $J \ge 1/\sqrt{C} \ge 2/C - 1$. Otherwise squaring the nonnegative sides of (15) gives $CJ^2 - 2aJ + 1 - C + a^2 \le 0$, so

$$J \ge j_-(a) := \frac{a - \sqrt{(C-1)(C-a^2)}}{C} \ge j_-(1) = \frac{2}{C} - 1.$$

The middle inequality uses $j_-'(a) > 0$ on $[1, \sqrt{C})$ and continuity at $a = \sqrt{C}$.

For sharpness, fix $C > 1$ and an integer $n \ge 4C/(C-1)$. Put $k = \lfloor n/C \rfloor + 2$, $\rho = k/n$, and use k copies of a_+ followed by $n-k$ copies of a_-, where

$$a_+ = 1 + \sqrt{\frac{(C-1)(1-\rho)}{\rho}}, \qquad a_- = 1 - \sqrt{\frac{(C-1)\rho}{1-\rho}}.$$

Their mean is one and second moment is C. Since $1/n < \rho - 1/C \le 2/n$ and $\rho \le (1+1/C)/2$, the two blocks are nonempty, $a_+ > 0 > a_-$, and their magnitudes belong to $[1/(3n), C]$; the lower bound follows by rationalizing $|a_-| = (C\rho - 1)/[(1-\rho)(|a_-| + 2)]$.

This also gives a rate uniform on compact memory regions. Fix $0 < \bar{b} < 1$, set $L_n = \lceil 4 \log n/(-\log \bar{b}) \rceil$ and $R_* = (1-\bar{b})^{-1/2}$. For $0 \le q \le b \le \bar{b}$, Lemma 2 gives $|m|/\sqrt{v} \le R_*$. At a phase preceded by L_n identical inputs c, the recursions give $|m - c| \le 2C\bar{b}^{L_n}$ and $|v - c^2| \le C^2\bar{b}^{L_n}$, whence

$$\left| \frac{m}{\sqrt{v}} - \operatorname{sgn} c \right| \le (6Cn + \tfrac{9}{2}C^2n^2)\bar{b}^{L_n} \le \frac{6C + (9/2)C^2}{n^2}.$$

There are at most $2L_n$ other phases, each with error at most $R_* + 1$. The average block sign is $2\rho - 1$, so

$$0 \le J_{q,b}^{(n)} - (2/C - 1) \le \frac{4 + 2(R_* + 1)L_n}{n} + \frac{6C + (9/2)C^2}{n^2}.$$

Thus the rate is $O_{C,\bar{b}}(\log n/n)$ uniformly on that region. For $q = b = 0$ the error is at most $4/n$ directly; for $C = 1$ a constant stream attains equality. All these streams satisfy the stated positivity and logarithmic-integrability assumptions. □

Corollary 2 (Finite-ϵ scalar sign below two)**.** *Under Theorem 9, define $J_\epsilon = \operatorname{sgn}(\mu)\mathbb{E}[m/(\sqrt{v} + \epsilon)]$ for any finite $\epsilon \ge 0$. If $C < 2$, then*

$$J_\epsilon \ge \frac{1 - \sqrt{C-1}}{\sqrt{C} + \epsilon/|\mu|} > 0. \tag{16}$$

At $q = b = 0$, zero phases are allowed, with $0/0 = 0$ when $\epsilon = 0$.

Proof. Normalize the mean and put $e = \epsilon/|\mu|$, $y = m/(\sqrt{v} + e)$, $z = \sqrt{v} + e$. Then $\mathbb{E}(yz) = 1$, $\mathbb{E}y^2 \le 1$, $\mathbb{E}z = a + e$ and $\operatorname{Var}(z) = C - a^2 \le C - 1$. Centered Cauchy–Schwarz gives $1 \le J_\epsilon(a + e) + \sqrt{C-1}$; use $a \le \sqrt{C}$. For the memoryless case, the same identities hold with $y = d/(|d| + e)$ and $z = |d| + e$, including zero phases. □

Corollary 3 (Coordinatewise stationary descent certificate)**.** *For a stationary vector stream, apply the hypotheses of Theorem 9 in each coordinate of nonzero mean μ_j, and write $C_j = \mathbb{E}d_{t,j}^2/\mu_j^2$. Then*

$$\sum_{j:\mu_j \ne 0} \mu_j \, \mathbb{E}\frac{m_{t,j}}{\sqrt{v_{t,j}}} \ge \sum_{j:\mu_j \ne 0} |\mu_j| \left(\frac{2}{C_j} - 1 \right).$$

In particular the pairing is positive when some coordinate is active and every active $C_j < 2$.

Proof. Multiply the scalar inequality by $|\mu_j|$ and sum. Coordinates of zero mean make no contribution. □

A.2 FINITE-PERIOD CONVERGENCE AND THE TWO MEMORY PARAMETERS

For $n \geq 2$, $0 < b < 1$, define

$$w_s = \frac{(1-b)b^s}{1-b^n} \quad (0 \leq s < n), \qquad \alpha_{n,b} = nw_{n-1}, \tag{17}$$

$$\gamma_{n,b} = \frac{1}{n}\left(\sum_s \sqrt{w_s}\right)^2 = \frac{1+\sqrt{b}}{n(1-\sqrt{b})}\frac{1-b^{n/2}}{1+b^{n/2}}, \tag{18}$$

$$V_{n,b}(C) = \min\{(1-\alpha_{n,b})(C-1), (1-\gamma_{n,b})C\},$$

$$C_{\mathrm{fin}}(n,b) = \max\left\{1 + \frac{1}{1-\alpha_{n,b}}, \frac{1}{1-\gamma_{n,b}}\right\} > 2. \tag{19}$$

Proposition 2 (Finite-period convergence)**.** *For an n-periodic scalar stream of mean $\mu \neq 0$, ratio C and $0 \leq q \leq b < 1$, $b > 0$, every finite $\epsilon \geq 0$ satisfies*

$$J_\epsilon \geq \frac{1-\sqrt{V_{n,b}(C)}}{\sqrt{C} + \epsilon/|\mu|} > 0 \qquad \text{if } C < C_{\mathrm{fin}}(n,b). \tag{20}$$

Consequently, in Theorem 12 one may replace $D_\square \leq C_0 n < 2n$ by $D_\square \leq C_0 n < nC_{\mathrm{fin}}(n,b)$, with all other hypotheses unchanged. For a strongly convex scalar average with globally Lipschitz component derivatives and $n^{-1}\sum_i f_i'(x)^2 \leq C_0|f'(x)|^2 + B$ globally, $1 \leq C_0 < C_{\mathrm{fin}}$, the diminishing-step protocol gives

$$\limsup_t |f'(x_t)| \leq \sqrt{\frac{B}{C_{\mathrm{fin}}(n,b) - C_0}}. \tag{21}$$

Both conclusions allow zero initial moments, either bias convention, and every fixed finite $\epsilon \geq 0$.

Proof. Normalize $\mu = 1$, and write $r = P_b d$, $v = P_b(d^2)$, $a = \mathbb{E}\sqrt{v}$. The smallest periodic weight yields

$$v_t - r_t^2 = \sum_s w_s (d_{t-s} - r_t)^2 \geq \alpha_{n,b}\{C - 1 + (1-r_t)^2\}.$$

Jensen applied to the convex function $r \mapsto \sqrt{r^2 + \alpha_{n,b}(C-1)}$ gives $a^2 \geq 1 + \alpha_{n,b}(C-1)$. Independently, $\Phi(s) = n^{-1}\sum_t \sqrt{(P_b s)_t}$ is concave on $\{s \geq 0 : \sum_i s_i = nC\}$. Its value at every vertex nCe_i is $\sqrt{\gamma_{n,b}C}$, so expressing d^2 as a convex combination of vertices gives $a^2 \geq \gamma_{n,b}C$. Thus $C - a^2 \leq V_{n,b}(C)$. Equation (13) and the centered argument in Corollary 2 give $1 \leq J_\epsilon(a + \epsilon/|\mu|) + \sqrt{V_{n,b}(C)}$. Since $V_{n,b}(C) < 1$ exactly when $C < C_{\mathrm{fin}}$, this proves the margin.

Under coordinatewise growth, each active frozen ratio is at most C_0. Replace the margin in Theorem 12 by $1 - \sqrt{V_{n,b}(C_0)} > 0$. The same absolute tracking and separation of large and small gradient coordinates apply: positive periodic weights bound the denominator on every large coordinate, and a zero-mean coordinate has zero inputs. Objective-level trapping then gives the claimed convergence.

For the additive scalar case, fix $r > \sqrt{B/(C_{\mathrm{fin}} - C_0)}$. Outside $|f'| < r$, the normalized ratio is at most $C_r = C_0 + B/r^2 < C_{\mathrm{fin}}$ and each normalized second moment is at least nw_{n-1}. Lemma 4 transfers the frozen inward margin, without a bounded-iterate premise, to a margin on all late epochs of at least $[1 - \sqrt{V_{n,b}(C_r)}]/[2(\sqrt{C_r} + \epsilon/r)]$. Nonsummable steps force entry into $[(f')^{-1}(-r), (f')^{-1}(r)]$; bounded updates and vanishing steps make subsequent overshoots $o(1)$. Decreasing r proves (21), also between epoch starts. □

Proposition 3 (First-moment refinement)**.** *Put $\beta = \alpha_{n,q}$ for $q > 0$ and $\beta = 0$ for $q = 0$. With $\alpha = \alpha_{n,b}$, define*

$$s_q = \frac{(1-q)^2}{(1-q)^2 + 4q\sin^2(\pi/n)},$$

$$H_q(C) = \min\left\{1 - \beta + \frac{\beta}{C}, \frac{1 + s_q(C-1)}{\alpha C}\right\}, \tag{22}$$

$$H_q\big(C_{\mathrm{fin}}^{(q)}\big)V_{n,b}\big(C_{\mathrm{fin}}^{(q)}\big) = 1, \qquad C_{\mathrm{fin}}^{(q)} = C_{\mathrm{fin}}^{(q)}(n,b) > 1. \tag{23}$$

The product is strictly increasing from zero to infinity on $C \geq 1$, so the last equation defines a unique threshold. All conclusions of Proposition 2 hold with $C_{\mathrm{fin}}^{(q)}$ in place of C_{fin} and margin

$$J_\epsilon \geq \frac{1 - \sqrt{H_q(C)V_{n,b}(C)}}{\sqrt{C} + \epsilon/|\mu|} > 0 \quad (C < C_{\mathrm{fin}}^{(q)}).$$

Proof. The filter $u = P_q(d^2)$ satisfies $u - m^2 \geq \beta(C-1)$, hence $\mathbb{E}(m^2/v) \leq 1 - \beta + \beta/C$ by (13) and Jensen. On zero-mean periodic vectors the largest squared Fourier multiplier of P_q is s_q; Parseval gives $\mathbb{E}m^2 \leq 1 + s_q(C-1)$, while $v_i \geq \alpha C$. This proves the second energy bound. Both remain valid with finite ϵ; centered Cauchy–Schwarz gives the margin and a nonnegative margin at equality. Each branch of $H_q V_{n,b}$ is strictly increasing, which also proves the asserted uniqueness. □

For $b = .999$, C_{fin} at $n = 100, 1000$ is approximately $4796.88, 49.15$; with $q = b$ the refined values are $1.79915 \cdot 10^7, 1117.56$. Both thresholds tend to two as $n \to \infty$ at fixed memories. At fixed n, putting $h = -\log b/4$ gives $\gamma_{n,b} = \tanh(nh)/(n \tanh h)$ and $(1-\gamma_{n,b})^{-1} \sim 48/[(n^2-1)(\log b)^2]$ as $b \uparrow 1$.

A.3 Scalar divergence with logarithmic overhead

For $b > 0$, set $M = \lceil R_{q,b} \rceil$ with $R_{q,b}$ from Lemma 2, $L_n = \lceil 6 \log n/(-\log b) \rceil$ and $r_n = 2(M+1)L_n$. For $q = b = 0$ set $r_n = 1$. Take even $n \geq 8$ with $r_n \leq n/4$, which holds for every sufficiently large even n, and define

$$p_n = \frac{1}{2} - \frac{r_n}{n}, \qquad a_n = \frac{1}{n}, \qquad B_n = \frac{1 + (1-p_n)a_n}{p_n}. \tag{24}$$

Use np_n copies of B_n followed by $n(1-p_n)$ copies of $-a_n$. Their mean is one and

$$C_n = \frac{1}{p_n} + \frac{1-p_n}{p_n}(2a_n + a_n^2), \qquad 0 < nC_n - 2n \leq 8r_n + 7. \tag{25}$$

The latter follows from $p_n \geq 1/4$ and $nC_n - 2n = [2r_n + (1-p_n)(2 + 1/n)]/p_n$.

Lemma 1 (Fixed-memory negative stationary drift)**.** *For $b > 0$, the profile (24) satisfies*

$$J_n = \frac{1}{n}\sum_{t=1}^{n} \frac{m_t}{\sqrt{v_t}} \leq -\frac{r_n}{2n} < 0. \tag{26}$$

Proof. All input magnitudes lie in $[1/n, 5]$, so $n^{-2} \leq v_t \leq 25$. At phases preceded by L_n equal inputs c, $|m_t - c| \leq 6b^{L_n}$ and $|v_t - c^2| \leq 25b^{L_n}$, giving

$$|m_t/\sqrt{v_t} - \operatorname{sgn} c| \leq (6n + \tfrac{25}{2}n^2)b^{L_n} \leq 19/n^4.$$

There are at most $2L_n$ remaining phases, with error at most $R_{q,b} + 1$. Both blocks have length at least $n/4$ and $r_n \geq 4L_n$. Their mean sign is $-2r_n/n$, so

$$J_n \leq -2r_n/n + 2L_n(R_{q,b}+1)/n + 19/n^4 \leq -r_n/n + 19/n^4 \leq -r_n/(2n).$$

The last inequality uses $n \geq 8$ and $r_n \geq 1$. □

Theorem 10 (Scalar divergence with logarithmic overhead)**.** *Fix $0 \leq q \leq b < 1$, $\lambda > 0$, finite $\epsilon \geq 0$, and positive nonincreasing epoch scales with divergent sum. For every sufficiently large even n, the scalar quadratics $f_i(x) = \lambda d_i x^2/2$ from (24) have average $\lambda x^2/2$ and optimal global heterogeneity*

$$2n < D_1 = nC_n \leq 2n + 8r_n + 7 = 2n + O_{q,b}(\log n).$$

With sample step η_k/n, zero initial moments, and either bias convention, every sufficiently large positive initialization yields $x_t \to +\infty$. At $q = b = 0$ the overhead is $O(1)$.

Proof. For $b > 0$, apply Theorem 17 with $H_i = \lambda d_i$, $e = 1$, $\ell(x) = x$. All ray inputs are nonzero, $U(e) = J_n < 0$, and the transverse space is $\{0\}$. Bounded within-epoch motion extends epoch divergence to every sample iterate. The exact heterogeneity is $\sum_i (\lambda d_i)^2/\lambda^2 = nC_n$.

For $q = b = 0$, put $\bar{\epsilon} = \epsilon/\lambda$ and $R_0 = \max\{1, \bar{\epsilon}n(n - r_n)/r_n\}$. Starting above $R_0 + \eta_0$ keeps every point in an epoch above R_0, because an update has magnitude at most η_k/n. Each negative-phase direction then has magnitude at least $1 - r_n/n$, whereas each positive one is at most one. The sum of phase directions is at most $np_n - n(1 - p_n)(1 - r_n/n) = -r_n(1 + p_n) \le -r_n$. Thus $x_{k+1} - x_k \ge \eta_k r_n/n$, closing the induction and proving divergence by nonsummability. Bias correction is one in this case. □

A.4 A COUNTEREXAMPLE OUTSIDE $q \le b$

Fix $q = .9$, $b = 0$, and a zero denominator constant. Use a period of 100 inputs: 40 copies of $-.1$, one $\varepsilon > 0$, and 59 copies of $B_\varepsilon = (104 - \varepsilon)/59$. The mean is exactly one and, as $\varepsilon \downarrow 0$, the second-moment ratio tends to

$$C_0 = \frac{40(.1)^2 + 59(104/59)^2}{100} < 1.84 < 2.$$

For small ε, all inputs and the periodic first moment before the negative block are below two. After 40 negative updates, that moment is at most $-.1 + 2.1(.9)^{40} < 0$. At the subsequent ε phase it is therefore bounded above by a fixed negative number. Since $b = 0$, the denominator there is ε: that phase contributes $-\Theta(1/\varepsilon)$, whereas the other 99 contributions remain $O(1)$. The stationary period-average drift thus tends to $-\infty$ with $C < 2$.

Online regret counterexamples (Heilman & Mohanty, 2026) apply to every memory pair; Theorem 9 concerns stationary input laws.

B COMMON TRAJECTORY TOOLS AND CONVERGENCE GUARANTEES

Unless stated otherwise, trajectory results use sample step η_k/n in epoch k, zero initial moments, a fixed finite $\epsilon \ge 0$, and either bias-correction convention. The order is fixed and the epoch scales satisfy

$$\eta_k > 0, \qquad \eta_k \to 0, \qquad \sum_k \eta_k = \infty. \tag{27}$$

The proof strategy is the same for each geometry: bound individual updates, track frozen moments without assuming bounded iterates, and use a proper potential to establish boundedness and convergence. Throughout, $g_i = \nabla f_i$ and $g = \nabla f$.

B.1 BOUNDED UPDATES, ABSOLUTE TRACKING, AND TRAPPING

Lemma 2 (Uniform normalized update bound)**.** *For $0 \le q \le b < 1$, $b > 0$, $m_0 = 0$, and $v_0 \ge 0$, every scalar input sequence satisfies*

$$|m_t| \le R_{q,b}\sqrt{v_t}, \qquad R_{q,b} := \frac{1 - q}{\sqrt{(1 - b)(1 - q^2/b)}}.$$

The same bound holds for periodic moments and, with Euclidean norm in the numerator, for a vector first moment with one shared second moment. At $q = b = 0$ take $R_{0,0} = 1$.

Proof. Weighted Cauchy–Schwarz in the moving-average formulas gives

$$|m_t|^2 \le (1 - q)^2 \sum_{h \ge 0} (q^2/b)^h \sum_{h \ge 0} b^h g_{t-h}^2 \le R_{q,b}^2 v_t.$$

Here $q^2 < b$. The infinite-history formula proves the periodic case, and vector Cauchy–Schwarz proves the shared-denominator case. The memoryless claim is immediate. □

Lemma 3 (Zero initialization and bias correction)**.** *Under zero initialization, a zero second moment always has zero first moment. Assigning $0/0 = 0$ therefore defines every update at $\epsilon = 0$. For $b > 0$, a coordinate's second moment stays positive after its first nonzero input. Corrected and raw normalized directions differ by a uniformly vanishing perturbation as the sample index tends to infinity.*

Proof. For $b > 0$, $v_t = (1-b)\sum_{s=1}^{t} b^{t-s} g_s^2$ vanishes exactly when all preceding inputs vanish. For $b = 0$ admissibility gives $q = 0$ and $(m_t, v_t) = (g_t, g_t^2)$. For bias correction, set $a_t = \sqrt{1-b^t}$:

$$\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} = \frac{a_t}{1-q^t} \frac{m_t}{\sqrt{v_t} + a_t \epsilon}, \qquad \left| \frac{m_t}{\sqrt{v_t} + a_t \epsilon} - \frac{m_t}{\sqrt{v_t} + \epsilon} \right| \le R_{q,b} \frac{1 - a_t}{a_t}. \tag{28}$$

The bound also holds at a zero moment and in vector norm for a shared denominator. The prefactor tends to one; at $q = b = 0$ correction is the identity. In particular all raw and corrected directions have a fixed bound, denoted below by B_u in Euclidean norm. □

Lemma 4 (Absolute filter tracking under vanishing steps)**.** *Suppose component gradients are globally Lipschitz, the memory pair is admissible, and $\eta_k > 0$ tends to zero. Freeze the realized index string at an epoch start, extending its past by complete permutations. On every fixed finite window, errors in the first and root-second moments tend to zero, without assuming bounded iterates. This holds for coordinatewise and shared normalization; a fixed order gives the periodic frozen states.*

Proof. The update bound gives vanishing motion over fixed windows and positions $O(t+1)$. At fixed lag h, the discrepancy between a historical component gradient and that same component at the frozen point tends to zero; Lipschitzness bounds it by $K(h + Nn + 1)$ on an N-epoch window. Dominated convergence against geometric weights controls first moments. For root second moments use reverse Minkowski,

$$\|\sqrt{v} - \sqrt{v^{\mathrm{fr}}}\| \le \Big((1-b) \sum_{h \ge 0} b^h \|G_{t-h} - G_{t-h}^{\mathrm{fr}}\|^2 \Big)^{1/2},$$

with vector norms inside the sum for shared normalization. The squared majorant is geometrically summable. The omitted preinitialization tails are $O((t+1)q^t)$ and $O((t+1)b^{t/2})$, hence vanish. For a fixed order, extend that same order into the past. At $q = b = 0$ only within-epoch gradient error remains. At positive epsilon, direction errors are bounded by $(\|\Delta m\| + R_{q,b}\|\Delta\sqrt{v}\|)/\epsilon$; (28) handles correction. □

Lemma 5 (Proper-potential trapping)**.** *Let $\|x_{k+1} - x_k\| \le B_u \eta_k$, with (27). Let $V \ge 0$ be continuous and proper, and let $v_* \ge 0$ with $\{V \le v_*\} \ne \varnothing$. If for every $a > v_*$ there are $c_a > 0$ and K_a such that*

$$k \ge K_a,\ V(x_k) \ge a \quad \Longrightarrow \quad V(x_{k+1}) - V(x_k) \le -c_a \eta_k,$$

then the trajectory is bounded, $\limsup_k V(x_k) \le v_*$, and $\operatorname{dist}(x_k, \{V \le v_*\}) \to 0$.

Proof. Every sublevel $\{V < a\}$ is entered arbitrarily late: otherwise telescoping and $\sum_k \eta_k = \infty$ would force $V < 0$. After a late entry, a step starting below a lies in a fixed compact neighborhood of that sublevel, whereas one starting above a decreases V. This proves boundedness. Uniform continuity on the attained compact set and $\eta_k \to 0$ make every later upward crossing overshoot a by $o(1)$. Hence $\limsup V \le a$. Let $a \downarrow v_*$; compactness then gives the distance conclusion. Within-epoch displacements $O(\eta_k)$ transfer each conclusion to sample iterates. □

For later use, if f is μ-strongly convex and L-smooth, an eventual margin $g(x_k)^\top U_k \ge c_r \|g(x_k)\|$ whenever $\|g(x_k)\| \ge r > 0$ proves convergence. Indeed, smoothness gives $\Delta f \le -c_r \eta_k \|g\| + L B_u^2 \eta_k^2 / 2$, and $\|g\|^2 \ge 2\mu(f - f_*)$ makes this a strict descent above every positive objective level. Apply Lemma 5 to $f - f_*$.

B.2 Cubic pairing and noninterpolating residual neighborhoods

Proposition 4 (Stationary cubic pairings)**.** *For a stationary stream satisfying the active-coordinate assumptions of Theorem 9, let $\mu = \mathbb{E}G \neq 0$, $S = \mathbb{E}\|G\|^2$, $C = S/\|\mu\|^2$, and $U_\epsilon = \mathbb{E}[m \oslash (\sqrt{v}+\epsilon\mathbf{1})]$, $\epsilon \geq 0$. Then*

$$(\mu \odot |\mu|)^\top U_0 \geq (2/C - 1)\|\mu\|^2, \tag{29}$$

$$(\mu \odot |\mu| + \epsilon\mu)^\top U_\epsilon \geq \|\mu\|^2 - S/2. \tag{30}$$

The first constant is sharp. In one dimension, writing $J_\epsilon = \operatorname{sgn}(\mu)U_\epsilon$, also

$$(1 + \epsilon/|\mu|)J_\epsilon \geq 1 - C/2, \tag{31}$$

whose coefficient $1/2$ in a universal bound $c(2 - C)$ is optimal. For frozen finite sums these include memoryless zero phases.

Proof. Put $s_j = \mathbb{E}G_j^2$. The scalar bound and Cauchy–Schwarz give $2\sum_{s_j>0} \mu_j^4/s_j - \|\mu\|^2 \geq 2\|\mu\|^4/S - \|\mu\|^2$. For the regularized bound normalize a scalar mean to one, set $y = m/(\sqrt{v} + e)$, $e = \epsilon/|\mu|$ and $a = \mathbb{E}\sqrt{v} \geq 1$. Since $\mathbb{E}y^2 \leq 1$,

$$(1 + e)\mathbb{E}y = 1 - \mathbb{E}[y(\sqrt{v} - 1)] \geq 1 - \sqrt{C - 2a + 1} \geq 1 - \sqrt{C - 1} \geq 1 - C/2.$$

Multiply by μ_j^2 and sum; including zero-mean coordinates on the right only weakens the bound. At memoryless zero phases retain $u \in [-1, 1]$: $d = u(|d| + e)$ and $u^2 \leq 1$ remain valid. At $e = 0$, mean-one inputs have positive fraction at least $1/C$ by Cauchy–Schwarz, giving the completed mean-sign bound $2/C - 1$. Scalar sharpness proves sharpness above, and $(2/C - 1)/(2 - C) = 1/C \to 1/2$ as $C \uparrow 2$. □

For a separable average define

$$\mathcal{W} = \sum_j \int_{x_{*,j}}^{x_j} F_j'(s)|F_j'(s)|\, ds, \qquad \mathcal{W}_\epsilon = \mathcal{W} + \epsilon(f - f_*). \tag{32}$$

Then

$$\nabla\mathcal{W} = g \odot |g|, \qquad \nabla\mathcal{W}_\epsilon = g \odot |g| + \epsilon g. \tag{33}$$

Applying the proposition to a frozen list with mean g and $S \leq C_0\|g\|^2$ gives respectively the positive margins $(2/C_0 - 1)\|g\|^2$ and $(1 - C_0/2)\|g\|^2$ when $C_0 < 2$.

