REPLAY-GAP THRESHOLDS: SHARP BURST GUARANTEES AND COUPLED RETENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse replay can retain training examples even when their update frequency vanishes, but the placement of updates matters. We prove a sharp worst-case replay-gap threshold for constant-step online logistic regression on any separable interfering pair. For a fixed burst of B rare updates, let $S_B(m)$ be its total scalar logistic response from margin m, let $\kappa = b/(ac^2)$ in coordinates $(a, 0), (-b, c)$, and set $R_{c,B} = \exp(S_B(0)/\kappa)$. Under an explicit sufficient step-size bound, every replay-time ratio sequence confined eventually to a fixed interval $1 < R_- \le R_k \le R_+ < R_{c,B}$ guarantees eventual all-time retention from every finite initialization. Geometric schedules above the threshold establish worst-case sharpness; equality remains open. The proof uses a common invariant region controlled by the derivative of the full burst response. We also extend the centered-map method to any finite number of orthogonal common examples coupled through one interfering rare example, obtaining an exact geometric sign criterion in the stated exposure regime and a computable local guarantee for arbitrary persistent ratio variation. In this family, two schedules whose rare-update counts differ by at most one at every prefix can have opposite retention outcomes, robustly under small bounded shared log-time jitter. The results concern logarithmically sparse replay and raw training margins, not population accuracy.

1 INTRODUCTION

How far apart can replay events be while still ensuring retention? A cumulative replay count does not specify whether its updates are spread out or concentrated into bursts. With logistic updates, this distinction affects not only the time of intervention but also its strength: an update raises the rare example's margin, reducing the response of the next update in the same burst. We quantify that effect and ask which conclusions survive persistent irregularity in replay timing.

Our central result is a sharp upper-gap guarantee for any separable, noncollinear interfering pair and every fixed burst size B. The relevant response is $S_B(0)$, the total scalar step of a burst starting at zero margin. It determines the same threshold for constant geometric replay and arbitrary switching within every fixed strictly subcritical ratio interval. The sufficient step range scales as $1/B$; it is not asserted optimal. The singleton formula $R_{c,1} = \exp(\eta D/(2I))$, where I is interference and D the Gram determinant, and the separate canonical unit-step result are recovered as special cases.

The distinction between constant-ratio stability and switching stability is essential. Stable geometric reference cycles alone do not supply a common set for arbitrary switches. We construct such a set and prove global attraction despite vanishing discrete approximation errors. Its invariance depends on a derivative bound for the entire burst response, which makes the extension from single updates to bursts possible. The resulting lower bound on the rare margin holds at every sufficiently late integer time, not just immediately after replay.

We then test the method beyond two examples. For any finite number of orthogonal common examples and one rare example interfering with each, the common coordinates remain coupled through a shared logistic response. We derive the exact geometric threshold and an explicit common-ball criterion for persistently irregular ratios. The latter guarantee is local in the ratio interval but global in initialization. We do not claim the sharp nonlocal pair guarantee for this larger family. A further consequence is a robust count obstruction: even here, rare-count functions differing by at most one at every prefix can yield opposite eventual retention outcomes.

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.

These results isolate timing at logarithmically sparse replay budgets. They do not propose a general replay algorithm or establish generalization. Complete prefix counts determine a binary schedule by their increments; our obstruction concerns bounded differences of those count functions and their leading logarithmic coefficient, not identical schedules. The new family demonstrates transfer to a coupled system of arbitrary finite size while retaining explicit structural assumptions.

Relation to prior work. Stationary separable logistic regression has an implicit max-margin bias (Soudry et al., 2018). The closest finite-budget continual-classification results already include one-step visits: Jung et al. (2025, Sections 2.1–2.2) allow $K \geq 1$ gradient steps per task. Their Theorems 3.1–3.2 establish joint-loss and directional conclusions under cyclic visits, and Theorems 4.1–4.2 treat independent uniform task sampling, under their respective step-size and loss assumptions. The directional statements additionally assume nondegenerate support vectors and tight exponential loss tails. Our rare example has zero limiting frequency and logarithmic count; the guarantees concern raw training signs under deterministic timing constraints. Their broader jointly separable datasets and our sparse switching schedules are different regimes.

Evron et al. (2023, Theorem 3.1) connect weakly regularized taskwise exponential-loss minimization with sequential unit-margin projections. Their Theorem 5.1 connects regularization-strength scheduling to scaled projection dynamics, and Proposition 5.2 shows that these dynamics can fail to reach joint feasibility even under cyclic task ordering. Section 7.1 also gives a normalized singleton projection relation for gradient flow stopped at a fixed positive target loss. Our update is one finite unregularized logistic step per occurrence. Neither one-step learning nor schedule sensitivity is claimed as new; the contributions are the sparse-gap thresholds, their switching guarantees, and the robust count obstruction for this update rule.