Assume now

$$S(x) \leq C\|g(x)\|^2 + B, \qquad 1 \leq C < 2, \quad B \geq 0. \tag{34}$$

Theorem 11 (Noninterpolating residual neighborhood)**.** *Suppose every component gradient is globally Lipschitz and the separable average is μ-strongly convex in fixed finite dimension p. Under (34), the cyclic protocol and (27), for every admissible memory pair and fixed finite $\epsilon \geq 0$, the entire trajectory is bounded. Set*

$$R = \sqrt{B/(2 - C)} = \sqrt{D_0/(2n - D_1)}, \tag{35}$$

$$M_R = \max_{\|g(x)\| \leq R} \mathcal{W}_\epsilon(x). \tag{36}$$

Both sample and epoch iterates satisfy

$$\limsup_t \mathcal{W}_\epsilon(x_t) \leq M_R, \qquad \limsup_t \|x_t - x_*\| \leq p^{1/6}R/\mu,$$
$$\operatorname{dist}(x_t, \{\mathcal{W}_\epsilon \leq M_R\}) \to 0. \tag{37}$$

In dimension one,

$$\limsup_t |f'(x_t)| \leq R, \qquad \limsup_t |x_t - x_*| \leq R/\mu. \tag{38}$$

For $B = 0$ every sample iterate converges to x_*, including at $q = b = 0$.

Proof. Translate x_* to zero and put $w = g \odot |g| + \epsilon g$. For a frozen period, its moments and direction obey $\mathbb{E}M = g$, $\mathbb{E}V_j = \mathbb{E}G_j^2$, $\mathbb{E}\sqrt{V_j} \geq |g_j|$, and $\mathbb{E}u_j^2 \leq 1$. At $b > 0$ every nonzero stream has positive periodic second moments; an identically zero stream has zero pairing weight. At $q = b = 0$ a zero-input, zero-ϵ phase can take any limit in $[-1, 1]$: $M_j = u_j(\sqrt{V_j} + \epsilon)$ and $u_j^2 \leq 1$ still hold. Consequently, by two Cauchy–Schwarz inequalities,

$$\begin{aligned} w^\top \mathbb{E}u &= \|g\|^2 - \sum_j g_j \mathbb{E}[u_j(\sqrt{V_j} - |g_j|)] \\ &\geq \|g\|^2 - \|g\|\sqrt{S - \|g\|^2} \geq (1 - C/2)\|g\|^2 - B/2. \end{aligned} \tag{39}$$

Transfer outside the residual radius. Fix $r > R$ and consider starts with $G_k = \|g(x_k)\| \geq r$. The normalized inputs $g_i(x_k)/G_k$ and $e_k = \epsilon/G_k$ range over a compact set, by $S/G_k^2 \leq C + B/r^2$. Lemma 4 makes their normalized moment errors vanish. In a subsequential limit with $b > 0$, a nonzero coordinate stream has positive frozen denominators, hence the actual directions converge to its periodic directions. A zero stream has limiting mean $h_j = 0$ and therefore zero pairing weight $h_j|h_j| + eh_j$. At $q = b = 0$, nonzero phases converge directly; when both input and effective denominator vanish, the completed zero-phase values above cover all possible limits. Every limiting normalized pairing is at least $a_r = 1 - (C + B/r^2)/2 > 0$. Compactness and (28) give the uniform margin on late epochs

$$w(x_k)^\top U_k \geq (a_r/2)G_k^2 \qquad (G_k \geq r, k \text{ large}). \tag{40}$$

Global Lipschitzness of g bounds the Taylor remainder by $K\eta_k^2(G_k + \epsilon + \eta_k)$, so

$$\mathcal{W}_\epsilon(x_{k+1}) - \mathcal{W}_\epsilon(x_k) \leq -(a_r/4)\eta_k G_k^2 \tag{41}$$

eventually on $G_k \geq r$. The potential is nonnegative and proper. For every $\rho > 0$, strong convexity makes $\inf_{\mathcal{W}_\epsilon \geq M_R + \rho} \|g\| > R$: a minimizing sequence with bounded gradient is compact, and a limiting gradient at most R would contradict the definition of M_R. Lemma 5 now gives the potential and distance claims, boundedness, and their sample-level versions.

Radii and positive regularization. Strong monotonicity of the marginal gradients gives

$$\mathcal{W}_\epsilon(x) \geq \frac{\mu^2}{3\sqrt{p}}\|x\|^3 + \frac{\epsilon\mu}{2}\|x\|^2. \tag{42}$$

For $x_j > 0$, $0 \leq F_j'(s) \leq g_j(x) - \mu(x_j - s)$ on $[0, x_j]$. Integrating the squared and linear upper bounds over the larger interval $[x_j - g_j(x)/\mu, x_j]$, and reflecting when $x_j < 0$, gives $M_R \leq R^3/(3\mu) + \epsilon R^2/(2\mu)$. Comparison with (42) proves the radius. In one dimension (40) gives an inward direction of magnitude at least $(a_r/2)r/(r + \epsilon)$ outside $|f'| < r$. Nonsummability forces entry; vanishing jumps bound every later overshoot by $o(1)$. Let $r \downarrow R$. □

B.3 Coordinatewise growth and shared normalization

Define $D_\square$ by

$$\sum_i g_{i,j}(x)^2 \leq D_\square g_j(x)^2 \quad \text{for every } x, j. \tag{43}$$

For diagonal quadratics its optimum is $\max_j \sum_i h_{ij}^2/(n^{-1}\sum_i h_{ij})^2$, with zero assigned to an all-zero row. Summing gives

$$\sum_i \|g_i(x)\|^2 \leq D_\square \|g(x)\|^2. \tag{44}$$

Theorem 12 (Cyclic convergence from scalar margins)**.** *Suppose component gradients are globally Lipschitz and f is strongly convex. Assume either coordinatewise growth $D_\square \leq C_0 n$, or shared normalization (103) with $S \leq C_0\|g\|^2$, where $C_0 < 2$. For every admissible memory pair, fixed finite $\epsilon \geq 0$, zero moments, either bias convention and (27), the cyclic protocol converges along every sample iterate. For $n \geq 2$, $b > 0$, one may use $C_0 < C_{\text{fin}}^{(q)}(n, b)$ instead. At $q = b = 0$, the order may change arbitrarily each epoch. The thresholds are dimension independent; the scalar family of Theorem 10 makes 2 asymptotically sharp.*

Proof. Outside $\|x - x_*\| < \delta$, normalize by $G = \|g\| \geq \mu\delta$: $h = g/G$, $D_i = g_i/G$, $e = \epsilon/G$ are bounded. Coordinatewise, the active ratios obey

$$C_j = \frac{n^{-1}\sum_i D_{i,j}^2}{h_j^2} \leq C_0, \tag{45}$$

and the scalar margin gives

$$\operatorname{sgn}(h_j)U_j^{\mathrm{fr}} \geq \frac{A_0|h_j|}{\sqrt{C_0}|h_j| + e}, \qquad A_0 = 1 - \sqrt{C_0 - 1}. \tag{46}$$

Some $|h_j| \geq p^{-1/2}$ ensures a uniform positive projected margin. For shared normalization use (104) directly. The finite-period bounds in both cases replace A_0 by $1 - \sqrt{H_q(C_0)V_{n,b}(C_0)} > 0$.

Absolute tracking gives normalized errors $o(1)$. For $b > 0$, large coordinates have the phasewise floor

$$V_{i,j}^{\mathrm{fr}} \geq (1-b)b^{n-1}\tau^2 \quad (|h_j| \geq \tau); \tag{47}$$

small coordinates, with direction bound $\bar{R}$, contribute at most

$$\left|\sum_{|h_j|<\tau} h_j(U_{k,j} - U_j^{\mathrm{fr}})\right| \leq 2\bar{R}p\tau. \tag{48}$$

Choose τ then k. The shared denominator has floor $(1-b)b^{n-1}$. At zero memory, pass to limits of normalized inputs and bounded directions: a zero-input, zero-e phase retains any limit in $[-1, 1]$ or the unit ball. The identities $D_i = u_i(|D_i| + e)$ coordinatewise, or $D_i = u_i(\|D_i\| + e)$ when shared, and the unit direction bounds preserve the centered inequality. Zero coordinates have zero pairing weight. Compactness therefore transfers the same strict margin in every order. Bias correction follows from (28). Finally $\Delta f \leq -\gamma_\delta \eta_k G_k + LB_u^2\eta_k^2/2$; proper-potential trapping gives sample-level convergence. □

B.4 Non-asymptotic bounds and arbitrary epoch orders

Theorem 13 (Non-asymptotic guarantee at positive regularization)**.** *Assume the hypotheses of Theorem 12 with $C_0 < 2$, fix $\epsilon > 0$, and let f be μ-strongly convex with L-Lipschitz gradient. Run the cyclic protocol with $\eta_k = a(k+1)^{-\alpha}$, $a > 0$, $0 < \alpha \leq 1$. Put $F = f - f_*$, $\Psi = \sqrt{1+F}$, $A_0 = 1 - \sqrt{C_0 - 1}$ and*

$$d(F) = \frac{A_0\mu F}{\sqrt{1+F}\left(\sqrt{2C_0\mu F} + \epsilon\right)}.$$

There are instance-dependent constants Γ_0, Γ_1, depending only on the components, x_0, ϵ, the memory pair, a and α, such that for every $K \geq 1$

$$\sum_{k<K} \eta_k\, d\big(F(x_{kn})\big) \leq \Psi(x_0) - 1 + \Gamma_0 + \Gamma_1 \sum_{k<K} \eta_k^2. \tag{49}$$

Moreover d is increasing and $d(F) \geq c_* \min(F, 1)$ with $c_* = A_0\mu/[\sqrt{2}(\sqrt{2C_0\mu} + \epsilon)]$. For $\alpha = 1/2$ this gives $\min_{k<K}\big(f(x_{kn}) - f_*\big) = O(\log K/\sqrt{K})$.

Proof. Let $X_k = x_{kn}$, let $U^{\mathrm{fr}}(x)$ be the frozen periodic direction with denominator constant ϵ, and let $\bar{U}_k = (X_k - X_{k+1})/\eta_k$, so that $\|\bar{U}_k\| \leq B$ by Lemma 2 and (28).

Tracking. At sample $t = kn + i$, compare the actual moments with the periodic moments frozen at X_k. With a common Lipschitz constant L_c and cumulative sample step A_s, each input differs by at most $L_cB|A_{s-1} - A_{kn}|$. For lags $l \leq kn/2$ this is at most $4L_cB\eta_k(1 + l/n)$, since $\eta_{\lfloor k/2 \rfloor} \leq 2\eta_k$; older lags and the omitted prehistory are multiplied by $q^{kn/2}$ or $b^{kn/4}$ and bounded by $G_0 + L_cBA_{kn}$, where $G_0 = \max_i \|g_i(x_0)\|$. Weighted summation for the first moment and reverse Minkowski for the root second moment bound the errors by $4L_cB\eta_kQ_1 + \rho_k$ and $4L_cB\eta_k\sqrt{B_2} + \rho_k$, where $Q_1 = 1 + q/[n(1-q)]$, $B_2 = 1 + 2b/[n(1-b)] + b(1+b)/[n^2(1-b)^2]$ and $\rho_k = O((k+1)b^{kn/4})$.

Since $|m^{\mathrm{fr}}| \le R_{q,b}\sqrt{v^{\mathrm{fr}}}$, each coordinate of the direction error is at most $(|\Delta m| + R_{q,b}|\Delta\sqrt{v}|)/\epsilon$, and bias correction adds $O(q^{kn} + b^{kn})$ by (28). Hence

$$\|\bar{U}_k - U^{\mathrm{fr}}(X_k)\| \le e_k := \kappa_1\eta_k + \kappa_0(k),$$

with $\kappa_1 = 4\sqrt{p}L_cB(Q_1 + R_{q,b}\sqrt{B_2})/\epsilon$ (without $\sqrt{p}$ for shared normalization) and $\sum_k \eta_k\kappa_0(k) < \infty$.

Descent. $\|\nabla\Psi\| \le M_1 = \sqrt{L/2}$ and $\nabla\Psi$ is L-Lipschitz, because $\nabla^2\Psi = \nabla^2 f/(2\sqrt{1+F}) - gg^\top/(4(1+F)^{3/2})$ and $\|g\|^2 \le 2LF$. Hence $\Psi(X_{k+1}) \le \Psi(X_k) - \eta_k\nabla\Psi(X_k)^\top\bar{U}_k + LB^2\eta_k^2/2$. By (46), or (104) for shared normalization, $g^\top U^{\mathrm{fr}} \ge A_0\|g\|^2/(\sqrt{C_0}\|g\| + \epsilon)$; with $\|g\|^2 \ge 2\mu F$ this gives $\nabla\Psi^\top U^{\mathrm{fr}} \ge d(F)$. Therefore

$$\eta_k d(F(X_k)) \le \Psi(X_k) - \Psi(X_{k+1}) + M_1\eta_k e_k + LB^2\eta_k^2/2.$$

Summing over $k < K$ and using $\Psi \ge 1$ proves (49) with $\Gamma_0 = M_1\sum_k \eta_k\kappa_0(k)$ and $\Gamma_1 = M_1\kappa_1 + LB^2/2$. Finally $\partial_F \log d \ge 1/(2F) - 1/(2(1+F)) > 0$, $d(F) \ge c_*F$ for $F \le 1$, and $d(F) \ge c_*$ for $F \ge 1$. For $\alpha = 1/2$, $\sum_{k<K}\eta_k^2 = O(\log K)$ and $\sum_{k<K}\eta_k \ge a\sqrt{K}$. □

The same argument with $\Psi = (1+\mathcal{W}_\epsilon)^{1/3}$ and (30) gives the analogous bound for separable averages.

Bounded conditioning. Under the hypotheses of Theorem 21, with $\epsilon > 0$ and the same schedule, there are instance-dependent Γ_0, Γ_1 such that

$$\sum_{k<K}\eta_k\|x_{kn} - x_*\|^2 \le \Gamma_0 + \Gamma_1\sum_{k<K}\eta_k^2 \qquad \text{for every } K \ge 1. \tag{50}$$

Translate x_* to zero and use V, λ and ν from (110). The tracking step above does not use the geometry, so $\|\bar{U}_k - U_\epsilon(X_k)\|_\infty \le e_k$. Since $\|\nabla V(x)\|_1 \le \lambda\|x\|^2 + \epsilon\sqrt{p}\|x\|$ and $\nabla^2 V = 2\lambda\,\mathrm{diag}\,|x_j| + \epsilon I$,

$$V(X_{k+1}) \le V(X_k) - \eta_k(\nu - \lambda e_k)\|X_k\|^2 + \eta_k a_k\|X_k\| + B^2\eta_k^2(\lambda B\eta_k + \epsilon/2),$$

with $a_k = \epsilon\sqrt{p}e_k + \lambda B^2\eta_k$. Let k_0 be the first epoch after which $\lambda e_k \le \nu/4$ and $a_k \le \nu/4$. For $k \ge k_0$, $a_k\|x\| \le (\nu/4)\|x\|^2 + a_k/4$ gives

$$V(X_{k+1}) \le V(X_k) - \tfrac{\nu}{2}\eta_k\|X_k\|^2 + \tfrac{1}{4}\eta_k a_k + B^2\eta_k^2(\lambda B\eta_k + \epsilon/2).$$

Sum from k_0 and use $\sum_k \eta_k\kappa_0(k) < \infty$. The epochs before k_0 and $V(X_{k_0})$ are bounded through $\|X_k\| \le \|x_0\| + B\sum_{j<k}\eta_j$, and $\|x_0 - x_*\| \le \|g(x_0)\|/m$. Since $f - f_* \le L\|x - x_*\|^2/2$, $\eta_k \propto (k+1)^{-1/2}$ again gives $\min_{k<K}(f(x_{kn}) - f_*) = O(\log K/\sqrt{K})$.

Without regularization the tracked moments are divided by $\sqrt{v}$ rather than by ϵ. A window floor and a dichotomy on the size of the gradient still give non-asymptotic bounds.

Theorem 14 (Non-asymptotic guarantee without regularization)**.** *Let $\epsilon = 0$ and $1 \le C_0 < 2$. Assume globally Lipschitz component gradients and a μ-strongly convex average f with L-Lipschitz gradient. Run the cyclic protocol with $\eta_k = a(k+1)^{-\alpha}$, $a > 0$, $0 < \alpha \le 1$, for any admissible pair. Write $X_k = x_{kn}$ and $G_k = \|g(X_k)\|$. There are instance-dependent constants Γ_0, Γ_1, depending only on the components, x_0, the memory pair, a, α and C_0, such that for every $K \ge 1$:*

(i) under the hypotheses of Theorem 12 (coordinatewise growth or a shared second moment),

$$\sum_{k<K}\eta_k G_k \le \Gamma_0 + \Gamma_1\sum_{k<K}\eta_k^2; \tag{51}$$

(ii) for a separable μ-strongly convex average with $D_0 = 0$ and $D_1 \le C_0 n$,

$$\sum_{k<K}\eta_k G_k^2 \le \Gamma_0 + \Gamma_1\sum_{k<K}\eta_k^3. \tag{52}$$

For $\alpha = 1/2$, since $f - f_* \le G^2/(2\mu)$, these give $\min_{k<K}(f(X_k) - f_*) = O(\log^2 K/K)$ in (i) and $O(1/\sqrt{K})$ in (ii).

Proof. Let U^{fr} be the frozen periodic direction at X_k and $\bar{U}_k = (X_k - X_{k+1})/\eta_k$, so $\|\bar{U}_k\| \le B$. The tracking paragraph in the proof of Theorem 13 does not use ϵ: coordinatewise, actual and frozen first and root-second moments differ by at most $E_k = \kappa_2\eta_k + \rho_k$, with $\sum_k \eta_k\rho_k < \infty$. At $\epsilon = 0$, (28) multiplies each normalized coordinate by a factor $1 + O(q^{kn} + b^{kn})$; since $G_k \le G_0 + LB\sum_{j<k}\eta_j$, its effect on the pairings below is an additive term that we include in ρ_k. Let $R = R_{q,b}$, let $\bar{R}$ bound every coordinate of every direction, and put $\varphi = \sqrt{(1-b)b^{n-1}}$. Since $b = 0$ forces $q = 0$, the cases $b > 0$ and $q = b = 0$ cover every admissible pair.

(i), $b > 0$. Fix $\tau \in (0,1)$, chosen below, and let J be the set of coordinates with $|g_j(X_k)| \ge \tau G_k$. Each $j \in J$ has a component with $|g_{ij}(X_k)| \ge |g_j(X_k)|$, so every frozen phase has $\sqrt{v_{ij}^{\mathrm{fr}}} \ge \varphi\tau G_k$. Suppose first $E_k \le \varphi\tau G_k/2$. On J the actual root second moments are then at least $\varphi\tau G_k/2$, and

$$\Big|\frac{m}{\sqrt{v}} - \frac{m^{\mathrm{fr}}}{\sqrt{v^{\mathrm{fr}}}}\Big| \le \frac{|m - m^{\mathrm{fr}}| + R|\sqrt{v} - \sqrt{v^{\mathrm{fr}}}|}{\sqrt{v}} \le \frac{2(1+R)E_k}{\varphi\tau G_k}.$$

Coordinates outside J contribute at most $2\bar{R}p\tau G_k$ to $g^\top(\bar{U}_k - U^{\mathrm{fr}})$. By (46) at $e = 0$, $g^\top U^{\mathrm{fr}} \ge A_0 G_k/\sqrt{C_0}$ with $A_0 = 1 - \sqrt{C_0 - 1}$; hence

$$g^\top\bar{U}_k \ge \Big(\frac{A_0}{\sqrt{C_0}} - 2\bar{R}p\tau\Big)G_k - \frac{2(1+R)\sqrt{p}\,E_k}{\varphi\tau}.$$

Take $\tau = A_0/(4\bar{R}p\sqrt{C_0})$ and $c_0 = A_0/(2\sqrt{C_0})$. Then $g^\top\bar{U}_k \ge c_0 G_k - K_1 E_k$. In the opposite case $G_k < 2E_k/(\varphi\tau)$, and $g^\top\bar{U}_k \ge -BG_k \ge c_0 G_k - K_1' E_k$. Shared normalization needs no truncation, since its frozen phases obey $\sqrt{s_i^{\mathrm{fr}}} \ge \varphi G_k$, and (104) at $e = 0$ gives the same $A_0/\sqrt{C_0}$.

(i), $q = b = 0$. The epoch direction is exactly the mean of $\mathrm{sgn}(G_{ij})$ over the actual inputs $G_{ij} = g_{ij}(x_{kn+i-1})$, which differ from $g_{ij}(X_k)$ by at most $\delta_k = L_c\sqrt{p}\,\eta_k$. For any list with mean $\hat{\mu} \ne 0$ and second moment $\hat{s}$, the positive fraction is at least $\hat{\mu}^2/\hat{s}$ by Cauchy–Schwarz, so the mean sign is at least $2\hat{\mu}^2/\hat{s} - 1$ in the direction of $\hat{\mu}$. On J, if $\delta_k \le \tau G_k/4$, the actual list has $\hat{\mu}^2/\hat{s} \ge (|g_j| - \delta_k)^2/(\sqrt{C_0}|g_j| + \delta_k)^2 \ge 1/C_0 - K\delta_k/(\tau G_k)$. Hence $g^\top\bar{U}_k \ge (2/C_0 - 1)(1 - p\tau)G_k - 2p\tau G_k - K\sqrt{p}\,\delta_k/\tau$, and a small fixed τ and the same dichotomy give $g^\top\bar{U}_k \ge c_0 G_k - K_1\eta_k$ with $c_0 = (2/C_0 - 1)/2$. Shared normalization is identical with unit vectors in place of signs and the centered bound (104) for the actual list.

In all cases smoothness gives $F(X_{k+1}) \le F(X_k) - \eta_k g^\top\bar{U}_k + LB^2\eta_k^2/2 \le F(X_k) - c_0\eta_k G_k + K_1\eta_k E_k + LB^2\eta_k^2/2$, where $F = f - f_*$. Summing and using $F \ge 0$ proves (51) with $\Gamma_0 = [F(x_0) + K_1\sum_k \eta_k\rho_k]/c_0$ and $\Gamma_1 = (K_1\kappa_2 + LB^2/2)/c_0$.

(ii). Use $\mathcal{W}$ from (32), with $\nabla\mathcal{W} = g \odot |g|$ and $\|\nabla^2\mathcal{W}\| \le 2L\|g\|_\infty$ almost everywhere. By (29) the frozen margin is $(2/C_0 - 1)G_k^2$. Weighting the same direction errors by $g_j|g_j|$, coordinates in J lose at most $K_1 G_k E_k/\tau$, and the others at most $2\bar{R}p\tau^2 G_k^2$; at $q = b = 0$ the actual-list bound above applies coordinatewise and is summed by Cauchy–Schwarz as in Proposition 4. Put $c_0 = (2/C_0 - 1)/2$, choose τ with $2\bar{R}p\tau^2 \le c_0$, and use $K_1 G_k E_k/\tau \le c_0 G_k^2/2 + K_1^2 E_k^2/(2c_0\tau^2)$; after the same dichotomy, $(g \odot |g|)^\top\bar{U}_k \ge (c_0/2)G_k^2 - K_2 E_k^2$. Taylor's bound along the step, $\|g\|_\infty \le G_k + LB\eta_k$ and $LB^2\eta_k^2 G_k \le (c_0/4)\eta_k G_k^2 + (L^2B^4/c_0)\eta_k^3$ give

$$\mathcal{W}(X_{k+1}) \le \mathcal{W}(X_k) - \tfrac{c_0}{4}\eta_k G_k^2 + K_2\eta_k E_k^2 + K_3\eta_k^3,$$

and summation proves (52), since $\eta_k E_k^2 \le 2\kappa_2^2\eta_k^3 + 2\eta_k\rho_k^2$.

For $\alpha = 1/2$, $\sum_{k<K}\eta_k \ge a\sqrt{K}$, $\sum_{k<K}\eta_k^2 = O(\log K)$ and $\sum_k \eta_k^3 < \infty$. Hence $\min_{k<K} G_k = O(\log K/\sqrt{K})$ in (i) and $\min_{k<K} G_k^2 = O(1/\sqrt{K})$ in (ii). □

Arbitrary orders within epochs.

Theorem 15 (Convergence under arbitrary epoch orders)**.** *Fix an admissible memory pair, a finite $\epsilon \ge 0$, zero moments, either bias convention, and (27); for $b > 0$, assume*

$$\eta_{k+1}/\eta_k \longrightarrow 1. \tag{53}$$

Allow any history-dependent epoch permutations. The convergence and residual conclusions of Theorems 2, 11, 7 and 12 remain valid with $D_1/n < 2$ or $D_\square/n < 2$. For $\epsilon > 0$, so do those of

Theorem 21 and Corollary 6, with (107). At $\epsilon = 0$, bounded conditioning gives convergence under (112), in particular for $C < 8\kappa/(\kappa+1)^2$ when $q = b$. All conclusions hold pathwise under random reshuffling.

Proof. For $b > 0$, freeze $d_1, \dots, d_n$ with mean μ, second moments s_j, and $S = \sum_j s_j$ along bi-infinite epoch permutations. Brackets average $T = Nn$ consecutive samples. Normalizing active coordinates by $\sqrt{s_j}$ bounds inputs by $\sqrt{n}$. Every backward $2n$ samples contain a complete permutation, bounding the oscillation of $\log v_j$ by $\Lambda = \log[1/((1-b)b^{2n-1})]$. Filter summation and Jensen give

$$\|\langle m\rangle - \mu\| \le K\sqrt{S}/T, \quad \sum_j \langle v_j\rangle \le (1 + K/T)S, \quad \langle\sqrt{v_j}\rangle \ge |\mu_j| - K\sqrt{s_j}/T.$$

For $a = P_q(d_j^2) \ge m_j^2$, (14) and logarithmic telescoping give $\langle v_{t-h,j}/v_{t,j}\rangle \ge e^{-h\Lambda/T}$, hence $\langle m_j^2/v_j\rangle \le 1 + K/T$. Inactive streams have zero moments; shared normalization uses squared norms. At $q = b = 0$, one epoch gives these bounds exactly: take $N = 1$.

With $u = m \oslash (\sqrt{v} + \epsilon\mathbf{1})$, the identity $m = u \odot (\sqrt{v} + \epsilon\mathbf{1})$ and Cauchy–Schwarz yield

$$\begin{aligned}(\lambda h \odot |h| + \epsilon h)^\top \langle u\rangle \ge h^\top\mu - \|h\|\sqrt{S - 2\lambda h^\top\mu + \lambda^2\|h\|^2} \\ - KT^{-1/2}\|h\|(\sqrt{S} + \lambda\|h\|).\end{aligned}$$

For $h = \mu, \lambda = 1$, the separable margin is $[1 - \sqrt{C-1} - O(T^{-1/2})]\|\mu\|^2$. For $h = x - x_*$, $\lambda = C(m+L)/4$, the conditioning margin (110) holds up to $O(T^{-1/2}\|x - x_*\|^2)$. Centered Cauchy–Schwarz, as in (104), gives, for shared normalization or an individual scalar stream,

$$\frac{\mu^\top\langle u\rangle}{\|\mu\|} \ge \frac{[1 - \sqrt{C-1} - O(T^{-1/2})]\|\mu\|}{\sqrt{(1 + O(T^{-1}))C}\,\|\mu\| + \epsilon}.$$

Apply this coordinatewise under coordinatewise growth. On residual exteriors substitute $C + B/r^2$, or use the conditioning residual calculation. A fixed large N preserves positive cyclic margins.

Freeze the realized indices. Lemma 4 applies: lag errors have linear majorants, geometric tails are summable, and the omitted past vanishes before boundedness is known. Positive ϵ transfers tracking to directions; at $\epsilon = 0$ see the final paragraph. Local flatness gives $x_{k+N} - x_k = -s_k[\langle u^{\mathrm{fr}}(x_k)\rangle + o(1)]$, where $s_k = \sum_{j=0}^{N-1}\eta_{k+j}$. The cyclic potential Taylor bounds absorb the errors uniformly on each exterior. On nonoverlapping windows, $s_k \to 0$ and $\sum s_k = \infty$. Lemma 5 gives convergence and coarse residual bounds, extended to all samples by $O(s_k)$ within-window motion.

Zero regularization. For $\epsilon = 0$ and $b > 0$, a window floor replaces ϵ in the transfer. Normalize by $G = \|g(x_k)\|$ as in Theorem 12. Every backward window of $2n$ samples contains a complete permutation, and in coordinate j some normalized input has magnitude at least $|h_j|$. Hence on coordinates with $|h_j| \ge \tau$ the frozen second moments are at least $(1-b)b^{2n-1}\tau^2$, and the actual ones at least a quarter of this eventually, so tracked moments give tracked directions there. The remaining coordinates contribute at most $2Rp\tau$ to the linear pairing and $2Rp\tau^2$ to the cubic pairing. Choosing τ before N and the late epoch preserves the margins. Shared normalization needs no truncation, since its window floor is $(1-b)b^{2n-1}\|h\|^2$. At $q = b = 0$ the completed zero-phase graph of Theorem 12 applies window by window.

Bounded conditioning at zero regularization. The truncation above uses weights $|h_j|$, whereas the conditioned potential weights coordinates by $x_j|x_j|$. We therefore argue by compactness, as in the cyclic zero-regularization paragraph after Theorem 21. Translate x_* to zero, let $V = \sum_j |x_j|^3/3$, and consider windows of N epochs starting at late x_k with $r_k = \|x_k\| \ge r_* > 0$. The normalized component gradients $g_i(x_k)/r_k$, the averaged Hessians $A(x_k)$ and the realized index strings (with the product topology on sequences of permutations) range over compact sets, so every subsequence has a further convergent one. If a coordinate's limiting normalized stream is not identically zero, every backward window of $2n$ samples contains its largest entry, because it contains a complete permutation; the frozen and actual second moments there are then at least a positive multiple of r_k^2, and absolute tracking makes the normalized directions converge to those of the limiting stream.

On identically zero limiting streams retain the actual limits, bounded by $R_{q,b}$ by Lemma 2. The bracket bounds above hold for every limiting stream, so the calculation of that paragraph, with $\langle u \rangle$ in place of U and errors $O(T^{-1/2})$, gives $(y \odot |y|)^\top \langle U \rangle \geq \delta - O(T^{-1/2})$ for every limit, where $\delta = 8\kappa/[C(\kappa+1)^2] - 1 - (R_{q,b}-1)(\kappa-1)^2/(\kappa+1)^2 > 0$. Fix N with the error below $\delta/2$. If late windows had $\nabla V(x_k)^\top (x_k - x_{k+N}) < (\delta/4) s_k \|x_k\|^2$ with $\|x_k\| \geq r_*$ infinitely often, a convergent subsequence would contradict this bound. Hence the window margin holds uniformly outside every ball, the Taylor remainder is $O(s_k^2(\|x_k\| + s_k))$, and Lemma 5 on nonoverlapping windows proves convergence. At $q = b = 0$ the completed zero-phase graph again applies. □

B.5 IID SAMPLING

Theorem 16 (IID convergence in the safe geometries)**.** *Assume globally Lipschitz component gradients, a strongly convex average, admissible memories, zero moments and either bias convention, with a finite $\epsilon \geq 0$ in cases (i)–(iii), and in case (iv) either $\epsilon > 0$ or $\epsilon = 0$ with (112). Under IID uniform sampling and sample steps $\gamma_t = a(t+1)^{-\alpha}$, $a > 0$, $0 < \alpha \leq 1$, the iterates from each fixed initial point converge almost surely to the minimizer in these cases:*

(i) coordinatewise Adam with $D_\square/n < 2$;

(ii) coordinatewise Adam with a separable average and $S \leq C\|g\|^2$, $C < 2$;

(iii) shared normalization (103) with that aggregate bound;

(iv) coordinatewise Adam under the Hessian and growth assumptions of Theorem 21.

No bounded-iterate assumption or smallness restriction on a is imposed.

Proof. Translate the minimizer to zero. Constants below are independent of labels. Lemma 2 and (28) give $\|x_{t+1} - x_t\| \leq B\gamma_t$, hence $\|x_t\| \leq K(1+t)$ before stability.

Global history control. Let $Z_t = (I_t, I_{t-1}, \dots)$, with fixed preinitialization labels. Write $u(x,z)$ for the raw infinite-filter direction and $U(x) = \mathbb{E}_\pi u(x, Z)$, where π is the IID history law. Reverse Minkowski for the second-moment roots and the moment cone give

$$\|u(x,z) - u(y,z)\| \leq \frac{\|\Delta m\| + R_{q,b}\|\Delta\sqrt{v}\|}{\epsilon} \leq K\|x - y\|.$$

Since $\|g_i(x)\| \leq K(1 + \|x\|)$, directions are bounded and histories sharing j labels differ by at most $K \min\{1, (1+\|x\|)\rho^j\}$, for some $\rho < 1$. Actual lag-ℓ inputs use $x_{t-\ell}$, and

$$\gamma_{t-\ell}/\gamma_t \leq (\ell+1)^\alpha, \qquad \|x_t - x_{t-\ell}\| \leq B\gamma_t \ell(\ell+1)^\alpha.$$

Summing these errors against the geometric first/root-second filters, including omitted preinitialization tails and geometric bias corrections, yields

$$\|u_{t+1}^{\mathrm{act}} - u(x_t, Z_{t+1})\| \leq K\gamma_t + K(1+t)\theta^t, \qquad \theta < 1. \tag{54}$$

The memoryless case is immediate.

Poisson windows without preboundedness. Take a proper potential Ψ with $\Psi(0) = 1 < \Psi(x)$ for $x \neq 0$, bounded Lipschitz gradient, and put $d = \nabla\Psi^\top U$, $\phi_x = \nabla\Psi(x)^\top(u(x,\cdot) - U(x))$. For the kernel P prepending an IID label, coupling gives $\|P^j\phi_x\|_\infty \leq K \min\{1, (1+\|x\|)\rho^j\}$; its x-modulus is $K\|x-y\|$. Summing the smaller bounds defines $G_x = \sum_{j\geq 0} P^j\phi_x$, $Q_x = PG_x$, with $G_x - Q_x = \phi_x$ and, for $\delta = \|x - y\| > 0$,

$$\|G_x\|_\infty \leq K\log(2 + \|x\|), \qquad \|G_x - G_y\|_\infty \leq K\delta\left[1 + \log\frac{2 + \|x\| + \|y\|}{\delta}\right]. \tag{55}$$

The same bounds hold for Q. Given $\mathcal{F}_t = \sigma(I_1, \dots, I_t)$, $\xi_{t+1} = G_{x_t}(Z_{t+1}) - Q_{x_t}(Z_t)$ is centered and $|\xi_{t+1}| \leq K\log(t+2)$; exactly $\phi_{x_t}(Z_{t+1}) = \xi_{t+1} + Q_{x_t}(Z_t) - Q_{x_t}(Z_{t+1})$. For fixed $H > 0$ let $\tau_H(T) = \min\{N > T : \sum_{t=T}^{N-1}\gamma_t \geq H\}$. Since $\gamma_t \geq a/(t+1)$, $\tau_H(T) \leq K_H(T+1)$ and $\sum_{t=T}^{\tau_H(T)-1}\gamma_t^2\log^2(t+2) \leq K_H\gamma_T\log^2(T+2)$. Azuma and a union bound imply

$$\mathbb{P}\left(\max_{T < N \leq \tau_H(T)}\left|\sum_{t=T}^{N-1}\gamma_t\xi_{t+1}\right| > r\right) \leq 2K_H(T+1)\exp\left[-\frac{r^2}{K_H\gamma_T\log^2(T+2)}\right].$$

This is summable in T for every $\alpha > 0$. Borel–Cantelli for rational $H, r > 0$ makes all these maxima vanish almost surely. Writing $Q_t = Q_{x_t}$, the remaining sum telescopes exactly to

$$\gamma_T Q_T(Z_T) - \gamma_{N-1} Q_{N-1}(Z_N) + \sum_{t=T+1}^{N-1} \big[(\gamma_t - \gamma_{t-1}) Q_{t-1}(Z_t) + \gamma_t (Q_t - Q_{t-1})(Z_t)\big].$$

Boundaries and step changes are $O_H(\gamma_T \log(T+2))$. By (55) and monotonicity of $\delta[1 + \log(A/\delta)]$ for $\delta \le A$, state changes are bounded by $K \sum \gamma_t \gamma_{t-1} \log(t+2)$, of the same order: use $\gamma_{t-1} \le \gamma_T$ and $\sum \gamma_t \le H + \gamma_T$. Taylor and tracking errors are $O_H(\gamma_T) + o(1)$, including initialization. Thus

$$\sup_{T < N \le \tau_H(T)} \left| \Psi(x_N) - \Psi(x_T) + \sum_{t=T}^{N-1} \gamma_t d(x_t) \right| \longrightarrow 0 \quad \text{almost surely.} \tag{56}$$

Suppose $d \ge 0$ and $\inf_{\Psi \ge v} d > 0$ for $v > 1$. Also $|\Delta\Psi| \le D\gamma_t$. Fix $e > 0$ and rational $H > 0$ with $DH < e/8$. Partition late times into $T_{k+1} = \tau_H(T_k)$; within-block variation is eventually below $e/4$. A block starting above $1 + e/2$ stays above $1 + e/4$, so (56) decreases its endpoint by at least $H \inf_{\Psi \ge 1+e/4} d/2 > 0$. Otherwise its endpoint is below $1 + 3e/4$. Since $\Psi \ge 1$, endpoints eventually enter and remain below $1 + 3e/4$; intermediate values are below $1 + e$. Taking $e = 1/j$ proves $\Psi \to 1$; properness and its unique minimum give boundedness and $x_t \to 0$.

Stationary margins and potentials. Active IID finite-alphabet streams have geometric last-nonzero ages, hence positive second moments and integrable logarithms when $b > 0$. Zero streams and $q = b = 0$ are direct. In (i),(iii), (104), coordinatewise in (i), gives

$$g^\top U \ge \frac{(1 - \sqrt{C-1})\|g\|^2}{\sqrt{C}\|g\| + \epsilon}, \qquad C = D_\square / n \text{ in (i).}$$

Use $\Psi = \sqrt{1+F}$, $F = f - f_*$. Smoothness gives $\|\nabla\Psi\| \le \sqrt{L/2}$ and $\mathrm{Lip}(\nabla\Psi) \le L$. Strong convexity gives properness and

$$d \ge \frac{(1 - \sqrt{C-1})\mu F}{\sqrt{1+F}(\sqrt{2C\mu F} + \epsilon)},$$

uniformly positive for $F \ge \delta > 0$. In (ii), use $W = \mathcal{W}_\epsilon$ from (32); in (iv), $W = \lambda \sum_j |x_j|^3/3 + \epsilon\|x\|^2/2$, $\lambda = C(m+L)/4$. The stationary proofs of (30) and (110) give $\nabla W^\top U \ge c\|x\|^2$. For $r = \|x\|$,

$$c_1 r^3 \le W \le K(r^3 + r^2), \quad \|\nabla W\| \le K(r^2 + r), \quad \|D^2 W\| \le K(1+r) \text{ a.e.}$$

Consequently $\Psi = (1+W)^{1/3}$ has bounded Lipschitz gradient: its Hessian terms $D^2W/[3(1+W)^{2/3}]$ and $-2\nabla W \nabla W^\top/[9(1+W)^{5/3}]$ are locally bounded and $O(r^{-1})$ at infinity. Its drift is at least $cr^2/[3(1+W)^{2/3}]$, uniformly positive on every exterior level. These potentials supply the preceding hypotheses.

Zero regularization in cases (i)–(iii). At $\epsilon = 0$ the direction is not Lipschitz in the moments, so a coupling over finite blocks replaces the Poisson equation. Fix $r > 0$ and late times t with $G_t = \|g(x_t)\| \ge r$, and normalize the component list $d_i = g_i(x_t)/G_t$, whose mean h has norm one. Splitting historical lags at $t/2$, bounded velocity, global Lipschitzness and geometric filter tails show that replacing every historical input by the same component at x_t changes the first and root-second moments by $O(\gamma_t + tb^{t/4})$, uniformly in the labels; after division by $G_t \ge r$ this is $o(1)$.

Couple the next updates with a stationary frozen run for the list d that uses the same future labels. For an active coordinate j choose a label with $|d_{ij}| = D_j := \max_i |d_{ij}|$. If it occurs among the last $\lfloor l/2 \rfloor$ of l updates, both second moments are at least $(1-b)b^{l/2}D_j^2$, while the moment differences are at most Kq^l and $Kb^{l/2}$. Hence, for $D_j \ge \delta$,

$$\mathbb{E}|u_l - u_l^*| \le K_\delta b^{l/4} + 2R_{q,b}(1 - 1/n)^{\lfloor l/2 \rfloor}, \tag{57}$$

which is summable in l; the vector version holds for shared normalization with $D = \max_i \|d_i\| \geq 1$. On that event the actual input also has magnitude at least $\delta/2$, so the actual directions obey the same bound up to $o(1)$. Coordinates with $|h_j| \geq \delta$ have $D_j \geq \delta$, and the others contribute at most $2R_{q,b}p\delta$ to $h^\top u$ and $2R_{q,b}p\delta^2$ to $(h \odot |h|)^\top u$. The stationary pairings are at least $\delta_0 = 2/C - 1$, with $C = D_\square/n$ in (i) (Corollary 3, Theorem 20, Proposition 4). Choosing δ, then a block length T, then late t, gives

$$\mathbb{E}\Big[p(h)^\top \frac{1}{T}\sum_{j=1}^{T} u_{t+j} \;\Big|\; \mathcal{F}_t\Big] \geq \frac{\delta_0}{2},$$

with $p(h) = h$ in (i),(iii) and $p(h) = h \odot |h|$ in (ii), uniformly over late starts with $G_t \geq r$. At $q = b = 0$ the same holds with $T = 1$: a limiting output at a zero input lies in $[-1, 1]$ (the unit ball when shared), and $d = u|d|$ is preserved.

Use $\Psi = \sqrt{1+F}$ in (i),(iii) and $\Psi = (1+W)^{1/3}$ with the cubic potential at $\epsilon = 0$ in (ii). For $X_k = x_{kT}$ and $\bar\gamma_k = T\gamma_{kT}$, the Taylor remainder is $O(\bar\gamma_k^2)$, so for each $e > 0$ there are deterministic $K_e, c_e > 0$ with

$$\mathbb{E}[\Psi(X_{k+1}) - \Psi(X_k) \mid \mathcal{F}_{kT}] \leq -c_e\bar\gamma_k \quad \text{on } \{\Psi(X_k) \geq 1 + e,\ k \geq K_e\},$$

and $|\Psi(X_{k+1}) - \Psi(X_k)| \leq D\bar\gamma_k$. The centered increments are bounded by $2D\bar\gamma_k$. Azuma's inequality on windows of fixed cumulative step, with the union bound used above, makes their maximal partial sums vanish almost surely for every $\alpha > 0$. The level-trapping argument above then gives $\Psi(X_k) \to 1$, hence boundedness and $X_k \to x_*$; vanishing within-block motion extends this to all samples.

Zero regularization in case (iv). The truncation by $|h_j|$ does not suit the weights $x_j|x_j|$ of the conditioned potential, so normalize by $r_t = \|x_t\| \geq r$ instead: $y = x_t/r_t$, $d_i = g_i(x_t)/r_t$, $g = \mathbb{E}d_i = A(x_t)y$ and $D_j = \max_i |d_{ij}|$. The frozen-input replacement above changes the normalized moments by $o(1)$, and (57) holds on every coordinate with $D_j \geq \delta$. Let $Z = \{j : D_j < \delta\}$ and treat it as the zero-stream set of the cyclic argument after Theorem 21: there both the stationary and the actual directions lie in $[-R_{q,b}, R_{q,b}]$. Since $|g_j| < \delta$ on Z, the steps of that argument hold up to $O(\delta)$: $\sum_{P\setminus Z} y_j^2/C_j \geq (a - \sqrt{p}\,\delta)^2/S$ and $\|y_Z\|^2 \leq 1 - (a - \sqrt{p}\,\delta)^2/\|g\|^2$, with $a = y^\top g \geq m$ and $\|g\| \leq L$. On $P \setminus Z$ the stationary IID scalar bound $\operatorname{sgn}(g_j)U_j \geq 2/C_j - 1$ applies, since active streams have positive second moments and integrable logarithms. Hence the stationary pairing is at least $\delta_1 - O(\delta)$ with $\delta_1 = 8\kappa/[C(\kappa+1)^2] - 1 - (R_{q,b}-1)(\kappa-1)^2/(\kappa+1)^2 > 0$, and choosing δ, then T, then late t gives

$$\mathbb{E}\Big[(y \odot |y|)^\top \frac{1}{T}\sum_{j=1}^{T} u_{t+j} \;\Big|\; \mathcal{F}_t\Big] \geq \frac{\delta_1}{2}$$

uniformly over late starts with $\|x_t\| \geq r$. With $\Psi = (1 + \sum_j |x_j|^3/3)^{1/3}$, whose gradient $(x \odot |x|)/[3(1+V)^{2/3}]$ is bounded and Lipschitz, the conditional block drift is at most $-c_e\bar\gamma_k$ on $\{\Psi(X_k) \geq 1 + e\}$, because $\|x\|^2/(1+V)^{2/3}$ is bounded below there. The same Azuma windows and level trapping prove almost sure convergence. At $q = b = 0$ take $T = 1$ and the completed zero-phase graph. □

For IID minibatches of size s, $S_s = \|g\|^2 + (S - \|g\|^2)/s$, also coordinatewise; the same conclusions hold with $C_s = 1 + (C-1)/s$ replacing C. Under the additive growth $S \leq C\|g\|^2 + B$ of Corollary 6 and $\epsilon > 0$, the proof of case (iv) gives its residual conclusion almost surely: the conditioned drift is uniformly positive above every level strictly larger than the residual level, and the same windows and level trapping apply.

C From frozen rays to divergent trajectories

This appendix turns an attracting outward ray of the frozen field into divergence of the actual iterates: a joint invariant neighborhood controls the moment error and the angular deviation while the radius grows. It then treats clipping, decoupled weight decay and running-maximum normalization.

C.1 A COMMON ATTRACTING-RAY TRANSFER THEOREM

For symmetric matrices $H_0, \dots, H_{n-1}$ and a direction y, let $(M_i(y), V_i(y))$ be the unique periodic moments

$$M_{i+1} = qM_i + (1-q)H_i y, \qquad V_{i+1} = bV_i + (1-b)(H_i y)^{\odot 2}. \tag{58}$$

Define $U(y) = n^{-1}\sum_i M_{i+1}(y) \oslash \sqrt{V_{i+1}(y)}$. For a linear functional ℓ satisfying $\ell(e) = 1$, put $T = \ker \ell$ and $F(y) = -U(y) + y\ell(U(y))$ on the section $\ell(y) = 1$.

Theorem 17 (Frozen-ray tracking)**.** *Fix $0 \le q \le b < 1$, $b > 0$. Suppose every coordinate of every $H_i e$ is nonzero, $U(e) = -ae$ with $a > 0$, and $DF(e)|_T$ is Hurwitz. For positive nonincreasing epoch scales with $\sum_k \eta_k = \infty$, cyclic Adam with sample step η_k/n, zero initial moments, and $x_0 = r_0 e$ with $r_0 \ge K_*(\eta_0 + \epsilon)$ satisfies*

$$\ell(x_k) \to \infty, \qquad x_k/\ell(x_k) \to e, \qquad \ell(x_{k+1}) - \ell(x_k) = a\eta_k + o(\eta_k). \tag{59}$$

This holds for every fixed finite $\epsilon \ge 0$ with bias correction either on or off. The constant $K_* > 0$ depends only on the matrices, ray, section functional, memory pair and bias convention, not on the schedule or ϵ. The same holds from $x_0 = r_0 y_0$, uniformly for y_0 in a fixed neighborhood of e in the section.

Proof. Lemma 2 makes within-epoch displacement $O(\eta_k)$, also with bias correction. At epoch starts write

$$r_k = \ell(x_k), \quad y_k = x_k/r_k, \quad h_k = \eta_k/r_k, \quad w_k = (m_k/r_k, v_k/r_k^2).$$

Choose compact $S_\delta = \{\ell(y) = 1 : \|y - e\| \le \delta\}$ with frozen second moments bounded below. The affine period moment map has linear part $\mathrm{diag}(q^n I, b^n I)$ and C^1 fixed point $w^*(y)$. Set $\rho = \max(q^n, b^n) < 1$ and $E_k = \|w_k - w^*(y_k)\|_*$ in a product norm.

Tracking and the prospective epoch map. With the tube inside S_δ, the update bound keeps prospective points there and gives $|r_{k+1}/r_k - 1| + \|y_{k+1} - y_k\| = O(h_k)$ and normalized inputs $H_i y_k + O(h_k)$. Moment recursion, radial rescaling and variation of w^* therefore give

$$E_{k+1} \le (\rho + C_T h_k)E_k + C_T h_k. \tag{60}$$

For some $W > E_0 + \sup_{S_\delta} \|w^*(y)\|_* + 1$ and small $\bar h$, this recursion preserves $[0, W]$ when $h_k \le \bar h$.

For small $E_k + h_k$, the actual second moments retain the frozen lower bound, so Lipschitz continuity gives

$$x_{k+1} = x_k - \eta_k[U(y_k) + \xi_k], \qquad \|\xi_k\| \le C\Delta_k, \qquad \Delta_k = E_k + h_k + \epsilon/r_k + q^{kn} + b^{kn}. \tag{61}$$

Without bias correction omit the last two terms; with it, (28) contributes $O(q^t + b^t)$, while the denominator contributes $O(\epsilon/r_k)$. For $z_k = y_k - e \in T$,

$$\begin{aligned} r_{k+1} - r_k &= \eta_k[-\ell(U(e + z_k)) + \zeta_k], & |\zeta_k| &\le C\Delta_k, \\ z_{k+1} &= z_k + h_k[F(e + z_k) + \omega_k], & \|\omega_k\| &\le C(\Delta_k + h_k). \end{aligned} \tag{62}$$

Strict tube margins and burn-in. Shrink the section until $-\ell(U(y)) \ge 3a/4$. Since $A_T = DF(e)|_T$ is Hurwitz, choose $P \succ 0$ with $A_T^\top P + PA_T = -I$. For $\mathcal{V}(z) = z^\top Pz$ on a smaller section, $\nabla\mathcal{V}(z)^\top F(e + z) \le -c\|z\|^2$. Nested sublevels $\Omega_- \Subset \Omega_+ \Subset S_\delta - e$ then satisfy

$$\mathcal{V}(z_{k+1}) - \mathcal{V}(z_k) \le -ch_k\|z_k\|^2 + Ch_k\|z_k\|\|\omega_k\| + Ch_k^2. \tag{63}$$

Choose $\tau > 0$ and reduce $\bar h$ so that the radial increment is at least $a\eta_k/2$, a step from Ω_- stays in Ω_+, the Lyapunov increment is negative on $\Omega_+ \setminus \Omega_-$, and

$$(\rho + C_T\bar h)\tau + C_T\bar h \le \tau. \tag{64}$$

Compact-annulus margins justify these choices; require ϵ/r_k and the bias terms below the same tolerance.