Theorem-level comparisons also distinguish recent replay analyses. Ding et al. (2026, Theorems 1–2) study signal/noise regimes and population classification under full replay in a feature-learning model. Peng et al. (2026, Theorem 1) bound weighted estimation error for global minimizers of weighted nonlinear-regression replay objectives under task-dependency assumptions, allowing zero regularization. Lanzillotta et al. (2026, Theorem 3 and Corollary 2) give feature-space separability bounds under a terminal neural-collapse replay-mixture model. Our fixed-feature, finite-step, zero-frequency setting addresses a different mechanism. These scoped comparisons do not establish exhaustive priority or exclude every possible reduction.

2 Model

Signed examples are $z_i = y_i x_i$. Given a deterministic schedule $i_n \in \{1, 2\}$, consider

$$w_n = w_{n-1} + \frac{\eta z_{i_n}}{1 + \exp(\langle z_{i_n}, w_{n-1} \rangle)}, \qquad n \geq 1. \tag{1}$$

The step size η is fixed. Strict correct classification means $\langle z_i, w_n \rangle > 0$. A schedule possesses *eventual all-time retention* if both inequalities hold at every sufficiently large integer n. Denote its rare-update count by $N(n) = \sum_{j=1}^{n} \mathbf{1}\{i_j = 2\}$. Infinite replay alone is not assumed to imply retention.

3 An exact criterion for geometric bursts

Any non-collinear interfering signed pair can be expressed in orthonormal coordinates by

$$z_1 = (a, 0), \quad z_2 = (-b, c), \quad a, b, c > 0.$$

In particular, $a = \|z_1\|$, $b = -\langle z_1, z_2 \rangle / a$, and $c = (\|z_2\|^2 - b^2)^{1/2}$. The orthogonal components of initialization are invariant and do not affect the margins. Let

$$Q = b^2 + c^2, \quad \alpha = \eta a^2, \quad \beta = \frac{b}{ac}, \quad \kappa = \frac{b}{ac^2}, \qquad 0 < \eta Q < 8.$$

A burst contains $B \geq 1$ successive rare updates, and all other updates are common. A rare update transforms its margin through

$$F(m) = m + \frac{\eta Q}{1 + e^m}, \qquad S_B(m) = \frac{F^B(m) - m}{Q}. \tag{2}$$

So S_B is the total scalar step over a burst. It decreases strictly from $B\eta$ to zero. This continues to hold even if F is not monotone: every $F'(m)$ lies in $(-1, 1)$, giving $S'_B = ((F^B)' - 1)/Q < 0$.

Proposition 1 (Reference cycle and exact strict sign test)**.** *Assume $0 < \eta Q < 8$. Let $R > 1$, $M = \log R < B\eta/\kappa$, and place bursts at $t_k = \lfloor R^k \rfloor$ after any sufficiently late start. For every finite initialization, let m_B be the unique solution of*

$$S_B(m_B) = \kappa M. \tag{3}$$

Then $au_n = \log n + O(1)$ and the rare margin $m_n = -bu_n + cv_n$ satisfies

$$\liminf_n m_n = m_B, \qquad \limsup_n m_n = m_B + Q\kappa M.$$

Hence $S_B(0) > \kappa M$ implies eventual all-time retention, and $S_B(0) < \kappa M$ implies infinitely many strict rare-example misclassifications. No sign assertion is made at equality.

In case of $B = 1$ and $R = \rho$ regular replay, the positive margin requirement is

$$\log \rho < \frac{\eta a c^2}{2b}. \tag{4}$$

On the other hand, for a given interfering pair, any $L = \log \rho$ that meets (4) causes failure of retention in sufficiently large fixed bursts with $R = \rho^B$. This is because starting F at zero results in $S_B(0) = O(\log B)$, while the required response κBL increases linearly. The regular and burst schedules can be paired with count discrepancy at most $B - 1$; Appendix B gives the details. The required threshold is therefore geometry-dependent, rather than a universal replay-count coefficient.

At the canonical parameter values in Theorem 4, both reference signs are strict:

$$S_1(0) = \tfrac{1}{2} > L, \qquad S_2(0) = \tfrac{1}{2} + \frac{1}{1+e} < \frac{17}{22} < \frac{4}{5} < 2L. \tag{5}$$

Hence regular and paired geometric replay have opposite retention outcomes. Theorem 8 proves that this remains true for persistently irregular replay locations.