The radius-independent $E_0 = \|w^*(e)\|_*$ admits finite B with $\rho^B(E_0 + 1) \le \tau/4$ and small bias terms. Since $\|x_k - r_0 e\| \le CB\eta_0$ for $k \le B$, large r_0 gives $r_k \ge r_0/2$, $h_k \le \bar h/2$, $z_B \in \Omega_-$ and small ϵ/r_k. From (60) and $E_k \le W$,

$$E_B \le \rho^B E_0 + \frac{2C_T(1 + W)\eta_0}{(1 - \rho)r_0} \le \tau/2$$

after a further increase of r_0.

For $x_0 = r_0 y_0$ with $y_0 - e$ in a smaller sublevel of $\mathcal{V}$, the same displacement bound places z_B in Ω_-, and E_0 stays bounded uniformly in y_0. After fixing the section, tolerances and B, every remaining requirement is $r_0 \geq K(\eta_0 + \epsilon)$: burn-in costs $CB\eta_0$ and later errors use only η_0/r_0 and ϵ/r_0. Taking the maximum of these finitely many constants gives schedule- and ϵ-independent K_*.

Joint first exit and the limit. After B, no first exit from $z_k \in \Omega_+$, $E_k \leq \tau$, $h_k \leq \bar{h}$, $r_k \geq r_0/2$ exists: (60)–(64) preserve $E \leq \tau$; the radial margin gives $r_{k+1} \geq r_k + a\eta_k/2$ and hence $h_{k+1} \leq h_k$; and the nested sublevels preserve $z_{k+1} \in \Omega_+$.

Thus nonsummability gives $r_k \to \infty$, whence $h_k \to 0$ and $E_k \to 0$ by (60). The update bound also gives $r_k \leq r_0 + C_r \sum_{j<k} \eta_j$, so

$$\sum_k h_k \geq \sum_k \frac{\eta_k}{r_0 + C_r \sum_{j<k} \eta_j} = \infty, \tag{65}$$

by comparison with the logarithm of the denominator. Because $\Delta_k \to 0$, on each angular annulus the error terms in (63) are eventually less than half its descent term. Equation (65) and vanishing steps imply $z_k \to 0$. Equation (62) gives the radial-speed limit, and bounded within-epoch motion gives sample-level divergence. □

C.2 Gradient clipping and decoupled weight decay

Practical implementations rescale gradients and decay weights. At $q = b = 0$ and $\epsilon = 0$ each sample update is a coordinatewise sign, so rescaling any sample gradient by a positive factor, in particular by global-norm clipping, leaves every iterate unchanged. The memoryless results at $\epsilon = 0$ therefore hold verbatim under clipping. For the families of Theorem 4, $d_i \perp Ae$ and $\|d_i\|^2 = \sigma^2/M$ for every i, so all ray inputs $H_i e = Ae + d_i$ have the same norm; clipping rescales them by a common factor and leaves the frozen field $U(e)$ at $\epsilon = 0$ unchanged. For the default-memory instance at $\kappa = 9$, Appendix H.1 reports clipped and decayed runs with their settings, horizons and radii.

Decoupled weight decay, $x \leftarrow (1 - \gamma_t\lambda)x - \gamma_t u_t$ with $\lambda > 0$, keeps every trajectory bounded: since $\|u_t\| \leq B_u$ (Lemma 3), one has $\|x_t\| \leq \max(\|x_0\|, B_u/\lambda)$ whenever $\gamma_t\lambda \leq 1$. The iterates nevertheless need not approach the minimizer. For the memoryless family the decay acts radially, the signs of the construction persist, and the radius converges to a/λ with $a = 1$ exactly:

Proposition 5 (Spurious limit under decoupled weight decay)**.** *Take the memoryless family of Appendix E with $n = 4$, $q = b = 0$ and $\epsilon = 0$, and replace each sample update by $x \leftarrow (1 - \gamma_t\lambda)x - \gamma_t \operatorname{sgn}(\nabla f_{i_t}(x))$, $\lambda > 0$, $\gamma_t = \eta_k/4$. Let η_k be positive and nonincreasing, with $\eta_k \to 0$ and $\sum_k \eta_k = \infty$, and let $x_0 = R_0 e$ with $200M\eta_0 \leq \min(R_0, 1/\lambda)$. Then $x_k \to \lambda^{-1} e \neq 0$, so $f(x_k) \to f(e)/\lambda^2 > 0 = f_*$. The same holds under any positive rescaling of the sample gradients, including clipping.*

Proof. Put $\theta = 1 - \lambda\eta_k/4$ and write $x_k = R_k e + z_k$. While the four signs s_i of the construction persist, the epoch map is exactly $x_{k+1} = \theta^4 x_k - (\eta_k/4)\sum_{i=0}^{3} \theta^{3-i} s_i$. As $\sum_i s_i = -4e$ and $|\theta^{3-i} - \bar{\theta}| \leq 3\lambda\eta_k/4$ for the mean $\bar{\theta}$ of these powers,

$$R_{k+1} = \theta^4 R_k + (1 - \theta^4)R_k^*, \qquad \|z_{k+1}\|_\infty \leq (1 - \lambda\eta_k/2)\|z_k\|_\infty + \tfrac{3}{4}\lambda\eta_k^2,$$

where $R_k^* = \eta_k\bar{\theta}/(1 - \theta^4) \in [1/(4\lambda), 2/\lambda]$ and $R_k^* \to 1/\lambda$. Hence $R_k \geq \min(R_0, 1/(4\lambda))$ and $\|z_k\|_\infty \leq \frac{3}{2}\eta_0$ for all k, while $R_k \to 1/\lambda$ and $z_k \to 0$. Decay rescales positions within an epoch by at least $\theta^4 \geq 1/2$, which does not change signs, so the within-epoch perturbation of $H_i(R_k e + z_k)$ is at most $25(\|z_k\|_\infty + 2\eta_k)$. The assumption on η_0 keeps it below $2R_k/M$, half the smallest input magnitude $4R_k/M$ at $R_k e$. Thus all signs persist, which closes the induction. At $q = b = \epsilon = 0$ positive rescaling of a sample gradient does not change its sign. □

Appendix H.1 runs the certified $\kappa = 9$ instance at the default memory with clipping and decoupled decay.

Clipping and decay leave the mechanism of Sections 3 and 4 intact. That mechanism rests on Adam's second moment itself: because v_t forgets large inputs, a phase with a small gradient still receives a

normalized update of order one. A running maximum of v_t, as in AMSGrad, removes exactly this mechanism.

Proposition 6 (Running-maximum normalization)**.** *Replace v_t in the denominator by $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$, $\hat{v}_0 = 0$, with any admissible pair, $\epsilon \geq 0$ and either bias convention. For a frozen periodic list at x the stationary running maximum is the period maximum of v^*, constant in time, and the frozen field is*

$$U^{\max}(x) = g(x) \oslash \Big(\sqrt{\max_i v_i^*(x)} + \epsilon \mathbf{1}\Big),$$

with $0/0 = 0$. Hence $g(x)^\top U^{\max}(x) > 0$ whenever $g(x) \neq 0$, for every heterogeneity, and no quadratic finite sum with a common minimizer has a direction e with $U^{\max}(e) = -ae$, $a > 0$.

Proof. In the stationary periodic regime the running maximum has absorbed a complete period, so it equals $\max_i v_i^*$ at every phase. The first moments m_i^* are exponential averages of a periodic input, so their period mean is g. Averaging $m_i^*/(\sqrt{\max_i v_i^*} + \epsilon)$ over phases gives the formula; bias correction tends to one. The pairing is $\sum_j g_j^2/(\sqrt{\max_i v_{ij}^*} + \epsilon)$, positive when $g \neq 0$. If $U^{\max}(e) = -ae$ with $a > 0$, then $g = Ae$ gives $g^\top U^{\max}(e) = -a\, e^\top Ae < 0$, a contradiction. □

D Positive-definite lifts with identity average

The scalar counterexamples use concave components. The lift below makes the components positive definite with identity average and keeps D_1: every coordinate sees a shifted copy of the scalar input. We verify positive definiteness and every transverse mode before applying the transfer theorem.

D.1 Lift algebra

Let $d = (d_0, \dots, d_{n-1})$ have mean one and second moment C. With indices modulo n, define

$$g_i[j] = d_{i+j}, \qquad Q_i = g_i g_i^\top/n, \qquad K = \frac{1}{n}\sum_i Q_i, \qquad H_i = I - K + Q_i. \tag{66}$$

Write $\widehat{d}_\ell = \sum_j d_j e^{-2\pi \mathrm{i}\ell j/n}$, $\mathbf{1} = (1, \dots, 1)^\top$, and $P_{\mathbf{1}} = \mathbf{1}\mathbf{1}^\top/n$.

Lemma 6 (Algebra of the cyclic lift)**.** *The eigenvalues of K are $\kappa_\ell = |\widehat{d}_\ell|^2/n^2$, with $\kappa_0 = 1$. If every nonconstant $\kappa_\ell < 1$, then $H_i \succ 0$, their average is I, $H_i\mathbf{1} = g_i$, and*

$$\frac{1}{n}\sum_i H_i^2 = I + CK - K^2, \qquad D_1 = n \max_\ell (1 + C\kappa_\ell - \kappa_\ell^2). \tag{67}$$

In particular, $C > 2$ and $K \preceq I$ imply $D_1 = nC$.

Proof. Cyclic shifts make K circulant with the stated Fourier eigenvalues. Since $g_i^\top \mathbf{1} = n$, $K\mathbf{1} = \mathbf{1}$ and $Q_i\mathbf{1} = g_i$, proving the ray action and average identity. If $z^\top H_i z = z^\top(I-K)z + (g_i^\top z)^2/n = 0$, positivity on $\mathbf{1}^\perp$ first forces z to be constant; then $g_i^\top z = 0$ forces $z = 0$. Thus $H_i \succ 0$. Finally $Q_i^2 = CQ_i$; expanding and averaging H_i^2 gives (67). The function $1 + C\kappa - \kappa^2$ is increasing on $[0, 1]$ for $C > 2$ and equals C at $\kappa = 1$. □

D.2 Positive memory: lifting the scalar profile

Fix $0 \leq q \leq b < 1$ with $b > 0$, and let d be the scalar profile (24): np_n copies of B_n followed by $n(1-p_n)$ copies of $-a_n$, with mean one and second moment C_n from (25). Since $r_n = O_{q,b}(\log n)$, we have $p_n \to 1/2$, $B_n \to 2$ and $C_n \to 2$. We lift d by (66) in dimension $p = n$. For a profile with one positive and one negative value, a common shift of the inputs never decreases the normalized update, and this response enters every transverse eigenvalue with the stabilizing sign. The lift therefore needs no heterogeneity beyond the scalar construction.

Lemma 7 (Uniform positive definiteness)**.** *For all sufficiently large even n, the lift has average I, $D_1 = nC_n$, $\lambda_{\min}(H_i) \geq 1/13$ and $\lambda_{\max}(H_i) < 4$. In particular $\mathrm{cond}(H_i) < 52$.*

Proof. For $\ell \neq 0$ and $r = \min(\ell, n-\ell)$,

$$\sqrt{\kappa_\ell} = \frac{B_n + a_n}{n} \frac{|\sin(\pi \ell p_n)|}{|\sin(\pi \ell / n)|}, \qquad B_n + a_n = \frac{1 + a_n}{p_n}. \tag{68}$$

For $r \geq 2$, $\sin(\pi r/n) \geq 2r/n$ gives $\sqrt{\kappa_\ell} \leq (B_n + a_n)/(2r) \leq 3/4$ once $B_n + a_n \leq 3$. For $r = 1$, this square root is $(B_n + a_n)\cos(\pi r_n/n)/[n \sin(\pi/n)] \to 2/\pi < 3/4$. Thus all nonconstant eigenvalues are eventually at most $3/4$ and Lemma 6 applies.

Put $\rho_* = 1 - \lambda_{\max}(K|_{\mathbf{1}^\perp}) \geq 1/4$. Since $I - K \succeq \rho_*(I - P_{\mathbf{1}})$, the lower comparison on $\operatorname{span}\{\mathbf{1}, g_i - \mathbf{1}\}$ is $\begin{pmatrix} 1 & \sqrt{C_n - 1} \\ \sqrt{C_n - 1} & C_n - 1 + \rho_* \end{pmatrix}$, whose determinant is ρ_* and trace $C_n + \rho_*$. The orthogonal complement has lower bound ρ_*. As $C_n < 3$ eventually, this gives $\lambda_{\min}(H_i) \geq \rho_*/(C_n + \rho_*) \geq 1/13$. Also $\lambda_{\max}(H_i) \leq 1 + C_n < 4$. □

Lemma 8 (Radial drift and signed transverse stability)**.** *Let J_n be a coordinate of $U(\mathbf{1})$ and let $\mu_{n,\ell}$ be the eigenvalue of $DU(\mathbf{1})$ in nonconstant Fourier mode ℓ. For all sufficiently large even n,*

$$J_n \leq -\frac{r_n}{2n} < 0, \qquad \min_{\ell \neq 0} \operatorname{Re} \mu_{n,\ell} \geq \frac{b}{32\sqrt{2}}. \tag{69}$$

Proof. Along the ray every coordinate receives a cyclic shift of d, so J_n is the scalar drift of Lemma 1.

Cyclic equivariance makes $DU(\mathbf{1})$ circulant. To identify its eigenvalue in a mode of frequency $\theta = 2\pi\ell/n$, put $v_\ell[j] = e^{\mathrm{i}\theta j}$ and $c_\ell = n^{-1} \sum_j d_j e^{\mathrm{i}\theta j}$. The lift gives

$$(H_i v_\ell)[j] = (1 - \kappa_\ell) v_\ell[j] + c_\ell d_{i+j} e^{-\mathrm{i}\theta i}.$$

After dividing by the spatial factor $v_\ell[j]$ and writing $t = i + j$, the temporal input perturbation is

$$\phi_t = (1 - \kappa_\ell) + c_\ell d_t e^{-\mathrm{i}\theta t}, \qquad |c_\ell|^2 = \kappa_\ell.$$

Writing P_q, P_b for temporal exponential filters, its update derivative is

$$\delta\psi = \frac{P_q \phi}{\sqrt{V}} - \frac{M P_b(2d\phi)}{2V^{3/2}}. \tag{70}$$

For each coordinate, averaging over one epoch averages these same temporal phases. Hence $\mu_{n,\ell} = n^{-1} \sum_t \delta\psi_t$. By linearity, $\delta\psi = (1 - \kappa_\ell)\delta\psi^{(1)} + \delta\psi^{\mathrm{osc}}$, where $\delta\psi^{(1)}$ is the response to $\phi \equiv 1$ and $\delta\psi^{\mathrm{osc}}$ the response to $c_\ell d_t e^{-\mathrm{i}\theta t}$. Thus

$$\mu_{n,\ell} = (1 - \kappa_\ell)\chi_n + \omega_{n,\ell}, \qquad \chi_n = \frac{1}{n} \sum_t \delta\psi_t^{(1)}, \qquad \omega_{n,\ell} = \frac{1}{n} \sum_t \delta\psi_t^{\mathrm{osc}}.$$

Since $1 - \kappa_\ell \geq 1/4$ by Lemma 7, it suffices to show $\chi_n \geq b/(4\sqrt{2})$ and $\sup_{\ell \neq 0} |\omega_{n,\ell}| = O_{q,b}(\log n/n)$.

Common shift. With $R_b = P_b d$, $\delta\psi^{(1)} = (V - MR_b)/V^{3/2}$. Write $a = a_n$, $B = B_n$, and let $u = P_q \mathbf{1}_+$ and $w = P_b \mathbf{1}_+$ be the filtered indicator of the positive block. Then $M = -a + (B + a)u$, $R_b = -a + (B + a)w$ and $V = a^2 + (B^2 - a^2)w$, so

$$V - MR_b = (B + a)\big[Bw(1 - u) + au(1 - w)\big] \geq 0. \tag{71}$$

Every common-shift response is therefore nonnegative. Let u_0, w_0 be the states at the last positive phase, so that $u_s = q^s u_0$ and $w_s = b^s w_0$ after s negative inputs. The positive block has length at least $n/4$, so $w_0 \geq 1 - b^{n/4} \geq 1/2$ for large n. As $B^2 w_0 \geq 2 > a^2$, the first s with $B^2 w_s \leq a^2$ is positive, satisfies $ba^2 < B^2 w_s \leq a^2$, and is $O_{q,b}(\log n)$, so it lies in the negative block. Since $q \leq b$ and $B \geq 2$, $u_s \leq b^s \leq 2w_s \leq 1/2$; also $V_s \leq a^2 + B^2 w_s \leq 2a^2$. Hence (71) gives

$$\delta\psi_s^{(1)} \geq \frac{B^2 w_s (1 - u_s)}{V_s^{3/2}} > \frac{ba^2/2}{(2a^2)^{3/2}} = \frac{bn}{4\sqrt{2}},$$

and discarding the other, nonnegative, phases gives $\chi_n \geq b/(4\sqrt{2})$.

Oscillation. Weighted Cauchy–Schwarz gives $|P_q(de^{-\mathrm{i}\theta\cdot})| \le R_{q,b}\sqrt{V}$, $|P_b(d^2e^{-\mathrm{i}\theta\cdot})| \le V$ and $|M| \le R_{q,b}\sqrt{V}$. Since $|c_\ell| \le 1$, $|\delta\psi_t^{\mathrm{osc}}| \le 2R_{q,b}$ at every phase, so the at most $4L_n$ phases whose latest $2L_n$ inputs are not all equal contribute at most $8R_{q,b}L_n/n$ to $|\omega_{n,\ell}|$. On an ideal constant block with value c the oscillatory response is

$$c_\ell \operatorname{sgn}(c)[\widehat{P}_q(\theta) - \widehat{P}_b(\theta)]e^{-\mathrm{i}\theta t}, \qquad \widehat{P}_\gamma(\theta) = \frac{1-\gamma}{1-\gamma e^{\mathrm{i}\theta}}.$$

The bounds $|c_\ell| \le (B_n + a_n)/(n|\sin(\theta/2)|)$ and $|\widehat{P}_q - \widehat{P}_b| \le 2|b-q||\sin(\theta/2)|/[(1-q)(1-b)]$ make this $O_{q,b}(1/n)$ uniformly in every nonconstant mode. To control the tail error, write $\Phi(A, M, E, V) = A/\sqrt{V} - ME/(2V^{3/2})$ with $A = P_q\phi$, $E = P_b(2d\phi)$ for the oscillatory ϕ. Its partial derivatives are $V^{-1/2}$, $-E/(2V^{3/2})$, $-M/(2V^{3/2})$, and $-A/(2V^{3/2})+3ME/(4V^{5/2})$. The states are bounded and $V \ge n^{-2}$ along the segment between true and ideal states, so $\|D\Phi\| = O(n^5)$. Those states differ by $O(b^{2L_n}) = O(n^{-12})$, giving tail error $O(n^{-7})$. Hence $\sup_{\ell\neq 0}|\omega_{n,\ell}| = O_{q,b}(\log n/n)$ and, for all sufficiently large n,

$$\operatorname{Re}\mu_{n,\ell} \ge \frac{\chi_n}{4} - |\omega_{n,\ell}| \ge \frac{b}{16\sqrt{2}} - O_{q,b}\Big(\frac{\log n}{n}\Big) \ge \frac{b}{32\sqrt{2}}.$$

□

Theorem 18 (Identity-average divergence with logarithmic overhead)**.** *For each fixed $0 \le q \le b < 1$ with $b > 0$, every sufficiently large even n gives positive-definite components in dimension $p = n$ with average Hessian I, component condition numbers below 52, and*

$$D_1 = nC_n = 2n + O_{q,b}(\log n),$$

the value of the scalar instance in Theorem 10, yet cyclic Adam diverges from every sufficiently large $r_0\mathbf{1}$. This holds for zero initial moments, either bias convention, every finite $\epsilon \ge 0$, and every positive nonincreasing epoch scale with divergent sum, using sample step η_k/n.

Proof. Take $e = \mathbf{1}$ and $\ell(x) = n^{-1}\mathbf{1}^\top x$. Cyclic equivariance gives $U(e) = J_n e$ with $J_n < 0$ and transverse angular eigenvalues $J_n - \mu_{n,\ell}$, whose real parts are below $-b/(32\sqrt{2})$ by Lemma 8. Lemma 7 and (25) give $D_1 = nC_n \le 2n + 8r_n + 7$. Every coordinate of $H_i e = g_i$ is nonzero. Theorem 17 gives divergence of the actual iterates with every stated option, and Lemma 7 gives the geometry. □

A finite standard-memory certificate. Take $(q, b) = (9/10, 999/1000)$, $p = D = 80$, $L = 1000$, $n = DL = 80{,}000$. The length-D template has 34 copies of $4023/1700$ followed by 46 copies of $-1/100$. For its cyclic shifts s_r, set $K = D^{-2}\sum_r s_r s_r^\top$ and $H_r = I - K + s_r s_r^\top/D$. The sample order uses L consecutive copies of each H_r. Exact arithmetic gives average I, $C = 404623/170000$ and $D_1 = nC = 3236984/17$. The all-mode interval certificate gives

$$J < -153/10000, \qquad \lambda_{\max}(K|_{\mathbf{1}^\perp}) < 271/500, \qquad \max \operatorname{Re}\operatorname{spec}(DF|_{\mathbf{1}^\perp}) < -117/100.$$

It covers all 79 transverse spatial modes and the full 80,000-step period, with Arb runs at 256 and 512 bits. The comparison matrix in Lemma 7, now with $\rho_* = 229/500$, gives $\lambda_{\min}(H_r) \ge \rho_*/(C + \rho_*)$ and $\lambda_{\max}(H_r) \le 1 + C$, hence condition numbers below 21. Theorem 17 applies. The supplement indexes the certificate and its recomputation commands.

E Divergence near minimum aggregate growth

We first prove the cyclic part of Theorem 4, then extend it to random sampling and give a finite cyclic default-memory instance with a convergent twin. Throughout, $A = n^{-1}\sum_i H_i$ and D_1 denotes the optimal constant in the global quadratic growth inequality. The construction has $\operatorname{cond}(A) \to \infty$.

E.1 POSITIVE SECOND-MOMENT MEMORY

Fix $0 \le q \le b < 1$ with $b > 0$. For a scalar period $a = (a_i)_{i=0}^{n-1}$, write $J(a)$ for the mean stationary direction $m/\sqrt{v}$. We first choose a period satisfying

$$\bar{a} > 0, \qquad J(a) = -c < 0, \qquad \chi := DJ(a)[\mathbf{1}] > 0, \qquad a_i \neq 0, \qquad c\bar{a} > 1. \tag{72}$$

Such a period exists for every fixed admissible pair. Start with a negative-drift positive-mean period from Lemma 1. The function $\phi(s) = J(a + s\mathbf{1})$ is smooth: the periodic second moments stay positive because the period is nonconstant. It is negative at zero and tends to one as $s \to \infty$. Before its first zero there is therefore a point with $\phi < 0$ and $\phi' > 0$. A small change of this shift avoids the finitely many zero-input phases. Positive rescaling preserves the drift and the sign of the shift derivative and makes $c\bar{a} > 1$. Homogeneity also gives

$$DJ(a)[a] = 0. \tag{73}$$

Average and component matrices. The tail coordinates carry the negative scalar drift, while the head has a positive gradient and moves in the opposite parameter direction. Replicating the tail dilutes its contribution to aggregate variance without changing its normalized Adam direction. Let M tend to infinity through multiples of n. The ambient space consists of M tail coordinates and one head coordinate, whose unit vector is e_0. Put $u = M^{-1/2}\mathbf{1}_M$ and let W be the tail-zero-sum subspace. Choose fixed $\lambda > 0$ and $\delta > 0$ with $\delta\chi/c^2 < 1$, and define

$$\gamma = \frac{1+\delta}{c\bar{a} - 1}, \qquad e = c\mathbf{1}_M - e_0, \qquad \ell(z) = -z_{\mathrm{head}}.$$

Thus $\ell(e) = 1$. Set $A = \lambda I$ on W and, on the ordered orthonormal basis (u, e_0), set

$$A_2 = \begin{pmatrix} (c\bar{a} + 1 + \gamma)/(c^2 M) & (1+\gamma)/(c\sqrt{M}) \\ (1+\gamma)/(c\sqrt{M}) & \gamma \end{pmatrix}. \tag{74}$$

Its determinant is $\delta/(c^2 M) > 0$, and

$$Ae = (\bar{a}/M)\mathbf{1}_M + e_0, \qquad e^\top Ae = c\bar{a} - 1 > 0. \tag{75}$$

Split the tail into n phase groups of size M/n. For a coordinate j in phase group r, put

$$d_{i,j} = \frac{a_{i+r} - \bar{a}}{M}, \qquad d_{i,\mathrm{head}} = 0, \qquad v = Ae_0, \qquad E_i = d_i v^\top + v d_i^\top, \qquad H_i = A + E_i.$$

All period indices are modulo n. With $\sigma^2 = n^{-1}\sum_i (a_i - \bar{a})^2$, one has

$$\sum_i d_i = 0, \qquad d_i \in W, \qquad \|d_i\|^2 = \sigma^2/M, \qquad H_i e = Ae + d_i.$$

Here $v^\top e = 1$ and $d_i^\top e = 0$. Since $Ad_i = \lambda d_i$, the only nonzero eigenvalues of $A^{-1/2}E_iA^{-1/2}$ are

$$\pm\sqrt{\frac{\gamma\sigma^2}{\lambda M}}.$$

Consequently all H_i are positive definite for sufficiently large M, and their average is exactly A.

Global aggregate constant. The mean perturbation vanishes, so

$$\frac{1}{n}\sum_i \|H_i z\|^2 = \|Az\|^2 + \frac{1}{n}\sum_i \|E_i z\|^2.$$

Writing $B_M = n^{-1}\sum_i d_i d_i^\top$, orthogonality gives the exact identity

$$\frac{1}{n}\sum_i (E_iA^{-1})^\top (E_iA^{-1}) = \frac{\sigma^2}{M} e_0 e_0^\top + \frac{\|v\|^2}{\lambda^2} B_M. \tag{76}$$

Indeed $E_iA^{-1} = d_i e_0^\top + \lambda^{-1} v d_i^\top$, with $d_i \perp v$. The two terms on the right act on orthogonal subspaces, and $\|B_M\| \le \operatorname{tr} B_M = \sigma^2/M$. Therefore

$$1 + \frac{\sigma^2}{M} \le \frac{D_1}{n} \le 1 + \frac{\sigma^2}{M}\max\{1, \|v\|^2/\lambda^2\} = 1 + O(M^{-1}). \tag{77}$$

Replication keeps the nonzero spectrum of MB_M fixed. Since $\|v\|^2 = \gamma^2 + (1+\gamma)^2/(c^2 M)$, the exact maximum in (76) decreases to zero, giving $D_1/n \downarrow 1$.

Frozen ray and transverse spectrum. At e, each tail coordinate sees a cyclic shift of a/M, and the head sees the constant input one. Thus every input coordinate is nonzero and

$$U(e) = -c\mathbf{1}_M + e_0 = -e. \tag{78}$$

For the section $\ell(y) = 1$, use $F(y) = -U(y) + y\ell(U(y))$. Its tangent space splits into within-group fluctuations, group means with zero total mean, and the uniform tail direction. We check these three blocks.

First, if a tail perturbation w sums to zero inside each phase group, then $d_i^\top w = v^\top w = 0$ and $H_i w = \lambda w$. A coordinate is perturbed by the constant temporal input λw_j. Since the base period is scaled by $1/M$,

$$DU(e)w = \lambda M\chi w, \qquad DF(e)w = -(1 + \lambda M\chi)w.$$

Next let $w_j = z_r$ within group r, with $\sum_r z_r = 0$. Define the fixed phase map

$$(Tz)_i = \frac{1}{n}\sum_r (a_{i+r} - \bar{a})z_r.$$

Then $d_i^\top w = (Tz)_i$, and the input perturbation in tail group r is

$$\lambda z_r \mathbf{1} + \frac{1+\gamma}{cM} Tz.$$

Consequently the tail derivative is $\lambda M\chi z + \mathcal{K}z$, where

$$(\mathcal{K}z)_r = \frac{1+\gamma}{c} DJ(\mathrm{shift}_r\, a)[Tz].$$

The matrix $\mathcal{K}$ is independent of M. The mean derivative of the head direction is zero: at the constant input one, linearizing $m/\sqrt{v}$ gives the difference of the q- and b-EMA filters, whose period means agree. Thus the phase block of $DF(e)$ is

$$-(1 + \lambda M\chi)I - \mathcal{K}, \tag{79}$$

which is Hurwitz for large M. Replication multiplies the input and output norms by the same factor. Shift invariance of J gives $\sum_r DJ(\mathrm{shift}_r\, a) = \chi\mathbf{1}^\top$, while $\mathbf{1}^\top T = 0$. Hence $\mathbf{1}^\top\mathcal{K} = 0$, so $\mathcal{K}$ preserves the phase-zero-sum subspace.

The remaining section direction is $y(\zeta) = \zeta\mathbf{1}_M - e_0$. Multiplying its tail input period by M gives

$$P_i(\zeta) = \frac{c\bar{a} + 1 + \gamma}{c^2}\zeta - \frac{1+\gamma}{c} + (a_i - \bar{a})\left(\frac{1+\gamma}{c}\zeta - \gamma\right).$$

At $\zeta = c$, $P = a$ and $P' = ((1+\gamma)/c)a - (\delta/c^2)\mathbf{1}$. Equation (73) yields $\frac{d}{d\zeta}J(P(\zeta))|_c = -\delta\chi/c^2$. The head direction remains one near the ray, so the uniform projective eigenvalue is

$$-1 + \delta\chi/c^2 < 0. \tag{80}$$

All section modes are Hurwitz. Theorem 17 now applies to (78): from zero moments and $x_0 = R_0 e$ with sufficiently large R_0,

$$\ell(x_k) \to \infty, \qquad x_k/\ell(x_k) \to e, \qquad \ell(x_{k+1}) - \ell(x_k) = \eta_k + o(\eta_k).$$

That theorem supplies both bias-correction choices and every fixed finite ϵ for the prescribed non-summable schedule.

E.2 An explicit memoryless construction

For $q = b = 0$, take $n = 4$, $M > 288$ divisible by four, and set $A = I$ on W and

$$A_2 = \begin{pmatrix} 18/M & 5/\sqrt{M} \\ 5/\sqrt{M} & 3/2 \end{pmatrix}, \qquad e = \tfrac{1}{2}\mathbf{1}_M - e_0.$$

Its determinant is $2/M$, and $Ae = (4/M)\mathbf{1}_M + e_0$. Split the tail into four equal groups I_i. Define d_i to equal $24/M$ on I_i, $-8/M$ on the other tail groups, and zero on the head. With $v = Ae_0$ and

$H_i = A + d_i v^\top + v d_i^\top$, the relative perturbation eigenvalues are $\pm\sqrt{288/M}$, hence $H_i \succ 0$. The nonzero eigenvalues of $\frac{1}{4}\sum_i d_i d_i^\top$ are $64/M$, whereas $\|v\|^2 = 9/4 + 25/M$. Equation (76) therefore gives exactly

$$D_1/n = 1 + 192/M. \tag{81}$$

At e, component i has gradient $28/M$ on I_i, $-4/M$ on the other tail groups, and one on the head. Their sign vectors s_i satisfy $\frac{1}{4}\sum_i s_i = -e$. The matrix row sums obey $\|H_i\|_{\infty\to\infty} < 25$: one has $\|A\|_{\infty\to\infty} \le 13/2$, $\|v\|_1 = 13/2$, $\|d_i\|_1 = 12$, the head row of the perturbation has absolute sum 18, and its tail rows have sum at most $216/M$. At $\epsilon = 0$, if an epoch starts at $R_k e$ with $R_k > (25/4)M\eta_0$, all four signs persist during its within-epoch motion. Consequently

$$x_{k+1} = x_k + \eta_k e, \qquad x_k = \left(R_0 + \sum_{h<k} \eta_h\right) e.$$

For finite ϵ, put $R_k = R_0 + \sum_{h<k} \eta_h$ and $x_k = R_k e + z_k$. Whenever $25(\|z_k\|_\infty + \eta_k) \le 2R_k/M$, all actual tail gradients have magnitude at least $2R_k/M$ and retain their signs. The scalar direction $g/(|g| + \epsilon)$ differs from its sign by at most $\epsilon/|g|$, so

$$\|z_{k+1}\|_\infty \le \|z_k\|_\infty + \eta_k \epsilon M/(2R_k).$$

Taking $R_0 \ge \eta_0$, a discrete logarithmic comparison gives $\sum_{h<k} \eta_h/R_h \le 2\log(R_k/R_0)$ and hence $\|z_k\|_\infty \le \epsilon M \log(R_k/R_0)$. Because $\log t/t \le 1/\mathrm{e}$ for $t \ge 1$, the choice $R_0 \ge 50M\eta_0 + 25\epsilon M^2$ closes the sign-tube induction. Thus $z_k = o(R_k)$, and the trajectory also diverges at finite ϵ. Bias correction is identically one at this corner.

E.3 IID SAMPLING AND FRESH RANDOM RESHUFFLING

For IID sampling, fix a batch size $s \ge 1$: each update averages s independent uniform labels and uses step η_k. For fresh reshuffling, a block is one independent uniform permutation with sample step η_k/n. Thus a block has $l = 1$ or $l = n$ updates. For an array a, let $J(a)$ be its stationary expected block-average direction. The original-data ratio remains $C = D_1/n$; independence gives the exact optimal batch ratio

$$C_s = 1 + (C-1)/s, \qquad \mathbb{E}\|s^{-1}\sum_{j=1}^{s} H_{I_j} x\|^2 = \|Ax\|^2 + s^{-1}n^{-1}\sum_i \|(H_i - A)x\|^2. \tag{82}$$

The finite sum is chosen after s, and different protocols may use different families.

A negative scalar seed. For $b > 0$ take $a = (2n-1, -1, \dots, -1)$, of mean one, with $n > s$. An IID batch input is $-1 + (2n/s)K$, $K \sim \mathrm{Binomial}(s, 1/n)$; it is never zero and has magnitude in $[1, 2n-1]$. If the last L updates had only negative inputs, stationary filters satisfy $m \le -1 + 2nq^L$ and $1 \le v \le 1 + 4n^2 b^L$. Choose $L = \lceil \log(16n^2)/(-\log b) \rceil$. Then $m/\sqrt{v} \le -4/5$. The exceptional probability is at most sL/n for IID batches and L/n averaged over reshuffling phases. With $s = 1$ for the latter, the update bound gives

$$J(a) \le -\frac{4}{5} + \left(R_{q,b} + \frac{4}{5}\right)\frac{sL}{n} < 0$$

for sufficiently large finite n. The function $J(a + t\mathbf{1})$ is smooth for $0 \le t < 1$ because every possible batch input is bounded away from zero on compact subintervals. At $t = 1$ the stream is nonnegative and its expected update is positive. Dominated convergence uses $R_{q,b}$; a last positive input exists almost surely, so its second moment is nonzero. The mean-value theorem before the first zero supplies a shift with $J = -c < 0$ and $DJ[\mathbf{1}] = \chi > 0$. Positive rescaling supplies (72), and homogeneity supplies (73).

Use exactly (74) and the matrices of the positive-memory construction. Positivity and the global constant (76) do not depend on sample order. Relabeling invariance gives $U(e) = -e$ and $\sum_r DJ(\mathrm{shift}_r\, a) = \chi \mathbf{1}^\top$. For IID batches J acts on the averaged inputs, so the same identity includes their cross-coordinate correlations. At the constant head stream the expected derivative is the difference of two EMA means, hence zero. Thus the same three transverse blocks (79)–(80), including the within-group block, are Hurwitz for large M. Every component head gradient is positive near e.

Stochastic ray transfer. Write a block label as $\omega = (H_{\omega,1}, \dots, H_{\omega,l})$; the labels are IID with finite support. Let $U(y)$ be the stationary expected frozen block-average direction at $\epsilon = 0$, and set $F(y) = -U(y) + y\ell(U(y))$, $\ell = -e_0^\top$.

Theorem 19 (Stochastic ray transfer)**.** *Fix $0 \le q \le b < 1$. Suppose $\ell(e) = 1$, every $H_{\omega,j}e$ has nonzero coordinates and the same positive head coordinate h_*, $U(e) = -e$, and $DF(e)|_{\ker \ell}$ is Hurwitz. There are a section neighborhood Ω_0 and constants $K_*, c_* > 0$, depending only on the fixed block law, matrices, memory pair and ray, such that uniformly for $y_0 \in \Omega_0$, every deterministic positive nonincreasing nonsummable schedule, fixed $\epsilon \ge 0$, and either bias convention, zero-moment Adam from $x_0 = r_0 y_0$ with sample step η_k/l satisfies*

$$r_k \to \infty, \qquad x_k/r_k \to e, \qquad r_k/\sum_{j<k} \eta_j \to 1, \qquad r_k = \ell(x_k), \tag{83}$$

with probability at least $1 - \exp(-c_* r_0/\eta_0)$ whenever $r_0 \ge K_*(\eta_0 + \epsilon)$. All sample iterates diverge on this event. The result includes $q = b = 0$.

Proof. On $T = \ker \ell$ choose $P \succ 0$ and $0 < \delta \le 1$ so that $\mathcal{K} = \{\ell(y) = 1 : |y - e|_P \le 2\delta\}$ has fixed input signs, magnitudes in $[a, B]$, positive heads, and $DV(z)F(e+z) \le -\lambda V(z)$ for $V(z) = |z|_P^2$, $\lambda > 0$. Set $\Omega_0 = \{|y - e|_P \le \delta/4\}$, $\mathcal{F}_k = \sigma(\omega_0, \dots, \omega_{k-1})$ and $\tau = \inf\{k : |y_k - e|_P \ge \delta\}$, where $y_k = x_k/r_k$. The update bound gives within-block motion $K_y \eta_k/r_k$. For $K_y \eta_0/r_0 \le \delta/4$, first exit keeps all samples of every block $k < \tau$, including its endpoint, in $\mathcal{K}$. Positive head moments make sample radii nondecreasing; raw moments obey $|m_j| \le Br$, $0 \le v_j \le B^2 r^2$, also for corrected v_j. The current head contributes $(1-q)ar$ to m, so $r_0 \ge \epsilon$ gives

$$c_r \eta_k \le r_{k+1} - r_k \le C_r \eta_k \quad (k < \tau), \qquad c_r = (1-q)a/(B+1) > 0. \tag{84}$$

With $h_k = \eta_k/r_k$ and $C_r h_0 \le 1$, telescoping gives

$$\sum_{k<\tau} \frac{\eta_k}{r_k^2} \le \frac{B_h}{r_0}, \qquad \sum_{k<\tau} h_k^2 \le B_h \frac{\eta_0}{r_0}, \qquad \sum_{k<\tau} |h_{k+1} - h_k| \le h_0, \quad B_h = 2/c_r. \tag{85}$$

Indeed $\eta_k/r_k^2 \le B_h(1/r_k - 1/r_{k+1})$ and h_k decreases. On no exit, $h_k \to 0$ and $\sum h_k = \infty$.

The incoming raw state $Z_k = (m_k/r_k, v_k/r_k^2)$ belongs to $\mathcal{Z} = [-B, B]^p \times [0, B^2]^p$. The frozen transition is $\Psi_{y,\omega}(Z) = DZ + B(y, \omega)$, $D = \mathrm{diag}(q^l I, b^l I)$, $\rho = \|D\| < 1$. Its reward is $f = -A + y\ell(A)$, where A averages the l post-input directions; their second moments are at least $(1-b)a^2$, even at $Z = 0$. Thus the maps have uniformly bounded local derivatives. Let P_y be the transition operator and $g_y = \mathbb{E}_\omega f$. The stationary series $Z^*(y) = \sum_{j \ge 0} D^j B(y, \omega_{-1-j})$ and its first two parameter derivatives converge uniformly. Common-block coupling contracts state and derivative differences as $O(\rho^j)$, hence F is smooth and

$$H_y = \sum_{j \ge 0} (P_y^j g_y - F(y)), \qquad H_y - P_y H_y = g_y - F(y)$$

converges uniformly in C^1 and takes values in T. For reshuffling, Z_k and an independent whole permutation form this Markov update; no within-block independence is used.

Local Taylor bounds and the moment recursions give, for $k < \tau$,

$$Z_{k+1} = \Psi_{y_k,\omega_k}(Z_k) + O(h_k), \qquad y_{k+1} = y_k + h_k f_k + R_{k+1},$$

where $\|R_{k+1}\| \le K(h_k^2 + \epsilon\eta_k/r_k^2 + h_k\theta^k)$; $0 < \theta < 1$ accounts for bias correction and is absent for raw moments. Write $A_k = H_{y_k}(Z_k)$, $B_k = H_{y_k}(\Psi_{y_k,\omega_k}(Z_k))$. Then $\xi_{k+1} = f_k - g_{y_k}(Z_k) + B_k - P_{y_k} H_{y_k}(Z_k)$ is bounded and $\mathcal{F}_k$-centered (both random terms use the same block), and $f_k - F(y_k) = \xi_{k+1} + A_k - B_k$. Consequently

$$\widetilde{y}_k = y_k + h_k A_k, \qquad \widetilde{y}_{k+1} = \widetilde{y}_k + h_k F(\widetilde{y}_k) + h_k \xi_{k+1} + e_{k+1}, \tag{86}$$

with exact remainder

$$e_{k+1} = R_{k+1} + h_k\{F(y_k) - F(\widetilde{y}_k)\} + (h_{k+1} - h_k)A_{k+1} + h_k(A_{k+1} - B_k).$$

Here $\|A_{k+1} - B_k\| \le K h_k$, so $\|e_{k+1}\| \le K(h_k^2 + |h_{k+1} - h_k| + \epsilon\eta_k/r_k^2 + h_k\theta^k)$ and $\sum_{k<\tau} \|e_{k+1}\| \le K_e(\eta_0 + \epsilon)/r_0$.

For $h_0 \sup|H|_P \le \delta/4$, corrected points through exit lie in $\mathcal{K}$, $V_0 \le \delta^2/4$, and $V_\tau \ge 9\delta^2/16$, where $V_k = V(\widetilde{y}_k - e)$. Taylor's formula gives

$$V_{k+1} \le V_k - \lambda h_k V_k + h_k M_{k+1} + E_{k+1}, \qquad M_{k+1} = DV(\widetilde{y}_k - e)\xi_{k+1}, \quad |M_{k+1}| \le M_0,$$

where $M_0 \ge 1$, $0 \le E_{k+1} \le K(h_k^2 + \|e_{k+1}\|)$ and $\sum_{k<\tau} E_{k+1} \le A_V(\eta_0 + \epsilon)/r_0$. Choose K_* to imply all preceding radius conditions and make this last budget at most $\delta^2/16$. For the stopped martingale $S_N = \sum_{k<N} \mathbf{1}_{\{k<\tau\}} h_k M_{k+1}$, conditional Hoeffding gives the supermartingale $\exp[tS_N - t^2 M_0^2 \sum_{k<N, k<\tau} h_k^2/2]$. By (85) and the maximal inequality,

$$\mathbb{P}(\tau < \infty) \le \mathbb{P}(\sup_N S_N \ge \delta^2/4) \le \exp\left[-\frac{\delta^4 r_0}{32 M_0^2 B_h \eta_0}\right].$$

Put $T_k = \mathbb{E}[\sum_{j \ge k, j<\tau} E_{j+1} \mid \mathcal{F}_k]$. Then $V_{k\wedge\tau} + T_k$ is a nonnegative supermartingale with drift at most $-\lambda \mathbf{1}_{\{k<\tau\}} h_k V_k$; this compensates the possibly nonpredictable errors exactly. The integrable total-error martingale minus its observed partial sum is T_k, so $T_k \to 0$ a.s. Thus $V_{k\wedge\tau}$ converges and $\sum_{k<\tau} h_k V_k < \infty$. On no exit, $\sum h_k = \infty$ forces $V_k \to 0$, hence $y_k \to e$. Within-block section points also tend to e, and consecutive sample radii have ratio tending to one. Geometric moment tails then give normalized head moments (h_*, h_*^2), so the head direction tends to one. Therefore $(r_{k+1} - r_k)/\eta_k \to 1$, proving (83); sample radii are nondecreasing. At $q = b = 0$, $H = 0$ and no argument divides by a memory parameter. For memoryless reshuffling in a sign tube the block reward is deterministic, hence the same radius condition gives probability one. □

Every IID batch matrix $s^{-1}\sum_{j=1}^{s} H_{I_j}$ has nonzero inputs at the constructed ray and head input one, so the theorem applies. At $q = b = 0$, reshuffling uses the preceding $n = 4$ family, whose sign tube has $U(y) = -e$ and $F(y) = e - y$. For IID batches use $n = 4(s+1)$, $c = 2(1-1/n)^s - 1 > 1/2$, $a = (2/c)(2n-1, -1, \dots, -1)$, and the same lift with $\lambda = \delta = 1$, $\gamma = 2$, and $M > 32(n-1)/c^2$ a multiple of n. The expected batch sign is $-c$; components are SPD, and (76) gives $C = 1 + 16(n-1)/(c^2 M) \downarrow 1$ exactly. Locally fixed batch signs give $U(y) = -e$ and $DF(e) = -I$ on all M transverse directions, completing Theorem 4 for every finite batch size.

A finite IID certificate at default memory. Take $q = 9/10$, $b = 999/1000$, $n = 32768$ and $M = 15n = 491520$. Let $w = P = 557055$ with probability $1/n$ and $w = -1$ otherwise, with mean $\nu = 16$ and second moment $s = (n-1)17^2 + 16^2$. Let $J = \mathbb{E}[m/\sqrt{v}]$ for the stationary filters. Condition on the age a of the last positive input, whose probability is $\pi_a = n^{-1}(1-1/n)^a$. The incoming history is independent of this age. Immediately after that input, $m_+ = (1-q)P + qm_{\mathrm{pre}} > 0$ and $v_+ = (1-b)P^2 + bv_{\mathrm{pre}}$. Subsequently $m_a = -1 + q^a(m_+ + 1)$ and $v_a = 1 + b^a(v_+ - 1)$. Put $D = (1-b)P^2 + bs - 1$, $K = 126$ and $L_0 = 50000$. Since $(P+1)q^K \le 1$, the conditional means satisfy

$$\mathbb{E}[m_a/\sqrt{v_a} \mid a] \le B_a := \begin{cases} \dfrac{q^a[(1-q)P + q\nu]}{\sqrt{1 + b^a(1-b)(P^2-1)}}, & a < K, \\ \dfrac{-1 + (P+1)q^a}{\sqrt{1 + Db^a}}, & a \ge K. \end{cases}$$

For the first case discard the negative numerator term. In the second, $m_a \le -1 + (P+1)q^a \le 0$ and Jensen's inequality for $v^{-1/2}$ gives the displayed upper bound, without assuming independence of m_a, v_a. The negative branch B_a decreases, so

$$J \le \sum_{a=0}^{L_0 - 1} \pi_a B_a + (1-1/n)^{L_0} B_{L_0} < -0.57486076214057508. \tag{87}$$

Directed evaluation at 60 and 100 digits certifies the infinite-tail bound; stationary moment comparison then gives $57/100 < c := -J \le 1$.

Shift derivative and matrices. If S_q, S_b filter the Bernoulli positive-input indicators, then $m = -1 + (P+1)S_q$, $r = -1 + (P+1)S_b$ and $v = 1 + (P^2-1)S_b$. Thus

$$v - mr = (P+1)\{PS_b(1-S_q) + S_q(1-S_b)\} > 0 \quad \text{almost surely.}$$

Thus $0 < \chi_w = \mathbb{E}[(v - mr)/v^{3/2}] \le 83/10$, using $v \ge 1$, $|r| \le \sqrt{v}$ and $|m|/\sqrt{v} < 73/10$. For $a = 2w/(c\nu)$, $\bar{a} = 2/c$, $J(a) = -c$ and $\chi = DJ(a)[\mathbf{1}] = (c\nu/2)\chi_w$. Take the matrices (74)

with $\lambda = 1$, $\delta = 1/1600$, $\gamma = 1 + \delta$, and the same cyclic phase groups and rank-two component perturbations. Then

$$\sigma^2 = \frac{4(n-1)17^2}{16^2c^2}, \qquad \frac{\gamma\sigma^2}{M} < 0.927118 < 1,$$

which proves $H_i \succ 0$. On the group-constant zero-sum subspace, B_M has eigenvalue $\sigma^2/[M(n-1)]$; its other eigenvalues vanish. Since $\|Ae_0\|^2 = \gamma^2 + (1+\gamma)^2/(c^2M) < n-1$, (76) gives the *optimal global* ratio

$$C = 1 + \frac{\sigma^2}{M} = 1 + \frac{9469663}{31457280c^2}. \tag{88}$$

Hence $1.3010324796 < C < 1.9265388723$; the two-dimensional mean block gives $\mathrm{cond}(A) < 7.88 \cdot 10^8$.

Full transverse stability. At $e = c\mathbf{1}_M - e_0$, the head input is one and each tail sees the IID profile a/M, hence $U(e) = -e$. The head's expected derivative is zero: the linearized direction there is the difference of two EMA filters, whose expectations agree. The section spectrum, with multiplicities, is

$$\begin{array}{c|c} -1 - M\chi & M-n \\ -1 - [M - 2(1+\gamma)/(c^2(n-1))]\chi & n-1 \\ -1 + \delta\chi/c^2 & 1. \end{array}$$

Indeed, for derivatives in the positive and each negative label, $D_+ + (n-1)D_- = \chi$ and $PD_+ - (n-1)D_- = 0$ by homogeneity. Thus $D_+ - D_- = -\nu\chi/[(n-1)(1+\nu)]$, which gives the group block; within-group perturbations give $H_i z = z$. On $y(z) = z\mathbf{1}_M - e_0$, the scaled tail profile is $[(1+\gamma)z/c - \gamma]a + \delta(c-z)\mathbf{1}/c^2$; homogeneity gives the last eigenvalue. The first two eigenvalues are below -1, and $\chi/c^2 \le \nu(83/10)/[2(57/100)]$ bounds the last by $-1057/1140 < -0.92719$. Nonzero ray inputs now permit Theorem 19, proving true IID divergence.

The constant c is computable: filling inputs before lag T with -1 changes the finite-history expectation by at most $(P+1)q^T + (73/20)P^2b^T \to 0$. All predicates are uniform for $57/100 < c \le 1$; `default_iid/` supplies exact endpoints and commands. Larger multiples M of n preserve the predicates and give $C = 1 + \sigma^2/M \downarrow 1$ at these default memories.

E.4 A FINITE DEFAULT-MEMORY CERTIFICATE AND A MATCHED PAIR

We use a five-block instance of the rank-two construction in Appendix F, with $(q, b) = (9/10, 999/1000)$ and

$$N = 5, \quad L_b = 4000, \quad n = NL_b = 20000, \quad M = 400000, \quad p = M + 1, \quad \kappa = 50000.$$

Repeat each entry of $w = (29/5, -1/5, -1/5, -1/5, -1/5)$ for L_b consecutive phases. Its mean is one and its variance is $\sigma^2 = 144/25$. Define $c = -J(w) > 0$ using these exact periodic moments, $u = M^{-1/2}\mathbf{1}_M$, and $e = c\mathbf{1}_M - e_0$. The average A equals κI on $W = \{u, e_0\}^\perp$ and has matrix

$$A_2 = \begin{pmatrix} \alpha & h \\ h & d \end{pmatrix}, \qquad \alpha = 1 + \frac{(\kappa-1)t_0^2}{1+t_0^2}, \quad h = \frac{(\kappa-1)t_0}{1+t_0^2}, \quad d = \kappa + 1 - \alpha, \quad t_0 = 3/250.$$

Its trace and determinant are $\kappa + 1$ and κ, respectively; thus $\mathrm{cond}_2(A) = 50000$ exactly. Write $g = Ae = g_u u + g_h e_0$ and $v = Ag/\|g\|^2 = v_u u + v_h e_0$. Partition the tail into N equal classes. In block i, let

$$(z_i)_j = \frac{g_u}{\sqrt{M}}(w_{i+r} - 1) \quad (j \text{ in class } r), \qquad (z_i)_{\mathrm{head}} = 0, \qquad H_i = A + z_i v^\top + v z_i^\top,$$

with indices modulo N, and repeat H_i for L_b phases. Since $z_i \perp e, v$, $v^\top e = 1$, and $\sum_i z_i = 0$, the average is A and $H_i e = g + z_i$. The certificate gives $g_u, g_h > 0$; the tail streams are positive multiples of cyclic shifts of w, while the head stream is positive and constant. Hence $U(e) = -e$ exactly.