4 A SHARP UPPER-GAP GUARANTEE FOR FIXED BURSTS

The geometric threshold extends to arbitrary persistent switching for every fixed burst size, with a sufficient step restriction that accounts for the full burst.

Theorem 2 (Sharp worst-case threshold for fixed bursts)**.** *Let $z_1 = (a, 0)$, $z_2 = (-b, c)$ with $a, b, c > 0$, fix an integer $B \geq 1$, and put*

$$R_{c,B} = \exp\left(\frac{S_B(0)}{\kappa}\right), \qquad \eta_0 = \min\left\{\frac{b}{aQ}, \frac{1}{Q}, \frac{2c^2}{Q^2}\right\}. \tag{6}$$

Assume $0 < \eta \leq \eta_0/B$, or $a = b = c = \eta = B = 1$. Update example 2 at $t_k, \ldots, t_k + B - 1$ and example 1 otherwise, with eventually disjoint bursts. If eventually

$$1 < R_- \leq R_k := t_k/t_{k-1} \leq R_+ < R_{c,B},$$

then every finite initialization and finite initial update prefix has eventual all-time retention, with

$$\liminf_{n\to\infty} z_2 \cdot w_n \geq \frac{b}{a} \log \frac{R_{c,B}}{R_+} > 0. \tag{7}$$

For every $C > R_{c,B}$, some geometric burst schedule has eventual ratios below C but infinitely many strict rare-example misclassifications. Equality at $R_{c,B}$ is not decided.

For $B = 1$, $S_1(0) = \eta/2$. With $I = -\langle z_1, z_2\rangle > 0$ and $D = \|z_1\|^2 \|z_2\|^2 - \langle z_1, z_2\rangle^2 > 0$, this becomes

$$R_{c,1} = e^{\eta D/(2I)}, \qquad \eta_0 = \min\left\{\frac{I}{\|z_1\|^2 \|z_2\|^2}, \frac{1}{\|z_2\|^2}, \frac{2D}{\|z_1\|^2 \|z_2\|^4}\right\}. \tag{8}$$

For $B > 1$, $S_B(0) < B\eta/2$: later updates in the burst see a larger margin. Thus its critical logarithmic gap per rare update, $\log R_{c,B}/B$, is smaller than the singleton value $\eta/(2\kappa)$. The theorem makes this response penalty exact even under arbitrary switching. Sharpness is worst-case, not necessity for each varying schedule. The step range is sufficient, not optimal; the canonical $B = 1, \eta = 1$ case holds beyond $\eta_0 = 1/2$.

Proof mechanism. Center the postburst state by $(1/a, b/(ac)) \log t_k$. Set $X = e^{aU}/\alpha$, $Y = e^{(ac/b)V}/\alpha$, $\chi = b/a$, $\mu = ab$, and $\nu = ac^2/b$. The limiting map is

$$Z = X + R - 1, \quad m = \chi \log(Y/Z), \quad s = S_B(m), \qquad X' = \frac{Z}{R} e^{-\mu s}, \quad Y' = \frac{Y}{R} e^{\nu s}. \tag{9}$$

Write $s_0 = S_B(0)$, $r = R_{c,B}$, $h = r - 1$, $E = e^{(\mu+\nu)s_0}$, $A = h/(E-1)$, $Y_* = A + h$, and $P = A^\nu Y_*^\mu$. The common invariant region is

$$K = \{(U, V) : X^\nu Y^\mu \geq P, \quad Y - X \geq h\}. \tag{10}$$

The rare burst preserves its weighted product. The bound $0 < -S_B'(m) \leq B\eta/4$ controls the second boundary throughout the region. Both constraints are invariant for every $1 \leq R \leq r$. Compactness and strict contraction toward the projection onto K yield attraction for the exact discrete iterates with vanishing errors. Appendix D proves this argument; Appendix E supplies the stronger canonical unit-step case.

5 Retention in a coupled multi-example family

The centered-map method also applies beyond a pair. Let $d \geq 2$ and take

$$z_i = a_i e_i \ (1 \leq i \leq d), \qquad z_* = (-b_1, \ldots, -b_d, c), \qquad a_i, b_i, c > 0. \tag{11}$$

At common times cycle through $z_1, \ldots, z_d$, pausing and resuming that cycle around bursts of B updates on z_*. Use update (1) with this larger index set. Put $Q_* = \sum_i b_i^2 + c^2$, $\kappa_* = (\sum_i b_i/a_i)/c^2$, and define $S_{B,*}$ as in (2) with Q_*.

Theorem 3 (Coupled geometric criterion and persistent-ratio bound)**.** *Assume $0 < \eta Q_* < 8$, fix B, and choose $R_0 > 1$ with $\kappa_* \log R_0 < B\eta$. Let m_* solve $S_{B,*}(m_*) = \kappa_* \log R_0$. For geometric burst starts $\lfloor R_0^k \rfloor$, every finite initialization satisfies*

$$\liminf_n z_* \cdot w_n = m_*, \qquad \limsup_n z_* \cdot w_n = m_* + Q_* \kappa_* \log R_0,$$

and $z_i \cdot w_n = \log n + O(1)$ for each common example. Within this exposure regime, the exact strict geometric sign threshold is $\exp(S_{B,*}(0)/\kappa_*)$.

For any $h > 0$ and interval $1 < R_- \leq R_0 \leq R_+$ with $\kappa_* \log R_+ < B\eta$, Appendix F defines explicit $q < 1$ and $\Delta \geq 0$. If

$$\Delta < (1 - q)h, \qquad \sqrt{Q_*} \left(\frac{\Delta}{1 - q} + \Delta \right) < |m_*|, \tag{12}$$

then every eventually disjoint burst schedule with eventual ratios in $[R_-, R_+]$ has eventual all-time retention when $m_* > 0$, and strict rare misclassification before every sufficiently late burst when $m_* < 0$, from every finite initialization and finite prefix. Every $m_* \neq 0$ admits a nondegenerate such interval. No equality claim is made.

This is a coupled system: the same rare response depends on all d common coordinates and changes every one of them. No equality among the $a_i b_i$ is imposed. For example, $d = 2$, $(a_1, a_2, b_1, b_2, c) = (1, 3/2, 1, 1/2, 1)$, $\eta = 1/4$, $B = 1$, and $R_0 = e^{1/20}$ give $m_* = \log(11/4) > 0$. The theorem supplies a persistent irregularity tolerance for this nonsymmetric three-example instance. Its ratio interval is local; we do not extend the sharp nonlocal guarantee of Theorem 2 to this family.

6 Count discrepancy and timing

Theorem 4 (Robust discrepancy-one separation)**.** *Set $z_1 = (1, 0)$, $z_2 = (-1, 1)$, $\eta = 1$, and $L = \log(3/2)$. There exist $\delta > 0$ and an integer k_0 such that, for every real sequence (ξ_j) satisfying $|\xi_j| \leq \delta$ and every common finite $w_0 \in \mathbb{R}^2$, the following holds. Put*

$$a_j = \lfloor \exp(jL + \xi_j) \rfloor.$$

The regular schedule updates example 2 at every a_j, $j \geq 2k_0$. The paired schedule updates it at a_{2k} and $a_{2k} + 1$, $k \geq k_0$. Both update example 1 at all remaining times. The regular schedule has

eventual all-time retention, whereas the paired schedule misclassifies example 2 infinitely often. At every prefix,

$$0 \le N_{\mathrm{paired}}(n) - N_{\mathrm{regular}}(n) \le 1. \tag{13}$$

Both counts are $\log n/L + O(1)$ *and agree at each block endpoint* $a_{2k+2} - 1$*. The start* k_0 *is independent of the sequence and initialization; the eventual classification time need not be.*

The timing variables perturb one common base sequence; the independent perturbation of the two generated schedules is not claimed to maintain discrepancy one. They can be nonconvergent, nonperiodic, or selected arbitrarily within the bounded range. The theorem is deterministic. Each block $[a_{2k}, a_{2k+2} - 1]$ of both schedules has two rare updates. The first update is the same for both schedules, the second update of the paired schedule is finished at $a_{2k} + 1$, and the regular schedule catches up at a_{2k+1}. Sufficiently large uniform gaps guarantee the distinction of these events. This establishes the count identity independently of the learning dynamics.

Corollary 5 (Limitation of coarse count characterizations)**.** *For the fixed data, step size, and initialization in Theorem 4, no binary characterization of eventual all-time retention that is invariant under* $\sup_n |N(n) - N'(n)| < \infty$ *is correct for all schedules. In particular, the coefficient of* $\log n$ *in* $N(n)$ *does not determine retention.*

Proof. These two schedules belong to the same bounded-discrepancy class and have identical leading coefficients, but opposing retention truth values. □

This rules out exact characterizations on such summaries. It does not contradict a sufficient condition that rejects both schedules or a criterion that retains extra timing information.