Global growth and positivity. For $B_M = N^{-1}\sum_i z_i z_i^\top$, the two terms in

$$N^{-1}\sum_i (H_i - A)^2 = \|v\|^2 B_M + g_u^2\sigma^2 vv^\top$$

act on orthogonal subspaces after conjugation by A^{-1}. The second has eigenvalue $g_u^2\sigma^2/\|g\|^2$; the first is no larger, since $\|Ag\| \le \kappa\|g\|$ and $\|B_M\| \le g_u^2\sigma^2$. Therefore the optimal global constant is

$$C = \frac{D_1}{n} = 1 + \frac{144 g_u^2}{25(g_u^2 + g_h^2)}, \qquad 1.0013391888424321 < C < 1.0013391888424326. \tag{89}$$

The relative perturbation $A^{-1/2}(H_i - A)A^{-1/2}$ has only two nonzero eigenvalues, $\pm\rho$, where

$$\rho^2 = (C-1)\frac{g^\top A g}{\kappa\|g\|^2} < .001339175 < 1.$$

Thus every component is positive definite.

All transverse modes. Let $d_i = \partial J(w)/\partial w_i$, where each block value is varied across its L_b phases, and put $\chi = \sum_i d_i$ and $t = g_u/\sqrt{M}$. On the section $-y_{\text{head}} = 1$, the $M - N = 399995$ within-class zero-sum directions have projective eigenvalue $-1 - \kappa\chi/t < -19979.47$. The remaining five directions are class-constant. Their circulant projective Jacobian is

$$K_{rs} = -\mathbf{1}_{r=s} - \frac{1}{t}\sum_i d_{i+r}\left[\kappa\mathbf{1}_{r=s} + \frac{\alpha-\kappa}{N} + \frac{v_u g_u}{N}(w_{i+s} + w_{i+r} - 2)\right].$$

Every Fourier eigenvalue has real part below $-1/2$. At 60 and 100 decimal digits, the supplement certifies the drift $J(w) < -.27055$, positivity, the global C and all 400000 transverse directions (module `integrated_refinement/geometry/`). An independent periodic-adjoint calculation checks all five drift derivatives. Intervals for J are propagated through the definition of e and every geometric quantity. Theorem 17 gives true divergence from sufficiently distant ray initializations.

Corollary 4 (Matched aggregate growth, opposite outcomes)**.** *Let $H_1, \dots, H_n \succ 0$, $n \ge 2$, have average A and optimal ratio $C = D_1/n < 2$. Put*

$$t^2 = \frac{C-1}{n-1}, \qquad a_1 = 1 + (n-1)t, \qquad a_i = 1 - t \quad (i \ge 2),$$

with $t \ge 0$. The components $\widetilde{H}_i = a_i A$ are positive definite and have the same n, p, A and optimal D_1, and cyclic Adam converges on them from every initial point, for every admissible memory pair, fixed finite $\epsilon \ge 0$, bias convention and diminishing schedule. In particular, at $(q, b) = (.9, .999)$ the certified instance above diverges and its twin converges from the same initial point, under the same nonincreasing diminishing schedule, finite ϵ, bias-correction choice and zero moments.

Proof. Since $0 \le t^2 < 1/(n-1) \le 1$, every a_i is positive, so $\widetilde{H}_i \succ 0$. The twin has the same average because $\sum_i a_i = n$, and

$$\sum_i a_i^2 = n + n(n-1)t^2 = nC.$$

Its component gradients are $a_i \nabla f$, so its optimal aggregate and coordinatewise constants are both exactly $nC < 2n$. Theorem 7 gives convergence from every initial point. For the certified instance, C is (89); choose the common initial point on its distant divergent ray. □

F THE SHARP CONDITIONING THRESHOLD FOR DIVERGENT RAYS

We prove Theorem 5; Appendix F.5 proves Theorem 6. Throughout this appendix, $f_i(x) = x^\top H_i x/2$, $H_i \succ 0$, and $A = n^{-1}\sum_i H_i$. The optimal global growth ratio is

$$C = \frac{D_1}{n} = 1 + \lambda_{\max}\left(A^{-1}\left[\frac{1}{n}\sum_i (H_i - A)^2\right]A^{-1}\right). \tag{90}$$

Indeed, $n^{-1}\sum_i \|H_i x\|^2 = \|Ax\|^2 + x^\top[n^{-1}\sum_i (H_i - A)^2]x$, and the substitution $g = Ax$ gives (90). Let U denote the periodic stationary field at $\epsilon = 0$. A ray is *regular and outward* if $U(e) = -ae$ for some $e \ne 0$, $a > 0$, and every coordinate of every $H_i e$ is nonzero. On a section $\ell(y) = 1$ with $\ell(e) = 1$, its angular field is

$$F(y) = -U(y) + y\ell(U(y)).$$

It is attracting when $DF(e)|_{\ker \ell}$ is Hurwitz, as in Theorem 17.

F.1 THE NECESSARY HETEROGENEITY COST

Write $g = Ae$ and $P = \{j : e_j g_j > 0\}$. For $j \in P$ the scalar stationary direction has sign opposite to its nonzero mean g_j. Uniform random phase makes this a stationary finite stream; regularity ensures positive second moments and log-integrability. Theorem 1 therefore implies

$$C_j := \frac{n^{-1}\sum_i (H_i e)_j^2}{g_j^2} > 2.$$

Consequently, with $E_i = H_i - A$,

$$C - 1 \geq \frac{n^{-1}\sum_i \|E_i e\|^2}{\|g\|^2} > \frac{\|g_P\|^2}{\|g\|^2} \geq \frac{(e^\top A e)^2}{\|e\|^2 \|Ae\|^2}. \tag{91}$$

The last step follows from $0 < e^\top g \leq \sum_{j\in P} e_j g_j \leq \|e\|\|g_P\|$. For a positive-definite matrix of condition number at most κ,

$$\frac{(e^\top A e)^2}{\|e\|^2\|Ae\|^2} \geq \theta_\kappa := \frac{4\kappa}{(\kappa+1)^2}. \tag{92}$$

For completeness, rescale the spectrum to $[1, \kappa]$ and take $\|e\| = 1$. The scalar inequality $\lambda^2 \leq (\kappa+1)\lambda - \kappa$ gives $\|Ae\|^2 \leq (\kappa+1)t - \kappa$, where $t = e^\top A e$. The minimum of $t^2/[(\kappa+1)t - \kappa]$ on $[1, \kappa]$ is θ_κ. Equations (91)–(92) prove the strict lower bound $C > 1 + \theta_\kappa$.

Other sampling protocols and zero inputs. The lower bound holds for every outward ray $U(e) = -ae$, $a > 0$, including zero input coordinates. Under IID minibatches of size s it strengthens to $C > 1 + s\theta_\kappa$; under fresh reshuffling take $s = 1$. Indeed, for $j \in P$, the stationary scalar stream has nonzero mean g_j and variance σ_j^2/s, where $\sigma_j^2 = n^{-1}\sum_i (E_i e)_j^2$. For $b > 0$, the last nonzero input has bounded age under cyclic order or reshuffling, and geometric age tails under IID sampling. Its nonzero values belong to a finite set, so $v_j > 0$ and $\mathbb{E}|\log v_j| < \infty$. Theorem 1 forces $\sigma_j^2 > s g_j^2$. At $q = b = 0$, normalize the mean to one; Cauchy–Schwarz gives $\Pr(d > 0) \geq 1/\mathbb{E}d^2$, hence $\mathbb{E}\,\mathrm{sgn}(d) \geq 2/\mathbb{E}d^2 - 1$, with the same conclusion. Summing over P in (91) proves the claim.

Writing $C = 1 + \delta$ with $0 < \delta < 1$ gives the equivalent necessary condition

$$\kappa > \frac{2 - \delta + 2\sqrt{1-\delta}}{\delta} = \frac{(1+\sqrt{1-\delta})^2}{\delta} = \frac{4}{\delta} - 2 + O(\delta).$$

Here $\delta(\kappa+1)^2 > 4\kappa$ and $\kappa \geq 1$ select the larger root.

Near minimal heterogeneity, outward rays also require many coordinates.

Proposition 7 (Dimension cost of outward rays)**.** *Under any of the three protocols, let $U(e) = -ae$ with $a > 0$ be an outward ray, write $C = D_1/n = 1 + \delta$, and let $P = \{j : e_j (Ae)_j > 0\}$. For every $0 < t < 1$,*

$$|P| > \Big(\frac{1-t}{1+t}\Big)^2 \frac{1 - \delta/t}{\delta}.$$

In particular, $p > |P| > (1 - 3\delta)/(4\delta)$ whenever $\delta < 1/3$, and $|P| \geq (1 - O(\sqrt{\delta}))/\delta$ as $\delta \downarrow 0$. Under IID minibatches of size s, δ is replaced by δ/s.

Proof. Write $g = Ae$ and $\sigma_j^2 = n^{-1}\sum_i (E_i e)_j^2$, so that $\sum_j \sigma_j^2 \leq \delta\|g\|^2$ by (91). Since $e^\top A e > 0$, P is nonempty; the bound is immediate for $t \leq \delta$. Hence assume $t > \delta$. Each coordinate stream satisfies (13), so $|U_j| \leq (\mathbb{E}[m_j^2/v_j])^{1/2} \leq 1$ and hence $|e_j| \leq 1/a$; at $q = b = 0$ this is immediate. For $g_j \neq 0$, Theorem 1 and the preceding paragraph give $J_j := \mathrm{sgn}(g_j) U_j \geq 2/C_j - 1$ with $C_j = 1 + \sigma_j^2/g_j^2$. On P, $J_j = -a|e_j| < 0$, so $\sigma_j^2 > g_j^2$; if $e_j = 0 \neq g_j$, then $J_j = 0$ and $\sigma_j^2 \geq g_j^2$. On $N = \{j : e_j g_j < 0\}$, $J_j = a|e_j|$. Put $u_t = (1-t)/(1+t)$ and $N_t = \{j \in N : C_j \leq 1 + t\}$. Then $|e_j| \geq u_t/a$ on N_t, while every other coordinate with $g_j \neq 0$ has $\sigma_j^2 \geq t g_j^2$. Consequently $\|g_{N_t}\|^2 \geq (1 - \delta/t)\|g\|^2$ and $\|g_P\|^2 < \delta\|g\|^2$. Since $e^\top g = e^\top A e > 0$,

$$\frac{u_t}{a}\sqrt{1 - \delta/t}\,\|g\| \leq \frac{u_t}{a}\sum_{j\in N_t}|g_j| \leq \sum_{j\in N}|e_j g_j| < \sum_{j\in P} e_j g_j \leq \frac{\sqrt{|P|}}{a}\|g_P\| < \frac{\sqrt{|P|\delta}}{a}\|g\|.$$

Rearranging proves the bound; $N_t \neq \varnothing$ gives $p > |P|$. Take $t = 1/3$ and $t = \sqrt{\delta}$ for the special cases. For minibatches the stream variance is σ_j^2/s. □

Together with the condition-number bound above, both the dimension and the condition number must grow like $1/(C-1)$. The construction of Theorem 4 attains both orders: there $C-1=\Theta(1/M)$, $p=M+1$, and the mean block (74) has condition number $\Theta(M)$.

The stronger obstruction at a fixed period. For $n\ge 2$ and $b>0$, the same argument and (23) give

$$C > 1 + \big(C_{\mathrm{fin}}^{(q)}(n,b) - 1\big)\theta_\kappa \ge 1 + \big(C_{\mathrm{fin}}(n,b) - 1\big)\theta_\kappa. \tag{93}$$

Indeed, $\operatorname{sgn}(g_j)U_j(e) < 0$ for every $j \in P$. The nonnegative endpoint margin in Appendix A.2 forces $C_j > C_{\mathrm{fin}}^{(q)}$. Summing $(C_j-1)g_j^2$ over P in (91) proves (93). At $n=1000$, $b=.999$ and $\kappa=9$, even the q-uniform lower bound is $C > 18.33461$. This is a necessary condition for an outward ray. As the period grows at fixed memories, it tends to $C > 1+\theta_\kappa$. At $(q,b)=(0.9,0.999)$, (93) forces $n \ge 1703$, 6340 and 4749 for the (C,κ) of the twins, the identity lift and the $\kappa=9$ instance in Table 2. Their certified periods are within factors 11.7, 12.6 and 15.8 of these necessary lower bounds.

F.2 A MATCHING POSITIVE-DEFINITE CONSTRUCTION

Fix $\kappa>1$, $0\le q\le b<1$, and a target $\bar{C} > 1+\theta_\kappa$. Choose an odd integer $N\ge 3$ and a positive τ. A scalar profile w of length N contains $(N-1)/2$ entries equal to B and $(N+1)/2$ entries equal to $-\tau$, where

$$P_N = \frac{N-1}{2N}, \qquad B = \frac{1+(1-P_N)\tau}{P_N}.$$

Its mean is one and its centered variance is

$$\sigma^2 = \frac{1}{N}\sum_i (w_i-1)^2 = \frac{N+1}{N-1}(1+\tau)^2. \tag{94}$$

Repeat each entry consecutively L times. The resulting period has length $n=NL$ and stationary drift $J_L \to -1/N$ as $L\to\infty$. This limit, including its first-derivative version used below, is proved in Lemma 9. For large enough L put $c_L = -J_L > 0$. At $q=b=0$, $c_L = 1/N$ for every L.

Let M be a multiple of N, put $p=M+1$, and let $u = M^{-1/2}\mathbf{1}_M$ and e_h denote the normalized tail and the head coordinate. On $W=\{u,e_h\}^\perp$ take $A=\kappa I$; on $\operatorname{span}\{u,e_h\}$ take

$$A_2 = \begin{pmatrix} \alpha & h \\ h & d \end{pmatrix}, \qquad \alpha = \frac{2\kappa}{\kappa+1}, \quad h = \frac{(\kappa-1)\sqrt{\kappa}}{\kappa+1}, \quad d = \frac{\kappa^2+1}{\kappa+1}. \tag{95}$$

The eigenvalues of A_2 are $1,\kappa$, so $\operatorname{cond}_2(A)=\kappa$. This orientation attains the angle bound (92) along the tail direction u: its ratio is $\alpha^2/(\alpha^2+h^2)=\theta_\kappa$. Define

$$e = c_L\mathbf{1}_M - e_h = su - e_h, \qquad s = c_L\sqrt{M}, \qquad g = Ae = g_u u + g_h e_h,$$

where $g_u = \alpha s - h$ and $g_h = hs - d$. We will choose $s > d/h$, which makes both coefficients positive. For prescribed deviations $z_i \perp e$, the symmetric perturbation $E_i = z_i v^\top + v z_i^\top$ satisfies $E_i e = z_i$ whenever $v^\top e = 1$. We choose v to minimize $\|A^{-1}v\|$ under this constraint. Since $(A^{-1}v)^\top g = v^\top e = 1$, Cauchy–Schwarz gives $\|A^{-1}v\| \ge 1/\|g\|$, with equality uniquely when $A^{-1}v = g/\|g\|^2$. Thus

$$v = \frac{Ag}{\|g\|^2}, \qquad v^\top e = 1, \qquad A^{-1}v = \frac{g}{\|g\|^2}. \tag{96}$$

This choice also aligns the growth operator's head–tail block with g. Partition the M tail coordinates into N equally sized classes, indexed by $r=0,\dots,N-1$. For $i=0,\dots,N-1$, set

$$(z_i)_j = \frac{g_u}{\sqrt{M}}(w_{i+r}-1) \quad \text{in class } r, \qquad (z_i)_h = 0, \qquad H_i = A + z_i v^\top + v z_i^\top,$$

with profile indices modulo N. The actual cyclic sample order is L copies of H_0, then L copies of H_1, and so on. Thus averaging over either N distinct matrices or NL samples gives A.

Each z_i belongs to W, $\sum_i z_i = 0$, and $\|z_i\|^2 = S := g_u^2\sigma^2$. Writing $\mathcal{B} = N^{-1}\sum_i z_i z_i^\top$, orthogonality gives

$$\frac{1}{N}\sum_i (H_i - A)^2 = \|v\|^2\mathcal{B} + Svv^\top,$$

$$A^{-1}\left[\frac{1}{N}\sum_i (H_i - A)^2\right]A^{-1} = \frac{\|v\|^2}{\kappa^2}\mathcal{B} + S\frac{gg^\top}{\|g\|^4}.$$

The two terms act on orthogonal subspaces. The second has eigenvalue $S/\|g\|^2$, and the first has norm at most $\|v\|^2 S/\kappa^2 \le S/\|g\|^2$. Hence the *optimal global* ratio is

$$C = 1 + \delta, \qquad \delta = \sigma^2\frac{g_u^2}{g_u^2 + g_h^2}. \tag{97}$$

The two nonzero eigenvalues of $A^{-1/2}(H_i - A)A^{-1/2}$ are $\pm r$, where

$$r^2 = (z_i^\top A^{-1} z_i)(v^\top A^{-1} v) = \delta\frac{g^\top A g}{\kappa\|g\|^2} \le \delta.$$

Thus $\delta < 1$ guarantees $(1 - \sqrt{\delta})A \preceq H_i \preceq (1 + \sqrt{\delta})A$ and $H_i \succ 0$. Finally $H_i e = g + z_i$: each tail coordinate sees a cyclic shift of w times the positive factor $g_u/\sqrt{M}$, and the head sees the constant $g_h > 0$. Therefore $U(e) = -e$ exactly, with all phase-gradient coordinates nonzero.

F.3 Sharp limit, attraction, and trajectories

Lemma 9 (First-derivative block limit)**.** *Fix finitely many linear scalar gradients and a compact neighborhood on which their magnitudes are bounded away from zero. For fixed $0 \le q, b < 1$, the stationary mean direction for consecutive blocks of length L converges in C^1 to the mean of the block signs, at rate $O(L^{-1})$. The bound is uniform over a compact family of such gradients.*

Proof. In a block with gradient $a(y)$, $\nu \le |a(y)| \le B_0$, the exact states are

$$m_j = a + q^j(m_0 - a), \qquad v_j = a^2 + b^j(v_0 - a^2), \qquad 1 \le j \le L.$$

The periodic block-start states and their first derivatives are uniformly bounded, since the full-period fixed-point denominators $1 - q^{NL}$ and $1 - b^{NL}$ are bounded away from zero. Stationary $v \ge \nu^2$ at every phase. Differentiation of the displayed formulas therefore gives

$$\left|m_j/\sqrt{v_j} - \operatorname{sgn} a\right| + \left\|D_y(m_j/\sqrt{v_j})\right\| \le K(q^j + b^j).$$

Sum over the blocks and divide by NL. At zero memory the corresponding error is identically zero. □

The parameters separate three tasks: (N, τ) set the scalar variance, M sets the mean-gradient angle, and L makes the ray attracting. First choose N large and τ small so that $\sigma^2\theta_\kappa < \min(\bar{C} - 1, 1)$. Next use

$$\frac{(\alpha s - h)^2}{(\alpha s - h)^2 + (hs - d)^2} \longrightarrow \frac{\alpha^2}{\alpha^2 + h^2} = \theta_\kappa \quad (s \to \infty)$$

to choose a finite multiple M of N for which, at $s_\infty = \sqrt{M}/N$, $s_\infty > d/h$ and the strict target bound on δ holds. Keep this M fixed while increasing L. Since $c_L \to 1/N$, the same bounds hold for all sufficiently large L.

The resulting matrices and rays converge to a finite nondegenerate limit. Lemma 9 applies on a common neighborhood and yields $DU_L(e_L) \to 0$ in the full ambient operator norm. Its constants may depend on the already fixed M. For $\ell(x) = -x_h$, put $P_T = I - e_L\ell$. On the entire $(p-1)$-dimensional section the transverse derivative is

$$DF_L(e_L)|_{\ker\ell} = -I - P_T DU_L(e_L)|_{\ker\ell}.$$

For large enough L the perturbation has norm below $1/2$, so the ray is attracting. This completes the construction with $1 + \theta_\kappa < C < \min(\bar{C}, 2)$. The order of choices is $(\kappa, q, b, \bar{C})$, then (N, τ), then M, then L; it establishes sharpness for each fixed memory pair.

For $b > 0$, Theorem 17 now applies directly to this finite instance. It includes zero initial moments, finite fixed ϵ, either bias-correction choice, and all nonsummable schedules in Setup. At $q = b = 0$, U is locally the constant $-e$ and the transfer is direct. Choose a small compact section around e on which every phase-gradient coordinate stays away from zero. Bounded coordinate updates give within-epoch motion at most $\sqrt{p}\,\eta_k$. At an epoch start $x_k = r_k y_k$, $r_k = -x_{k,h}$, the actual epoch map is

$$x_{k+1} = x_k + \eta_k(e + \xi_k), \qquad \|\xi_k\| \le K(\eta_k + \epsilon)/r_k.$$

Here the sign pattern is constant on the section, and the difference between $a/(|a| + \epsilon)$ and $\operatorname{sgn} a$ is bounded by $\epsilon/|a|$. These estimates also hold for one prospective step in a slightly larger section, by taking r_0 large enough. With $z_k = y_k - e$ and $h_k = \eta_k/r_k$, this gives

$$r_{k+1} - r_k = \eta_k[1 + O((\eta_k + \epsilon)/r_k)], \qquad z_{k+1} = z_k + h_k[-z_k + O((\eta_k + \epsilon)/r_k + h_k)].$$

Two nested small balls in the section close a first-exit argument: the squared norm decreases on the annulus when the error is sufficiently small, and a step from the inner ball cannot cross the outer boundary. At the same time $r_{k+1} - r_k \ge \eta_k/2$. Starting at $r_0 e$, the nonincreasing schedule and increasing radius preserve both error tolerances. Hence $r_k \to \infty$ and $f(x_k) \to \infty$. Bias correction is the identity at this corner.

F.4 A FINITE DEFAULT-MEMORY INSTANCE AT CONDITION NUMBER NINE

Use the rank-two construction of Appendix F.2 with $(q, b) = (.9, .999)$, $N = 5$, $L = 15000$, $M = 4900$, and

$$w = (103/40, 103/40, -1/20, -1/20, -1/20), \qquad A_2 = \begin{pmatrix} 13/5 & 16/5 \\ 16/5 & 37/5 \end{pmatrix}.$$

This finite orientation has eigenvalues $1, 9$; take $A = 9I$ on W. The profile has mean one and centered variance $\sigma^2 = 1323/800$. With $c_L = -J_L$ defined by its exact periodic drift, the same formulas for e, g, v, z_i, H_i give $n = 75000$, $p = 4901$, and

$$\operatorname{cond}_2(A) = 9, \qquad 1.8348135501173188 < C < 1.8348135501173190 < 2.$$

Here C is the exact global expression (97); the squared relative component perturbation is below 0.758457, so every component is positive definite. The directed-interval computations use 60 and 100 decimal digits.

To verify the full transverse space, set $t = g_u/70$ and $d_j = \partial J_L/\partial w_j$, with each derivative perturbing all L copies in block j. At the constant positive head stream, the derivative of the mean normalized direction is $\operatorname{mean}(P_q h - P_b h)/g_h = 0$, since both EMA filters preserve the period mean. The 4895 within-class zero-sum directions have $H_i y = 9y$ and hence eigenvalue $-1 - 9(\sum_j d_j)/t < -24.45796$. The remaining five modes use the circulant class Jacobian in Appendix E.4, with $N = 5$, $\kappa = 9$ and $\alpha = 13/5$. The interval certificate places every Gershgorin right endpoint below -0.18572. Thus all 4900 transverse modes are strictly stable. An independent 70-digit reverse-adjoint computation confirms the block derivatives and every transverse eigenvalue real part. The ray is regular and satisfies $U(e) = -e$, so Theorem 17 yields divergence with zero initial moments, sufficiently large ray initialization, finite fixed ϵ, and either bias-correction choice.

F.5 NON-RAY DIVERGENCE AT FIXED CONDITIONING

The ray lower bound does not constrain trajectories whose direction keeps changing. We construct such trajectories at fixed conditioning, proving Theorem 6 and its consequence for the region of Theorem 21. Put

$$\lambda_\pm = 1 \pm 1/\sqrt{2}, \qquad \kappa_\star = \lambda_+/\lambda_- = 3 + 2\sqrt{2}.$$

At $q = b = 0$ both bias conventions give $\hat{m}_t \oslash (\sqrt{\hat{v}_t} + \epsilon\mathbf{1}) = \phi_\epsilon(G_t)$ coordinatewise, with $\phi_\epsilon(u) = u/(|u| + \epsilon)$ and $\phi_0 = \operatorname{sgn}$. At $\epsilon = 0$ each sample moves every coordinate by $-\eta_k/n$ times the sign of its current component-gradient coordinate. The frozen terminal direction restores the coordinate toward its input median; Step 5 controls the actual cyclic iterates around this moving target. The construction designs these medians. At $\epsilon > 0$, inputs of size comparable to the radius make ϕ_ϵ agree with sgn up to controlled errors.

Corollary 5 (Sharpness of $3+2\sqrt{2}$ without memory)**.** *Let $q=b=0$ and fix $\epsilon \geq 0$. For $\kappa \geq 1$ let $C_{\mathrm{safe}}(\kappa)$ be the supremum of all $C_0 \geq 1$ such that cyclic Adam, with every diminishing schedule and from every initialization, converges on every finite sum, in every dimension, that satisfies the hypotheses of Theorem 21 with $L/m=\kappa$ and $C<C_0$. Then*

$$C_{\mathrm{safe}}(\kappa) \geq \frac{8\kappa}{(\kappa+1)^2} > 1 \quad (\kappa < \kappa_\star), \qquad C_{\mathrm{safe}}(\kappa) = 1 \quad (\kappa > \kappa_\star).$$

Proof. For $\kappa < \kappa_\star$, Theorem 21 applies when $\epsilon > 0$, and (112) with $R=1$ gives $T_0 = 8\kappa/(\kappa+1)^2 > 1$ when $\epsilon = 0$. For $\kappa > \kappa_\star$, the instances of Theorem 6 have positive-definite quadratic components, so they satisfy these hypotheses with m, L the extreme eigenvalues of A, and they diverge with C arbitrarily close to one. □

Reductions and order of choices. Quadratic problems are jointly homogeneous: the trajectory from $R\bar{x}$ with scales η_k and constant ϵ is R times the trajectory from $\bar{x}$ with η_k/R and ϵ/R. It therefore suffices to treat $\sup_k \eta_k \leq 1$, $\epsilon \leq 1$ and $R \geq R_0$ with one R_0; rescaling by $\max(\sup_k \eta_k, \epsilon)$ gives the condition $R \geq R_0(\sup_k \eta_k + \epsilon)$. Steps 1–3 build a finite relay sign system and a finite median law that realizes it; Step 4 embeds this law in fixed positive-definite components and computes C; Steps 5–6 show that the actual cyclic iterates follow the relay system up to a bounded input error, and that this error cannot stop the escape. Every choice in Steps 1–4 selects a finite object, and all of them are made before R. The only infinite limit is the time limit for one fixed instance.

Step 1: a chain with small wrong-sign cost. For a curve $x(t) \in \mathbb{R}^d$ put $\|x\|_\star^2 = \sum_j \sup_t |x_j(t)|^2$, with the supremum taken before the sum. Consider continuous, periodic, antipodal piecewise-affine curves, $x(t+H) = -x(t)$, whose coordinates have velocities $x_j' \in \{-1, 1\}$, and a symmetric $G \succ 0$. Sign descent $x' = -\operatorname{sgn}(\text{input})$ realizes such a curve from an input that equals Gx up to a correction of size β_j in coordinate j, where

$$\beta_j = \sup_t \big[x_j'(t)\,(Gx(t))_j\big]_+$$

is the *wrong-sign cost* of coordinate j. For every $N \geq 2$ there are such a curve X and a matrix $G = A_N$ with

$$\lambda_- I \preceq A_N \preceq \lambda_+ I, \qquad \lambda_{\min}(A_N) = \lambda_-, \quad \lambda_{\max}(A_N) = \lambda_+, \qquad \frac{\sum_j \beta_j^2}{\min_t \|X(t)\|^2} \leq \frac{24}{N+1}. \tag{98}$$

The construction rests on a replication lemma. Let b be any continuous, periodic, antipodal, piecewise-affine curve, of arbitrary speed, and fix tolerances. Track each coordinate of b by k^2 copies of a fast antipodal zigzag z with slopes $\pm k$, uniform error e and $O(k)$ turns, divided by k and shifted in phase across a short interval $[-h, h]$. This gives a normalized replication P ($P^\top P = I$), a unit-speed curve x and a continuous antipodal target $\tilde{c}$ with the period of b, such that $x_j'(x_j - \tilde{c}_j) \leq 0$, $\|\tilde{c}\|_\star \leq \|x\|_\star$, and $\|x - Pb\|_\star$, $\|PP^\top \tilde{c}\|_\star$ are as small as desired. Indeed the squared tracking errors of the copies sum to at most $(e + \mathrm{Lip}(b)h)^2$, and uniform quadrature over the shifts gives $P^\top(|x|x') = O(1/k)$, because $|z|z'$ has total variation $O(k^2)$. The target $\tilde{c}$ is $|x|x'$, linearly interpolated to x near the finitely many turns, which are placed at distinct times.

Start from a scalar unit-speed triangle wave x_0 and its target $\tilde{c}_0$, formed from $|x_0|x_0'$ in the same way. For $\ell = 1, \dots, N$ apply the lemma to $b = \tilde{c}_{\ell-1}$, which is piecewise affine but in general not of unit speed; this gives x_ℓ, $\tilde{c}_\ell$ and the replication $P_{\ell-1}$ into layer ℓ, with tolerances at most $1/(100N)$. Put $E_0 = I$, $E_\ell = P_{\ell-1}P_{\ell-1}^\top$, $Q_\ell = I - E_\ell$ and

$$(A_N)_{\ell\ell} = \tfrac{3}{2}I - E_\ell, \qquad (A_N)_{\ell,\ell+1} = -\tfrac{1}{2}Q_\ell P_\ell^\top.$$

Distinct off-diagonal blocks pair mutually orthogonal subspaces, and each coupled pair carries $\left(\begin{smallmatrix} 3/2 & -1/2 \\ -1/2 & 1/2 \end{smallmatrix}\right)$, whose eigenvalues are $\lambda_\pm$; uncoupled parts have eigenvalues $1/2$ or $3/2$. This proves the spectral part of (98) for every depth and replication. In each interior layer, $(A_N x)_\ell = \frac{1}{2}(x_\ell - \tilde{c}_\ell) + r_\ell$ with $\|r_\ell\|_\star \leq 3/(100N)$, and the leading term has the sign that the velocity requires. Only the root and the last layer have wrong-sign costs of order one, so $\sum_j \beta_j^2 \leq 2$, whereas each of the $N+1$ layers keeps time-averaged squared norm at least $1/6$. Many equally spaced phase copies of the whole curve, under a block-diagonal copy of A_N, turn this time average into a pointwise lower bound on $\|X(t)\|^2$, which proves (98).

Step 2: an exact relay system. Let X have half-period H. For $\Lambda > 1$ put $a_\Lambda = (1+\Lambda)/2$, $t_0 = a_\Lambda H/(\Lambda - 1)$ and

$$\chi(t) = a_\Lambda X\big((t - t_0)/a_\Lambda\big) + (1 - a_\Lambda)X(0), \qquad t_0 \le t \le \Lambda t_0,$$

extended by $\chi(\Lambda t) = -\Lambda\chi(t)$. This path is continuous, has coordinate velocities ± 1, and as $\Lambda \downarrow 1$ its normalized wrong-sign ratio tends to that of X. For $\tau > 0$ and $c_\Lambda = (\Lambda - 1)/(\Lambda + 1)$ define

$$T_\tau(t) = (-1)^k\big[t - (1 + c_\Lambda)\Lambda^k\tau\big], \qquad \Lambda^k\tau \le t < \Lambda^{k+1}\tau.$$

It is continuous with slopes ± 1, has kinks exactly at the points $\Lambda^k\tau$, and satisfies $T_\tau(\Lambda t) = -\Lambda T_\tau(t)$. With $G_\Lambda = \sqrt{\Lambda}$ and $\omega = (G_\Lambda - 1)/(G_\Lambda + 1)$, the clock $Z_\tau(t) = (T_\tau(t), G_\Lambda T_\tau(t/G_\Lambda))$ solves

$$Z'_\tau = -\operatorname{sgn}(M_\omega Z_\tau), \qquad M_\omega = \begin{pmatrix} -\omega & 1 \\ -1 & -\omega \end{pmatrix}$$

exactly. Each quarter turn multiplies its radius by G_Λ, and at each switching wall both adjacent normal speeds have the same strict sign, with magnitudes $1 \mp \omega$, so there is no sliding.

Every continuous piecewise-affine u with $u(\Lambda t) = -\Lambda u(t)$ and finitely many knots per half-scale satisfies

$$u(t) = \sum_\tau \frac{\Delta u'(\tau)}{2}\, T_\tau(t),$$

summed over its knots τ in one half-scale $[t_1, \Lambda t_1)$: both sides have the same derivative jumps, so they differ by an affine function, and the scaling rule forces it to vanish. On each constant-velocity interval of χ_j place a tent of sign $-\chi'_j$ with slopes ± 1 that vanishes at the endpoints, and extend it anti-self-similarly. The formula realizes all tents by one finite clock bank Z and one linear readout F of first clock coordinates, so

$$Z' = -\operatorname{sgn}(M_c Z), \qquad \chi' = -\operatorname{sgn}(FZ)$$

is an exact, autonomous, triangular and finite sign system. The readout zeros are simple, and its active kinks lie at tent midpoints, so each zero has a kink-free neighborhood. Take R_c copies of χ sharing this clock bank, put $Y = (\chi, \dots, \chi, Z)$ and $A_\circ = \operatorname{diag}(A_N, \dots, A_N, I)$, where A_N now denotes the block-diagonal copy from Step 1. Let $\mathcal{T}$ be the triangular matrix of this system, multiplied by a small constant $\lambda_0 > 0$, so that $Y' = -\operatorname{sgn}(\mathcal{T}Y)$ and $\|\mathcal{T}\| \le 1$. Copying the source leaves its normalized wrong-sign ratio unchanged and divides the clock's share of $\|Y\|^2$ by R_c.

Step 3: a finite median law with small covariance. Fix a coupling $\nu > 0$ and a rational $\rho \in (0, 1/2)$, and put

$$L = \frac{A_\circ - \mathcal{T}}{\nu(1-\rho)}, \qquad C_\circ = \frac{A_\circ - \rho\mathcal{T}}{1-\rho}, \qquad z_\rho = \Phi^{-1}\Big(\frac{1}{2(1-\rho)}\Big).$$

For output j give the coefficient vector ℓ_j the law $(1-\rho)N(0, t_j^2 I) + \rho\,\delta_{L_j}$, centered by subtracting ρL_j. With Φ the standard normal distribution function and $\operatorname{clip}(a, l, u) = \min(\max(a, l), u)$, the median of $\ell_j^\top y$ is $\mu_j(y) = \operatorname{clip}(L_j y, -z_\rho t_j\|y\|, z_\rho t_j\|y\|) - \rho L_j y$, and with $\gamma_j = \nu z_\rho t_j$,

$$A_\circ y - \nu\mu(y) = \operatorname{clip}\big(\mathcal{T}y,\ C_\circ y - \gamma\|y\|,\ C_\circ y + \gamma\|y\|\big) \tag{99}$$

coordinatewise. The total covariance of the laws is

$$K = \frac{1-\rho}{\nu^2 z_\rho^2}\Big(\sum_j \gamma_j^2\Big) I + \frac{\rho}{\nu^2(1-\rho)}(A_\circ - \mathcal{T})^\top(A_\circ - \mathcal{T}). \tag{100}$$

Choose each clock width γ_j to cover the full envelope of $((C_\circ - \mathcal{T})Y)_j/\|Y\|$ along the path; then every clock input in (99) equals $\mathcal{T}Y$. Choose each source width strictly larger than the wrong-sign cost of its coordinate, including the correction $-\rho\mathcal{T}/(1-\rho)$ in $C_\circ$, divided by the minimum radius. Then (99) has the sign required by the velocity along the path. At a source turn the two adjacent velocities have opposite signs, so the wrong-sign cost bounds $|(C_\circ Y)_j|$ there, while $(\mathcal{T}Y)_j$ vanishes; by the strict slack the clip is inactive nearby, and the input equals $(\mathcal{T}Y)_j$ exactly. The choices are noncircular: fix ν; take ρ small for the second term of (100); then make $\sum_j \gamma_j^2$ small relative to $\nu^2 z_\rho^2$

by a large depth N, a dilation Λ close to one, the finite clock bank, many source copies R_c, strict width slack and a small λ_0.

The Gaussian parts are then replaced by a finite law. For every finite dimension and accuracy there is a finite, equally weighted, centrally symmetric cloud with mean zero, covariance exactly I, and one-dimensional projection distribution functions uniformly close to the standard Gaussian: truncate and round a rotation-invariant Gaussian to a symmetric grid, rationalize the weights over signed-permutation orbits, and rescale. Using such a cloud for every Gaussian part, keeping the rational atom weight, centering exactly, clearing denominators and adding one zero centered entry gives, for each output j, an equally weighted centered profile $(\ell_{jr})_{r\le m}$ of one common odd length m, with

$$K_{\mathrm{fin}} = \sum_j K_j, \qquad K_j = \frac{1}{m}\sum_r \ell_{jr}\ell_{jr}^\top, \qquad K_{\mathrm{fin}} = (1-1/m)K.$$

Where the atom is strictly central, the left and right rank gaps at the atom stay uniformly positive, so the finite median $\hat{\mu}_j(y)$ equals the atom value exactly. Hence the clock inputs are exactly linear on a conic neighborhood of the whole path, each source input equals its linear readout near its turns, and the strict signs persist elsewhere.

Step 4: fixed positive-definite components and their growth constant. Let $d = \dim Y$, replicate each profile to a multiple M_p of m, and let P be the normalized replication of $\mathbb{R}^d$ into d groups of M_p terminal coordinates. For every tuple $\pi = (\pi_1, \dots, \pi_d)$ of permutations of the group indices, let V_π have rows $\ell_{j,\pi_j(r)}^\top/\sqrt{M_p}$, and set

$$\mathcal{H}_\pi = \begin{pmatrix} A_\circ + \nu^2 I & \nu P^\top + V_\pi^\top \\ \nu P + V_\pi & I \end{pmatrix}, \qquad \mathcal{A} = \begin{pmatrix} A_\circ + \nu^2 I & \nu P^\top \\ \nu P & I \end{pmatrix}.$$

The finite sum uses every tuple once, in a fixed cyclic order, with repetitions allowed; its mean is $\mathcal{A}$. Centering gives, for every π,

$$P^\top V_\pi = 0, \qquad V_\pi^\top V_\pi = K_{\mathrm{fin}}, \qquad \mathrm{avg}_\pi\, V_\pi = 0.$$

The Schur complement of the terminal block of $\mathcal{H}_\pi$ is $A_\circ - K_{\mathrm{fin}}$, so every component is positive definite once $\|K_{\mathrm{fin}}\| < \lambda_-$, and $\|\mathcal{H}_\pi - \mathcal{A}\|_2 = \|K_{\mathrm{fin}}\|^{1/2}$. The mean has eigenvalue 1 on the complement of range P in the terminal block, and on an eigenvector of $A_\circ$ with eigenvalue λ it has the eigenvalues of $\left(\begin{smallmatrix} \lambda+\nu^2 & \nu \\ \nu & 1 \end{smallmatrix}\right)$. Its condition number depends only on ν and tends to $\kappa_\star$ as $\nu \downarrow 0$. In (90), $\mathrm{avg}_\pi(\mathcal{H}_\pi - \mathcal{A})^2$ is block diagonal with blocks K_{fin} and $\bigoplus_j \mathrm{tr}(K_j)(I - \mathbf{1}\mathbf{1}^\top/M_p)/(M_p-1)$, and $\min_z \|\mathcal{A}(y,z)\|^2 = y^\top A_\circ^2 y/(1+\nu^2)$ on the source and consensus coordinates. Therefore

$$C - 1 = \max\Big\{(1+\nu^2)\,\lambda_{\max}\big(A_\circ^{-1} K_{\mathrm{fin}} A_\circ^{-1}\big),\ \max_j \frac{\mathrm{tr}\, K_j}{M_p - 1}\Big\}. \tag{101}$$

Given $\kappa > \kappa_\star$ and $\delta > 0$, fix ν with $\mathrm{cond}_2(\mathcal{A}) < \kappa$, make both terms of (101) smaller than δ by Step 3 and a large M_p, and append to every component the same scalar $\kappa\lambda_{\min}(\mathcal{A})$ on one extra coordinate. The mean then has condition number exactly κ; the extra coordinate adds no covariance, and it stays at its initial value zero. Since the widths are positive, $K_{\mathrm{fin}} \neq 0$ and $C > 1$.

Step 5: the actual iterates follow the median port. Write the terminal state as $Pz + w$ with $P^\top w = 0$. Since $P^\top V_\pi = 0$, every component gives the source gradient $(A_\circ + \nu^2 I)y + \nu z + V_\pi^\top w$; on $w = 0$ the phase-dependent rows cancel before any averaging. Actual samples need not keep $w = 0$. However, if π' exchanges the profile entries that π assigns to terminal rows r and v of one group, then

$$e_r^\top \mathcal{H}_\pi - e_v^\top \mathcal{H}_{\pi'} = (e_r - e_v)^\top.$$

Within epoch k every coordinate moves by at most η_k, so gradient coordinates change by at most $H_\infty \eta_k$, with $H_\infty = \max_\pi \|\mathcal{H}_\pi\|_\infty$. If a within-group gap $u_r - u_v$ exceeds $\max(2, 2H_\infty)\eta_k$ at the start of an epoch, the identity makes the input of (π, r) larger than that of (π', v) at their actual evaluation points. Monotonicity of ϕ_ϵ and the bijection $\pi \leftrightarrow \pi'$ show that the epoch does not increase the gap, and any epoch changes it by at most $2\eta_k$. Induction from $w_0 = 0$ gives $\|w_t\|_\infty \le W \sup_k \eta_k$ for all t, with W independent of time and of R.

Let $\hat{\mu}_j$ be the finite profile median and $b_j(y) = -\nu y_j - \hat{\mu}_j(y)$, at which the median input of group j vanishes. Suppose $z_j - b_j(y)$ exceeds $T = T_1 \sup_k \eta_k + 2m\sqrt{M_p}\,\epsilon$, with T_1 covering W and the within-epoch motion. Then strictly more than half of the entries of profile j give physical terminal inputs above $T/(2\sqrt{M_p}) \geq m\epsilon$ throughout the epoch, so each of them contributes $\phi_\epsilon \geq m/(m+1)$, the others at least -1, and the group average of ϕ_ϵ is at least $1/(2m)$. Hence z_j decreases at rate at least $\sqrt{M_p}/(2m)$, while $b_j(y)$ moves at rate at most $V_j = \nu + \max_r \|\ell_{jr}\|_1$. With $\sqrt{M_p}/(2m) > \max_j V_j$, a condition that does not involve ϵ, the same drift-and-overshoot induction bounds $e = z - b(y)$ by a fixed multiple of $\sup_k \eta_k + \epsilon$. Consequently, at every sample,

$$\text{source gradient} = A_\circ y - \nu\hat{\mu}(y) + e_t, \qquad \|e_t\|_\infty \leq E, \tag{102}$$

with E a fixed multiple of $\sup_k \eta_k + \epsilon$. This is a samplewise identity for the actual iterates, not a statement about averaged dynamics.

Step 6: bounded input errors do not stop the clocks. Where Step 3 applies, (102) and (99) make the actual source–clock samples the system $Y \leftarrow Y - (\eta_k/n)\phi_\epsilon(\mathcal{T}Y + \xi)$ with $\|\xi\|_\infty \leq E$ at clock rows and near source turns; on the strict-sign parts of Step 3 the velocity has the sign of $-\mathcal{T}Y$ and magnitude at least $1 - \epsilon/(c\|Y\| + \epsilon)$. Dividing inputs by λ_0 replaces ϵ by $\epsilon' = \epsilon/\lambda_0$. After a reflection, a clock has inputs $l_1 = \omega x + y$ and $l_2 = x - \omega y$. Linear interpolation of the samples and the input error create strips of the fixed width $D = E/\lambda_0 + \max(1+\omega, \|F\|_\infty) \sup_k \eta_k$ around the two switching walls. Inside a strip the affected velocity may take any value in $[-1, 1]$; outside, it has the ideal sign and magnitude at least $1 - \epsilon'/(|l_i| - D + \epsilon')$. When $|l_i| \geq r/2$ and $r \geq 4D$, its deficit is at most $4\epsilon'/r$. Thus, for $r \geq C_\omega(D + \epsilon')$, each wall has normal speed at least $(1-\omega)/2$, so the strip crossings take time $O(D)$, there is no sliding, and the soft-sign deficits integrate to $O(\epsilon' \log(e + r/\epsilon'))$ over a quarter. Comparing with the ideal quarter path, whose endpoint equations are linear in the accumulated deficits, a quarter turn that starts at radius r ends at radius r_+ after time T with

$$|r_+ - G_\Lambda r| + |T - \tau_\omega r| \leq C_\omega\big(D + \epsilon' \log(e + r/\epsilon')\big), \qquad \tau_\omega = \frac{1+\omega^2}{1-\omega};$$

at $\epsilon = 0$ one may take the separate bounds $4D/(1-\omega)^2$ and $2(1+\omega)D/(1-\omega)^2$ for $r \geq 8D/(1-\omega)$. Choose R so the quarter error is at most $(G_\Lambda - 1)r/2$ for $r \geq R$. Then r_{j+1}/r_j lies between $(G_\Lambda+1)/2 > 1$ and $(3G_\Lambda-1)/2$, so $\log r_j = O(\log R + j)$. The section radii obey $r_{j+1} = G_\Lambda r_j + \rho_j$ with $|\rho_j| \leq C(D_R + \epsilon' j)$ and $D_R = D + \epsilon' \log(e + R/\epsilon')$. Hence $r_j = c_\infty G_\Lambda^j + O(D_R + \epsilon' j)$, where $c_\infty = r_0 + \sum_{j \geq 0} \rho_j / G_\Lambda^{j+1}$ converges absolutely, and section times carry errors $O(jD_R + \epsilon' j^2)$. Thus each neutral clock phase converges to a nearby fixed phase, and no attraction between clocks is needed. With the age $A_t = 1 + t/R$, the clock bank stays within $O(D_R(1+\log A_t) + \epsilon'(1+\log A_t)^2)$ of an exact bank whose phase vector θ_∞ lies within $O(D_R/R)$ of the nominal one. The readout zeros are simple, so for nearby phases the set where $|FZ_\theta| \leq u$ has length $O(u)$ per half-scale. Source signs therefore disagree with those of the exact system for time $O(D_R(j+1) + \epsilon'(j+1)^2)$ on half-scale j, the soft-sign deficits of the source add $O(\epsilon'(\log(e + R/\epsilon') + j))$ there, and the source stays within $O(D_R(1+\log A_t)^2 + \epsilon'(1+\log A_t)^3)$ of $R\chi_{\theta_\infty}(A_t)$, the anti-self-similar source driven by the exact clocks. At $\epsilon = 0$ these are the bounds $O(D(1+\log A_t))$ and $O(D(1+\log A_t)^2)$. Write Y_{θ_∞} for this phase-adjusted reference, the exact source–clock path with clock phase θ_∞. Then $\|Y(t) - RY_{\theta_\infty}(A_t)\|_\infty$ obeys the same sublinear bounds. Since the normalized path $Y_\theta(A)/A$ is Lipschitz in the phase, uniformly in $A \geq 1$, it also satisfies $\|Y_{\theta_\infty}(A) - Y_{\mathrm{ref}}(A)\|_\infty / A = O(D_R/R)$.

These estimates assume the error model, which holds only near the reference path. Use the time-normalized tube $\|Y(t)/(R+t) - Y_{\mathrm{ref}}(A_t)/A_t\|_\infty < \zeta$, with ζ smaller than the conic margin of Step 3 and than half the minimum normalized clock norm. Before a first exit, the tube gives $\|Y(t)\| \geq c(R+t)$, so the strict source signs of Step 3 dominate E and the error model holds. Continue the sequence after the putative exit by the error model with $\xi = 0$. The bounds above, $\sup_{A \geq 1}(1+\log A)^3/A < \infty$ and $D_R/R \to 0$ keep $Y(t)/(R+t)$ within $\zeta/4$ of $Y_{\theta_\infty}(A_t)/A_t$, which lies within $\zeta/4$ of $Y_{\mathrm{ref}}(A_t)/A_t$, for all large R. This contradicts the exit. Hence the actual iterates stay in the tube for all time, and $\|x_t\| \geq \|Y(t)\| \geq c(R+t) \to \infty$, because the cumulative step sum is infinite. The initial point $R\bar{x} = (RY_{\mathrm{ref}}(1),\ Pb(RY_{\mathrm{ref}}(1)),\ 0)$ has zero within-group spread and lies on the median graph.

Finally, the phase-adjusted reference satisfies $Y_{\theta_\infty}(\Lambda A) = -\Lambda Y_{\theta_\infty}(A)$ with a nonzero clock part, $Y(t) - RY_{\theta_\infty}(A_t)$ is sublinear in $R + t$, and the terminal coordinates equal $Pb(y)$ up to bounded

errors with b odd and positively homogeneous. The normalized iterates therefore have two opposite subsequential limits and no limiting direction. The supplement (`nonray_fixed_condition/`) contains the complete notes and exact rational checks of the clock formulas, the spline identity, the port identities and the quarter estimates.

G GEOMETRY, SHARED NORMALIZATION, AND BOUNDED CONDITIONING

Proof of Corollary 1. Take a $C < 2$ family from Theorem 5 and a sufficiently distant divergent initialization x_0. Choose orthogonal Q with $Q^\top AQ = \Lambda$ diagonal, and put $\widetilde{f}_i(z) = f_i(Qz)$, $z_0 = Q^\top x_0$. Then $\nabla \widetilde{f}_i(z) = Q^\top \nabla f_i(Qz)$. Orthogonality preserves each Hessian spectrum and both squared norms in the growth inequality; the bijection $x = Qz$ therefore preserves its optimal D_1 exactly. The average $\widetilde{f}(z) = z^\top \Lambda z/2$ is separable and strongly convex. Theorem 7 gives $z_t \to 0$ for the same algorithmic choices and diminishing schedule, whereas the original run satisfies $\|x_t\| \to \infty$. Thus $\|Qz_t\| \to 0$ for the same finite sum and initial point. No componentwise separability is required.

G.1 DIMENSION-FREE CONVERGENCE WITH ONE SHARED SECOND MOMENT

The rotation-equivariant comparator uses a single second moment:

$$m_t = qm_{t-1} + (1-q)G_t, \quad s_t = bs_{t-1} + (1-b)\|G_t\|^2, \quad x_t = x_{t-1} - \gamma_t \frac{m_t}{\sqrt{s_t} + \epsilon}. \tag{103}$$

Here G_t is evaluated before the update. Bias correction, when used, acts on both moments in the usual way.