Corollary 6 (Discrepancy one for every interfering pair)**.** *For every* $a, b, c > 0$ *and* $0 < \eta Q < 8$*, choose*

$$\frac{S_2(0)}{2\kappa} < L < \frac{\eta}{2\kappa}.$$

There exist $\delta > 0$ *and an integer* k_0*, depending only on the pair, step size and* L*, such that the regular and paired schedules of Theorem 4, formed with this* L *and any shared sequence* $|\xi_j| \le \delta$*, have opposite retention outcomes on the present pair from every common finite initialization. Their prefix-count discrepancy is zero or one, their counts are* $\log n/L + O(1)$*, and they agree at each block endpoint, exactly as in that theorem.*

Proof. The interval is nonempty since

$$S_2(0) = \frac{\eta}{2} + \frac{\eta}{1 + e^{\eta Q/2}} < \eta.$$

The selection provides $S_1(0) = \eta/2 > \kappa L$ but $S_2(0) < 2\kappa L$, and satisfies the two reference exposure conditions. Consequently the $B = 1, R_0 = e^L$ root is positive, and the $B = 2, R_0 = e^{2L}$ root is negative. Theorem 8 yields sufficiently small ratio neighborhoods preserving both strict signs from every finite starting point. The floor-error and event-ordering arguments from Appendix C work for any fixed $L > 0$: select a shared δ fitting both neighborhoods and then a uniform late k_0. They give the stated sign separation and exact count identities. □

Corollary 7 (Coupled discrepancy-one separation)**.** *For every geometry* (11) *and* $0 < \eta Q_* < 8$*, choose*

$$\frac{S_{2,*}(0)}{2\kappa_*} < L < \frac{\eta}{2\kappa_*}.$$

There exist $\delta > 0$ *and an integer* k_0*, depending only on the geometry,* η *and* L*, such that the following holds for every shared sequence* $|\xi_j| \le \delta$ *and every common finite initialization. Form the regular and paired rare schedules of Theorem 4 using this* L *and* k_0*; at all other times resume the prescribed common cycle, starting both streams in the same phase. Regular replay has eventual all-time retention; paired replay has infinitely many strict rare misclassifications. Their rare counts differ by zero or one at every prefix, agree at block ends, and both equal* $\log n/L + O(1)$*. Each common example's counts differ by at most one. The start* k_0 *is uniform over the jitter sequence and initialization; the eventual classification time need not be.*

Proof. $S_{2,*}(0) < \eta$, so the interval is nonempty. The two geometric roots have opposite strict signs. Apply Theorem 3 and the shared-jitter event-order argument of Appendix C. Each schedule resumes its own common cycle. When both common streams start in the same phase, their counts for each common example also differ by at most one: the total common counts differ by at most one, and the round-robin streams are prefixes of the same cycle. No identical common update at each integer time is asserted. □

7 A COMPUTABLE SUFFICIENT CONDITION FOR IRREGULAR TIMING

Throughout this section, $0 < \eta Q < 8$ is assumed; the extra small-step assumption in (6) is unnecessary. Fix B and reference $R_0 > 1$ such that $M = \log R_0 < B\eta/\kappa$. Suppose m_* is the solution to (3) at R_0 and set

$$A = \frac{\alpha(R_0 - 1)}{R_0^{1+b^2/c^2} - 1}, \quad U_* = \frac{\log A}{a}, \quad V_* = \frac{m_* + bU_* + Q\kappa M}{c}, \quad p_* = (U_*, V_*). \tag{14}$$

For $R > 1$, define the centered common-gap and rare maps by

$$f_R(U) = \frac{1}{a}\log\left(\frac{e^{aU}}{R} + \alpha(1 - R^{-1})\right), \quad C_R(U, V) = (f_R(U), V - \beta \log R), \tag{15}$$

$$G(U, V) = (U - bD, V + cD), \qquad D = \frac{\eta}{1 + e^{-bU+cV}}, \qquad T_R = G^B \circ C_R. \tag{16}$$

Direct substitution shows $T_{R_0}(p_*) = p_*$ and $z_2 \cdot C_{R_0}(p_*) = m_*$. The geometry of these maps provides explicit bounds, rather than a trajectory-dependent stability condition.

Take $h > 0$ and $1 < R_- \le R_0 \le R_+ < \infty$, along with $\log R_+ < B\eta/\kappa$. Define the following finite quantities:

$$\begin{aligned} \theta &= \frac{e^{a(U_*+h)}}{e^{a(U_*+h)} + \alpha(R_- - 1)}, & g(R) &= c(V_* - \beta \log R) - bf_R(U_*), \\ M_0 &= \max_{R\in\{R_-,R_+\}} |g(R)|, & H &= M_0 + \sqrt{Q}h + B\eta Q, \\ s_{\min} &= \frac{e^H}{(1 + e^H)^2}, & \zeta &= \left[\max\{|1 - \eta Q s_{\min}|, |1 - \eta Q/4|\}\right]^B. \end{aligned} \tag{17}$$

Let $P = z_2 z_2^\top / Q$, $D_\theta = \mathrm{diag}(\theta, 1)$, and

$$E = D_\theta[I - (1 - \zeta^2)P]D_\theta, \qquad q = \sqrt{\frac{\operatorname{tr} E + \sqrt{(E_{11} - E_{22})^2 + 4E_{12}^2}}{2}}, \tag{18}$$

$$d = \max_{R\in\{R_-,R_+\}} \sqrt{(f_R(U_*) - f_{R_0}(U_*))^2 + \beta^2(\log R - \log R_0)^2}. \tag{19}$$

These expressions give $q < 1$. They use a unique monotone scalar root and finite formulas; no simulated or actual iterate enters them.

Theorem 8 (Persistent-ratio timing bound)**.** *Let integer burst starts t_k be eventually disjoint and satisfy $t_k/t_{k-1} \in [R_-, R_+]$ eventually; the ratios need not converge. Define*

$$p_k = \left(u_{t_k+B-1} - \frac{\log t_k}{a},\ v_{t_k+B-1} - \beta \log t_k\right).$$

If

$$d < (1 - q)h, \tag{20}$$

then, for every finite initialization and every finite initial update prefix,

$$\limsup_{k\to\infty} \|p_k - p_*\| \le \frac{d}{1 - q}, \qquad \limsup_{k\to\infty} |z_2 \cdot w_{t_k-1} - m_*| \le \sqrt{Q}\left(\frac{d}{1 - q} + d\right). \tag{21}$$

Table 1: Numerical evaluation of the sufficient interval condition. Here $R_0 = (3/2)^B$; the last column is the right-hand side of the margin bound in (21). Values are rounded, not certified enclosures.

B	m_*	q	d	Margin-error bound
1	+0.382745	0.946933	0.00000361	0.00010119
2	-0.120146	0.978260	0.00000328	0.00021820

Also $au_n = \log n + O(1)$ *at all late times. If* $m_* \neq 0$ *and*

$$\sqrt{Q}\left(\frac{d}{1-q} + d\right) < |m_*|, \tag{22}$$

then $m_* > 0$ *implies eventual all-time retention, while* $m_* < 0$ *implies strict rare-example misclassification before every sufficiently late burst. For every admissible nonzero-margin reference there is a nondegenerate interval* $R_- < R_0 < R_+$ *satisfying* (20) *and* (22)*.*

Global entry. The proof first establishes $e^{au_n} = \Theta(n)$, then obtains the centered recursion $p_k = T_{R_k}(p_{k-1}) + o(1)$ without presupposing transverse boundedness. Global nonexpansiveness and contraction on the reference ball give a fixed inward drift outside it; the strict slack in (20) dominates the vanishing error. The resulting limsup recurrence proves (21). Appendix A gives the discrete indexing, matrix bound, and global-entry argument.

7.1 From the interval condition to arbitrary jitter

For the two strict reference signs indicated in (5), select intervals provided in Theorem 8. Small bounded ξ_j lead to

$$\log a_j = jL + \xi_j + O(e^{-jL+\delta})$$

uniformly for all permissible sequences. Thus the regular successive log ratios differ from L by no more than $2\delta + o(1)$, and the paired-start log ratios from $2L$ by no more than $2\delta + o(1)$. One sufficiently late start places both ratio sequences within their respective intervals. This proves Theorem 4; Appendix C checks the uniform floor and indexing details.

8 Interpretation, evaluation, and limits

This comparison of counts identifies the temporal positioning of rare events at asymptotically matched update budgets. However, it does not fix the state evolution completely since shifting a rare update in time also shifts the value of its margin and, consequently, its logistic response. Later rare updates in a burst encounter margins increased by earlier rare updates. The sum of their responses $S_B(m)$ is the value that compensates the interference caused by the common interval. The root equation (3) makes this balance exact, and the timing bound controls its perturbation.

A concrete evaluation of the bound. Table 1 checks the displayed inequalities at the canonical pair with $h = 0.25$ and $[R_-, R_+] = R_0 \exp([-3 \cdot 10^{-6}, 3 \cdot 10^{-6}])$. Both strict inequalities exhibit positive numerical slack. Selection of $\delta = 10^{-6}$ allows eventual floor errors as $2\delta < 3 \cdot 10^{-6}$. These numbers are deliberately conservative, fixed prior to computation and unoptimized. The computation uses Python Decimal arithmetic with 80 and 120 digits and 256 bisection steps for roots; 14 output comparisons agree to 40 decimal places. It is precision replication of a deterministic numerical calculation, not an interval certificate or an empirical training experiment. The existence of a positive admissible tolerance is proved separately by continuity.