Theorem 20 (Sharp shared-scalar stationary barrier)**.** *Let a strictly stationary vector stream have mean $\mu \neq 0$ and finite second moment. Suppose $0 \le q \le b < 1$, the stationary $s_t > 0$ almost surely, and the log-integrability assumption of Theorem 9 holds. For $C = \mathbb{E}\|G\|^2/\|\mu\|^2$, at $\epsilon = 0$,*

$$J := \mathbb{E}\frac{\langle \mu/\|\mu\|, m\rangle}{\sqrt{s}} \ge 2/C - 1.$$

The bound is sharp by embedding the scalar sharpness family in a line.

Proof. Normalize $\|\mu\| = 1$ and write $h = \mu$. Let $r = P_b G$ and $u = P_q\|G\|^2$, where P_c is the geometric filter with parameter c. Jensen gives $\|r\|^2 \le s$ and $\|m\|^2 \le u$. The scalar difference-recursion argument (14) gives $\mathbb{E}(u/s) \le 1$, hence $\mathbb{E}(\|m\|^2/s) \le 1$. Also $a = \mathbb{E}\sqrt{s} \ge \|\mathbb{E}r\| = 1$. The variables $y = h^\top m/\sqrt{s}$ and $z = \sqrt{s}$ obey $\mathbb{E}yz = 1$, $\mathbb{E}y^2 \le 1$, $\mathbb{E}z^2 = C$, and $\mathbb{E}z \ge 1$. The centered scalar calculation proves the result. □

For finite ϵ, normalized $e = \epsilon/\|\mu\|$ and $J_e = \mathbb{E}[h^\top m/(\sqrt{s} + e)]$ obey $1 \le J_e(a+e) + \sqrt{(1-J_e^2)(C-a^2)}$, so

$$J_e \ge \frac{1-\sqrt{C-1}}{\sqrt{C}+e} > 0 \quad (C < 2). \tag{104}$$

For a frozen finite sum, in particular,

$$\left\langle \frac{g}{\|g\|}, U_0^{\mathrm{fr}} \right\rangle \ge 2n/D_1 - 1 \qquad (D_0 = 0). \tag{105}$$

Finite-period gain. For $n \ge 2$, $b > 0$, use $\alpha, \beta, \gamma_{n,b}$ and $H_q, V_{n,b}$ from Appendix A.2. With normalized mean of norm one, minimum filter weights give $s_i - \|r_i\|^2 \ge \alpha(C-1)$ and $u_i - \|m_i\|^2 \ge \beta(C-1)$. Jensen yields $a^2 \ge 1 + \alpha(C-1)$; concavity of the square root on the simplex of squared input norms gives $a^2 \ge \gamma_{n,b}C$. Furthermore, $\mathbb{E}(\|m\|^2/s) \le 1 - \beta + \beta/C$ by $u_i - \|m_i\|^2 \ge \beta(C-1)$, (14), and Jensen. Vector-valued Parseval also gives $\mathbb{E}\|m\|^2 \le 1 + s_q(C-1)$; together with $s_i \ge \alpha C$ this yields the second branch of $H_q(C)$. Thus $\mathbb{E}(\|m\|^2/s) \le H_q(C)$, and centering $y = h^\top m/(\sqrt{s} + e)$ gives the same dimension-free margin:

$$J_e \ge \frac{1-\sqrt{H_q(C)V_{n,b}(C)}}{\sqrt{C}+e} > 0 \quad \text{if } C < C_{\mathrm{fin}}^{(q)}(n,b). \tag{106}$$

Theorem 12 turns these stationary margins into cyclic convergence of the actual iterates, including the finite-period threshold.

G.2 An aggregate convergence region under bounded conditioning

For a nonseparable average, a cubic potential in the parameter coordinates can replace the cubic-gradient potential. Here Hessian bounds hold globally, rather than only at a single point.

Theorem 21 (Aggregate convergence under bounded conditioning)**.** *Suppose every component gradient is globally Lipschitz, f is C^2, and for constants $0 < m \le L < \infty$,*

$$mI \preceq \nabla^2 f(x) \preceq LI, \qquad S(x) = n^{-1} \sum_i \|g_i(x)\|^2 \le C\|g(x)\|^2 \quad \textit{for all } x.$$

Let $\kappa = L/m$ and

$$1 \le C < 8\kappa/(\kappa+1)^2. \tag{107}$$

For any admissible memory pair, $\epsilon > 0$, zero moments, either bias convention, and (27), cyclic coordinatewise Adam converges to the minimizer along every sample iterate. The average need not be separable and components need not be convex.

Proof. Translate the minimizer to zero. Write $g(x) = A(x)x$ with $A(x) = \int_0^1 \nabla^2 f(tx)\,dt$, so $mI \preceq A \preceq LI$. For frozen periodic directions $u_{ij} = m^*_{ij}/(\sqrt{v^*_{ij}} + \epsilon)$, the scalar moment argument gives

$$\mathbb{E}u_j^2 \le 1, \quad \mathbb{E}m^* = g, \quad \sum_j \mathbb{E}v_j^* = S, \quad \mathbb{E}\sqrt{v_j^*} \ge |g_j|. \tag{108}$$

These identities include zero streams and memoryless phases. For $w = x \odot |x|$, $\lambda > 0$, the identity $m^* = u(\sqrt{v^*} + \epsilon)$ and Cauchy–Schwarz give

$$(\lambda w + \epsilon x)^\top U_\epsilon \ge x^\top g - \|x\|\sqrt{S - 2\lambda x^\top g + \lambda^2\|x\|^2}. \tag{109}$$

Indeed the squared factors are at most $\|x\|^2$ and $S - 2\lambda\sum_j |x_j||g_j| + \lambda^2\|x\|^2$; replacing the middle sum by $x^\top g$ increases the second. Its radicand is at least $\|g - \lambda x\|^2$.

Put $c = (m+L)/2$, $a = x^\top g/\|x\|^2$, and $s = \|g\|^2/\|x\|^2$. The spectral interval gives $m \le a \le L$ and $s \le 2ca - mL$. Choose $\lambda = Cc/2$ and set $d = CmL - C^2c^2/2 > 0$. For $Q = Cs - 2\lambda a + \lambda^2 \ge 0$,

$$a^2 - Q \ge (a - Cc/2)^2 + d \ge d, \qquad a - \sqrt{Q} \ge d/(2L) =: \nu > 0.$$

Hence

$$\nabla V(x)^\top U_\epsilon(x) \ge \nu\|x\|^2, \qquad V(x) = \frac{\lambda}{3}\sum_j |x_j|^3 + \frac{\epsilon}{2}\|x\|^2. \tag{110}$$

Absolute tracking transfers directions globally because

$$\left|\frac{m}{\sqrt{v}+\epsilon} - \frac{m^*}{\sqrt{v^*}+\epsilon}\right| \le \frac{|m - m^*| + R_{q,b}|\sqrt{v} - \sqrt{v^*}|}{\epsilon}.$$

Including the vanishing bias perturbation gives

$$x_{k+1} = x_k - \eta_k[U_\epsilon(x_k) + e_k], \qquad \|e_k\|_\infty \to 0. \tag{111}$$

Outside $\|x\| < r$, its pairing error is at most $\|e_k\|_\infty(\lambda + \epsilon\sqrt{p}/r)\|x\|^2$. The Taylor remainder is bounded by $K\eta_k^2(\|x\| + \epsilon + \eta_k)$, and is absorbed uniformly there. Eventually $\Delta V \le -(\nu/4)\eta_k\|x\|^2$ on every such exterior. Since

$$\frac{\lambda}{3\sqrt{p}}\|x\|^3 + \frac{\epsilon}{2}\|x\|^2 \le V(x) \le \frac{\lambda}{3}\|x\|^3 + \frac{\epsilon}{2}\|x\|^2,$$

every positive potential level is separated from zero. Proper-potential trapping proves $V \to 0$, hence convergence. □

The sufficient region is dimension independent and nonempty precisely when $\kappa < 3 + 2\sqrt{2}$. At $q = b = 0$ this threshold is sharp for every $\epsilon \ge 0$ (Corollary 5). In a quadratic problem m, L can be the extreme eigenvalues of the average Hessian; for nonlinear averages they are common global spectral bounds.

Zero regularization and vanishing coordinate streams. Keep the Hessian, growth, initialization, and step assumptions, but set $\epsilon = 0$. Define $h = 4\kappa/(\kappa+1)^2$ and $R = R_{q,b}$, with $R = 1$ at $q = b = 0$. Every sample iterate converges whenever

$$1 \le C < T_0 := \frac{2h}{1 + (R-1)(1-h)}. \tag{112}$$

For $q = b$, $R = 1$ recovers the full region $C < 2h$, including memoryless Adam. The following argument accounts for possible limiting zero streams.

Take late starts with $r_k = \|x_k\| \ge r_* > 0$. Strong growth and Lipschitzness make the normalized arrays compact; along a subsequence,

$$x_k/r_k \to y, \qquad g_i(x_k)/r_k \to D_i, \qquad A(x_k) \to A,$$
$$\|y\| = 1, \qquad g := \mathbb{E}D_i = Ay, \qquad S := \mathbb{E}\|D_i\|^2 \le C\|g\|^2.$$

Absolute tracking remains valid after division by $r_k \ge r_*$. For $b > 0$, nonzero limiting coordinate streams have positive periodic denominators; their actual phase limits are periodic and satisfy $\mathbb{E}u_j^2 \le 1$. On identically zero streams Z, retain the actual limits, bounded in magnitude by R, rather than replacing them by zero. At $q = b = 0$, arbitrary limits in $[-1, 1]$ at zero phases still obey $u_{ij}|D_{ij}| = D_{ij}$ and $\mathbb{E}u_j^2 \le 1$, so the scalar bound applies to this completed graph.

Let U be the limiting epoch direction, $P = \{j : y_j g_j > 0\}$, $a = y^\top g$, and $C_j = \mathbb{E}D_{i,j}^2/g_j^2$ on P. There $\operatorname{sgn}(g_j)U_j \ge 2/C_j - 1$; outside $P \cup Z$, $|U_j| \le 1$. Weighted Cauchy–Schwarz and the spectral angle bound give

$$\sum_P \frac{y_j^2}{C_j} \ge \frac{a^2}{S}, \qquad \|y_Z\|^2 \le 1 - \frac{a^2}{\|g\|^2} \le 1 - h.$$

As $R \ge 1$, these imply

$$(y \odot |y|)^\top U \ge \frac{2a^2}{C\|g\|^2} - 1 - (R-1)\|y_Z\|^2 \ge \frac{2h}{C} - 1 - (R-1)(1-h) =: \delta > 0.$$

For $b > 0$, $R^2 - 1 = (b-q)^2/[b(1-b)(1-q^2/b)]$ verifies $R \ge 1$. Compactness gives a uniform margin $\nabla V(x_k)^\top U_k \ge (\delta/2)\|x_k\|^2$ outside every fixed ball, now with $V = \sum_j |x_j|^3/3$. Its Taylor remainder is bounded by $K\eta_k^2(\|x_k\| + \eta_k)$, so the same proper-potential trapping proves convergence.

Corollary 6 (Additive heterogeneity under bounded conditioning)**.** *Under the Hessian, Lipschitz, memory, positive-ϵ, initialization, and step assumptions of Theorem 21, replace strong growth by $S(x) \le C\|g(x)\|^2 + B$, $B \ge 0$, still with (107). Translate the minimizer to zero and set*

$$\lambda = C(m+L)/4, \quad d = CmL - C^2(m+L)^2/8, \quad R_0 = \sqrt{B/d}, \quad V = \lambda \sum_j |x_j|^3/3 + \epsilon\|x\|^2/2.$$

The whole trajectory is bounded and

$$\limsup_t V(x_t) \le \lambda R_0^3/3 + \epsilon R_0^2/2, \qquad \limsup_t \|x_t\| \le p^{1/6} R_0. \tag{113}$$

Here $C = D_1/n$ and $B = D_0/n$ in (1).

Proof. The same shifted pairing gives, for $r = \|x\| > R_0$,

$$\nabla V^\top U_\epsilon \ge r^2(a - \sqrt{Q + B/r^2}) \ge (dr^2 - B)/(2L),$$

since $a^2 - Q \ge d$ and $\sqrt{Q + B/r^2} < a \le L$. Absolute tracking needs only Lipschitz gradients, so it and the Taylor bound remain valid without interpolation. For every $R > R_0$ there is uniform eventual descent by a positive multiple of $\eta_k r^2$ outside that ball. Above any level greater than $M(R_0) = \lambda R_0^3/3 + \epsilon R_0^2/2$, continuity of $M(R) = \lambda R^3/3 + \epsilon R^2/2$ supplies such an R. Lemma 5 gives the potential bound and boundedness. The inequality $V \ge \lambda\|x\|^3/(3\sqrt{p}) + \epsilon\|x\|^2/2$ proves the radius. □

H EXPERIMENTS AND VERIFICATION

Protocols and records are indexed in `MANUSCRIPT_MAP.md` and `COHORT_AND_VALIDATION_INDEX.md`.

H.1 TRAJECTORIES OF THE CERTIFIED INSTANCE AT $\kappa = 9$

Ray tracking, matched statistics and bases. We run the $\kappa = 9$ instance of Appendix F.4 ($n = 75000$, $p = 4901$, $C \approx 1.834814$) with $(q, b) = (.9, .999)$, $\epsilon = 10^{-8}$, zero moments, bias correction, $x_0 = 1000e$ and sample step η_k/n. Over 3000 epochs with $\eta_k = 1$ and $\eta_k = (k+1)^{-1/2}$, respectively, the final mean radial speeds are 1.0000059 and 1.0000011 (prediction $a = 1$), the final moment errors are .04080 and .002704, and the maximum cross-section deviations are .01982 and .002586 (below .290% and .038% of $\|e\|_2 \approx 6.84407$). For this diminishing-step setting, `initialized_radius` certifies divergence whenever $x_0 = r_0 e$ with $r_0 \geq 2 \cdot 10^6$. The exact six-class reduction uses multiplicities $(980, 980, 980, 980, 980, 1)$. Applying the construction in the proof of Corollary 4 to this instance preserves n, p, A, C; after 1000 diminishing-step epochs, the objective ratios are 1.127894 for the certified instance and .100501 for its twin.

Figure 1(b) compares the original and eigenbases, with $\widetilde{H}_i = Q^\top H_i Q$ and $z_0 = Q^\top x_0$. At epoch 3000, the objective ratios are 1.228376 and .527186; the latter includes .49120 from the eigenvalue-one coordinate, which moves 108.09 from initial magnitude 6503.04. Full-dimensional operator checks and 15010 stock-Adam updates across the first block switch agree within $3.6 \cdot 10^{-15}$ parameter error.

Clipping and decoupled weight decay. The exact class reduction also runs the $\kappa = 9$ instance at $(q, b) = (.9, .999)$, $\epsilon = 10^{-8}$, with the optimizer settings of the Transformer study (Appendix C.2). With global-norm clipping at one, active at every sample, and $\eta_k = (k+1)^{-1/2}$ from $x_0 = 1000e$, the radius at epoch 3000 is 1108.093285, against 1108.093410 without clipping. With clipping and decoupled decay $\lambda = 0.01$ at constant $\eta_k = 0.1$, the radius falls from 1000 to 102.23 by epoch 6000 (prediction $a/\lambda = 100$), with normalized class coordinates within $1.9 \cdot 10^{-3}$ of the ray; with $\lambda = 0.05$ and $\eta_k = 0.02$ it reaches 22.43 (prediction 20). Started nearer the minimizer, at $30e$ and $5e$, the iterates move outward and approach the opposite ray $-e$, reaching radii 93.40 and 19.83 by epoch 6000. The code and records are in `optimizer_variants/`.

H.2 FUNCTION-PRESERVING TRANSFORMER BASES

Protocol. The 65280-parameter character Transformer has two pre-normalized blocks (width 48, four heads, feed-forward width 192), context 48, learned token and position embeddings, and no dropout. Tiny Shakespeare (Karpathy, 2015) is split chronologically 80/10/10, with the training-set vocabulary. Both arms start from identical FP32 logits using $\theta = \theta_0 + Qz$, $z_0 = 0$. The global Q combines a permutation and independent 64-coordinate QR blocks; tensor rotations act within each of the 29 tensors, using smaller final blocks.

AdamW (Loshchilov & Hutter, 2019) uses $(q, b) = (.9, .999)$, $\epsilon = 10^{-8}$, bias correction and global gradient clipping at norm one. Its internal decay is zero; instead, before each adaptive update we apply $z \leftarrow (1 - \gamma\lambda)z - \gamma\lambda Q^\top \theta_0$ with $\lambda = .01$, which decays the physical weights θ toward zero in both arms alike. Development seeds 9101–9102 use 600 updates with the equivalent parametrization $\theta = Qu$, $u = Q^\top \theta_0 + z$. Mean final validation NLL across both bases selects peak rate .008 from $\{.0005, .002, .008\}$.

Primary seeds 9201–9206 and independent seeds 9401–9406 use 1000 updates, batch size 16 (768000 characters per arm), shared batches, 50-update linear warmup and cosine decay to 10% of the peak rate. Validation uses 256 fixed windows; only final models see the 512 test windows (24576 positions). Splits, hashes and seed formulas are supplied.

Outcomes, sensitivity and implementation controls. All six test gaps are positive in each cohort. The independent mean is .04200 (SD .01028; paired 95% t_5-interval $[.03121, .05280]$); Table 4 gives the primary statistics. The primary training-loss gaps, averaged over all 1000 shared pre-update batches, are also all positive (mean .05210, SD .01034).

The secondary grid of Table 4 was fixed after the primary study and before its 1000-update runs; its global .008 row replays the primary setting. Seeds 9301–9306 also index an earlier unanchored replication with different rotations. Neither set is pooled with the independent seeds 9401–9406. In 30-update FP64 controls, momentum SGD and the shared-second-moment update agree across the two bases, after mapping between them, within $5 \cdot 10^{-17}$ relative parameter error, and AdamW agrees exactly under signed permutations. The supplement also retains checks with active clipping, physical decay, endpoints, precision and computing environments.

Table 4: Complete secondary grid: rotated minus original test NLL. Each cohort uses six pairs with separate paired Student-t intervals.

	Original: 9201–9206			Follow-up: 9301–9306		
Setting	Mean	SD	95% interval	Mean	SD	95% interval
Tensor .008	.01497	.00736	[.00724, .02269]	.01442	.00582	[.00831, .02053]
Global .004	.06201	.01481	[.04647, .07755]	.05321	.01855	[.03374, .07268]
Global .008	.04269	.01345	[.02858, .05680]	.03908	.01376	[.02464, .05352]
Global .016	.02298	.02370	[−.00189, .04785]	.02197	.01785	[.00323, .04070]

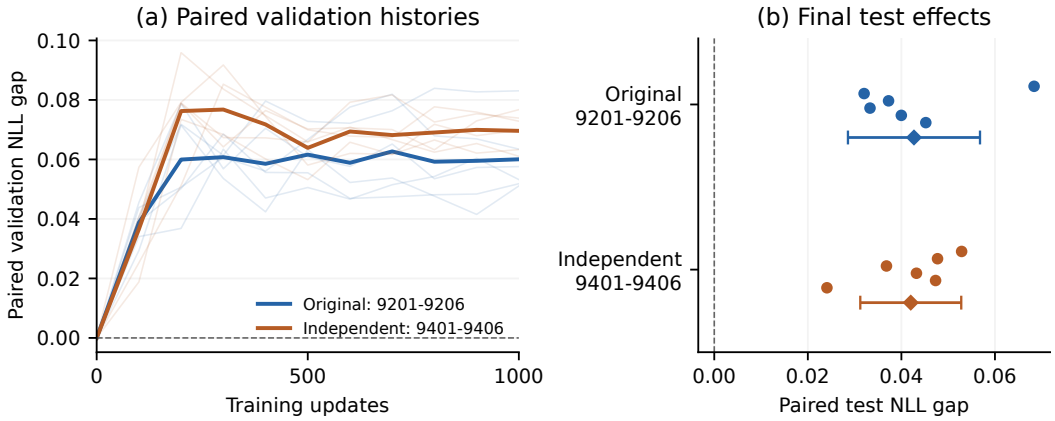


Figure 2: Paired Transformer NLL (rotated minus original), in nats per character. Left: all validation histories and cohort means. Right: final gaps (circles), means (diamonds) and paired 95% Student-t intervals.

H.3 Quantitative basis dependence at matched Transformer states

Use each endpoint's ordered N-minibatch bank d_i in both bases. Let $\mu = \mathbb{E}d$, $s_j = \mathbb{E}d_j^2$, and $W = \mu^\top \mathbb{E}[m \oslash (\sqrt{v}+\epsilon\mathbf{1})]$ for periodic raw moments. The pointwise bank ratios are $C_{\mathrm{bank}} = \sum_j s_j / \|\mu\|^2$ and $C_{\square,\mathrm{bank}} = \max_{s_j>0} s_j/\mu_j^2$, taking a zero active mean as $+\infty$. Only C_{bank} is rotation invariant. For each coordinate let $\widehat{d}_k = N^{-1}\sum_i d_i e^{-2\pi \mathrm{i} ki/N}$ and $h_r(k) = (1-r)/(1-re^{-2\pi \mathrm{i}k/N})$. By Parseval,

$$R = \mu^2 + \sum_{k=1}^{N-1} |h_q(k)\widehat{d}_k|^2 = \mathbb{E}m^2, \qquad T = \sum_{k=1}^{N-1} |h_b(k)\widehat{d^2}_k|^2 = \mathrm{Var}(v).$$

Proposition 8 (Spectral bounds for frozen work)**.** *Fix $N \geq 2$, $0 \leq q \leq b < 1$, $b > 0$, and $\epsilon \geq 0$. Let $\alpha = \alpha_{N,b}$, $\gamma = \gamma_{N,b}$ from (17)–(18), $\beta = N(1-q)q^{N-1}/(1-q^N)$ (zero at $q = 0$), and*

$\sigma_q = (1-q)^2/[(1-q)^2 + 4q\sin^2(\pi/N)]$. *In every coordinate with* $s > 0$*, set*

$$H = \min\left\{1 - \beta + \beta\mu^2/s, \frac{\mu^2 + \sigma_q(s-\mu^2)}{\alpha s}, \frac{R}{(\sqrt{\alpha s}+\epsilon)^2}\right\},$$

$$A = \max\left\{\sqrt{\mu^2 + \alpha(s-\mu^2)}, \sqrt{\gamma s}, \sqrt{s}\left[1 - \frac{T}{2s^2(1+\sqrt{\alpha})^2}\right]_+\right\}, \qquad Z = |\mu|\sqrt{H(s-A^2)}.$$

Define $u = (\mu^2 + Z)/(A+\epsilon)$ *and* $l = (\mu^2 - Z)/(\sqrt{s}+\epsilon)$ *when* $\mu^2 \geq Z$*, and* $l = (\mu^2 - Z)/(A+\epsilon)$ *otherwise. Then* $\sum_j l_j \leq W \leq \sum_j u_j$*; coordinates with* $s_j = 0$ *contribute zero.*

Proof. Write $y = m/(\sqrt{v}+\epsilon)$ and $a = \mathbb{E}\sqrt{v}$. Propositions 3 and 2, and $v \geq \alpha s$, give $\mathbb{E}y^2 \leq H$ and the first two terms in A. Also,

$$2\sqrt{s}(\sqrt{s} - a) = \mathbb{E}(\sqrt{v} - \sqrt{s})^2 \leq \frac{T}{s(1+\sqrt{\alpha})^2},$$

so $A \leq a \leq \sqrt{s}$ and $\mathrm{Var}(\sqrt{v}) \leq s - A^2$. The identity $\mu = (a+\epsilon)\mathbb{E}y + \mathrm{Cov}(y, \sqrt{v})$ implies $|\mu^2 - (a+\epsilon)\mu\mathbb{E}y| \leq Z$ by Cauchy–Schwarz. Choosing the appropriate denominator for each numerator sign proves the bounds. Zero means contribute zero without division by μ. □

The 144 cases cross six seeds, two endpoints, both bases, $N \in \{32, 128, 1024\}$ and $\epsilon \in \{0, 10^{-8}\}$ at $(q, b) = (.9, .999)$. Every FP64 interval encloses the computed work. At $\epsilon = 10^{-8}$, relative widths $(\sum u_j - \sum l_j)/W$ are .223–.256%, .507–.631% and 2.091–2.830% for the three periods. At $N = 1024$, C_{bank} spans 8.206–21.929 and $C_{\square,\mathrm{bank}}$ spans $1.175 \cdot 10^{10}$–$1.281 \cdot 10^{13}$, with every active coordinate included. At every period and both values of ϵ, all twelve same-state basis pairs have disjoint work bounds, although their aggregate ratios agree to $1.1 \cdot 10^{-14}$; at $N = 1024$ the bounds of the two bases are separated by a factor of at least 1.277. Full records and bank reconstruction are in `spectral_bridge/`.

Aggregate and coordinatewise thresholds at larger batches. For IID minibatches of s bank elements, the identity after Theorem 16 gives exactly $C_s = 1 + (C_{\mathrm{bank}} - 1)/s$ and, coordinatewise, $C_{j,s} = 1 + (C_j - 1)/s$ with $C_j = s_j/\mu_j^2$. Let s^* be the smallest s with $C_s < 2$; over all 72 moment sets s^* ranges from 7 to 21. At s^* the median of C_{j,s^*} over active coordinates lies between 3.03 and 6.40, and its 0.9-quantile also exceeds 2, in every set and both bases. On each bank, a moderate batch multiple therefore brings the aggregate ratio below the scalar threshold 2, while more than half of the coordinate ratios stay above it. The aggregate and the coordinatewise conditions of Section 5, evaluated on the same bank, thus give different answers (`spectral_bridge/threshold_range.py`).