What retention means here. For the pair, the centered bounds in Theorem 2 give

$$w_n = (1/a, b/(ac)) \log n + O(1)$$

within its span. The leading direction is orthogonal to z_2, and therefore $(z_2 \cdot w_n)/\|w_n\| \to 0$. Hence, positive raw training margins are consistent with convergence of the normalized rare margin to zero. Fixed cyclic replay has bounded gaps and ratios converging to one, outside $R_- > 1$; the sparse-recurrence theorem does not recover joint max-margin behavior. Discrepancy one pertains to counts and permits long displacements of individual replay events.

Where the guarantee stops. The general sharp theorem uses the explicit sufficient step range (6). Its threshold is optimal in the stated worst-case sense; we do not claim that the step range is optimal. The canonical unit-step example shows that this range can be enlarged. For every fixed burst size B, the sharp pair theorem holds when $\eta \leq \eta_0/B$; the local theorem covers the larger range $0 < \eta Q < 8$. The coupled multi-example switching guarantee remains local around a nonzero-margin geometric reference. Equality cases, $R_- = 1$, collinear opposition, nonnegative signed inner product, variable burst sizes, and unbounded ratio variability are outside the theorem. The replay growth considered here is logarithmic, so the limiting replay fraction is zero. The multi-example theorem assumes orthogonal common examples, one rare example, and a fixed common cycle. Arbitrary larger datasets, multiple rare directions, stochastic replay laws, representation learning, and population accuracy remain outside the results. It also establishes no uniform bound on classification transients.

The sufficient local constants can be conservative. In particular, H includes a worst-case estimate for burst intermediates, and the sign estimate uses Euclidean norm upper bounds. The small illustrative radius determines neither the maximum tolerable jitter nor robustness at a prescribed larger scale. The sharp fixed-burst pair theorem removes this local-radius restriction within its stated step range. More informative timing summaries may still yield necessary conditions for individual varying schedules. What is established here is that a geometry-dependent upper-gap bound guarantees retention in each strictly subcritical ratio interval, while bounded closeness of counts alone does not preserve retention.

REPRODUCIBILITY STATEMENT

These results are deterministic. Full proofs follow below; the accompanying standard-library Python script `reproduction/tolerance.py` reproduces the formula evaluation in Table 1. There are no training runs, random seeds, external datasets, or empirical baseline scores that support the theoretical claims. Numerical evaluation is distinct from proof.

REFERENCES

Meng Ding, Jinhui Xu, and Kaiyi Ji. Provable effects of data replay in continual learning: A feature learning perspective. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL https://virtual.aistats.org/virtual/2026/poster/13789.

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/evron23a.html.

Hyunji Jung, Hanseul Cho, and Chulhee Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/94ab02a30b0e4a692a42ccd0b4c55399-Paper-Conference.pdf.

Giulia Lanzillotta, Damiano Meier, and Thomas Hofmann. Heads collapse, features stay: Why replay needs big buffers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9b9521a6c8ead802f1b5dbd8cca34918-Abstract-Conference.html.

Liangzu Peng, Uday Kiran Reddy Tadipatri, Ziqing Xu, Eric Eaton, and Rene Vidal. Recovery guarantees for continual learning of dependent tasks: Memory, data-dependent regularization, and data-dependent weights. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL https://virtual.aistats.org/virtual/2026/poster/13773.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Proof dependencies. Appendix A establishes the scalar response, discrete growth and centered recursion before proving local contraction and global entry for Theorem 8. Appendix B derives Proposition 1; Appendix C supplies the shared-jitter counts. Appendix D uses the growth and recursion, then proves the common invariant region and compact attraction for Theorem 2; Appendix E treats its canonical unit-step case. Appendix F proves Theorem 3 by extending the centered-map argument; Corollary 7 combines it with the shared-jitter construction.

A Complete proof of the persistent-ratio theorem

A.1 The scalar root and reference fixed point

For finite m, $F'(m) = 1-\eta Q e^m/(1+e^m)^2 \in (-1,1)$. Every factor of $(F^B)'$ therefore lies strictly inside that interval; their product is less than one, even when some factors are negative. Hence $S'_B < 0$. Since $m \le F^j(m) \le m + j\eta Q$, the limits of S_B at negative and positive infinity are respectively $B\eta$ and zero. The exposure bound gives a unique finite root, whose sign is the sign of $S_B(0) - \kappa M$.

The constant A for (14) is such that it meets $0 < A < \alpha$ and

$$A + \alpha(R_0 - 1) = AR_0^{1+b^2/c^2}.$$

Therefore, $f_{R_0}(U_*) = U_* + b\kappa M$. Substitution of V_* gives $z_2 \cdot C_{R_0}(p_*) = m_*$. The rare burst has total scalar step κM, hence vector displacement $(-b\kappa M, c\kappa M)$. Since $c\kappa = \beta$, this cancels the common-map shifts and proves $T_{R_0}(p_*) = p_*$.

A.2 All-time growth without an assumed attracting set

Write $w_n = (u_n, v_n)$ and $y_n = e^{au_n}$. A common update has

$$0 < y_{\text{new}} - y = y\left(e^{\alpha/(1+y)} - 1\right) \le \alpha e^{\alpha}.$$

Rare updates reduce y, hence $y_n \le y_0 + \alpha e^{\alpha} n$. Each rare update lowers u by at most $b\eta$, or multiplies y by a factor at least $e^{-ab\eta}$.

Discard a finite prefix after which the ratio constraints apply. Then $k = O(\log t_k)$, and the common gap that ends at $t_k - 1$ has $t_k - t_{k-1} - B \ge (1 - R_-^{-1})t_k - B$ steps. The lower bound $u \ge -b\eta Bk + O(1)$ is true prior to and inside the gap, because common steps cause u to grow. During $u < 0$, each common step contributes at least $\eta a/2$, so $O(k)$ common steps are enough to raise it to zero if needed. Once $y \ge 1$, a common step adds at least $\alpha y/(1+y) \ge \alpha/2$. The remaining order-$t_k$ steps give $y_{t_k - 1} \ge Ct_k$ for all large k.

During the subsequent burst, this lower bound retains at least the factor $e^{-ab\eta B}$. Following that y grows until the next burst, while $t_{k+1} \le R_+ t_k$. This process results in a lower bound $y_n \ge C'n$ for all following gaps and burst intervals. Thus

$$y_n = \Theta(n), \qquad au_n = \log n + O(1), \tag{23}$$

holds uniformly at late times, with constants depending on initialization. No estimate of v_n was assumed.

Because of $y \to \infty$, its common step increment is $\alpha + O(1/y)$. Because of (23), the summation of the error through any whole or partial bounded-ratio gap is $O(1)$, since the sum of $1/n$ over that interval is bounded. The precise number of common steps in between bursts is $t_k - t_{k-1} - B$, so

$$y_{t_k - 1} = y_{t_{k-1} + B - 1} + \alpha(t_k - t_{k-1}) + O(1).$$

After division by t_k, its leading expression is $e^{aU_{k-1}}/r_k + \alpha(1 - r_k^{-1})$, at least $\alpha(1 - R_-^{-1}) > 0$. Taking logarithms gives the centered first preburst coordinate $f_{r_k}(U_{k-1}) + o(1)$. The second is exactly $V_{k-1} - \beta \log r_k$.

Centering is the subtraction of $(\log t_k/a, \beta \log t_k)$, a vector perpendicular to z_2. Therefore, it commutes with rare increments. The Jacobian $DG = I - \eta s z_2 z_2^{\top}$, where $s = e^m/(1+e^m)^2 \in (0, 1/4]$,

has eigenvalue one on $z_2^\perp$ and $1 - \eta Qs \in (-1, 1)$ on z_2. Integration along segments shows global nonexpansiveness of G. Applying G^B to the preburst error gives, without any prior bound on v_n,

$$p_k = T_{r_k}(p_{k-1}) + \varepsilon_k, \quad \|\varepsilon_k\| \to 0. \tag{24}$$

Also $DC_R = \mathrm{diag}(f'_R, 1)$ with $0 < f'_R < 1$, so each C_R and T_R is globally nonexpansive.

A.3 Uniform contraction and endpoint forcing

On $K = \{p : \|p - p_*\| \le h\}$,

$$0 < f'_R(U) = \frac{e^{aU}}{e^{aU} + \alpha(R-1)} \le \theta < 1.$$

Since $A < \alpha$, $f_R(U_*) = a^{-1}\log(\alpha + (A-\alpha)/R)$ grows with R. Then $g(R)$ decreases and its absolute value is no greater than the endpoint value M_0. Nonexpansiveness of C_R ensures that its output margin over K changes by no more than $\sqrt{Q}h$. Each rare step increases that margin by a quantity in $(0, \eta Q)$. All intermediate margins entering the B rare derivatives have absolute value at most H, and their curvatures lie in $[s_{\min}, 1/4]$. Thus $0 \le \zeta < 1$.

The rare Jacobians share the eigenspaces $z_2^\perp$ and $\mathrm{span}(z_2)$ that are orthogonal. Their product equals $I - P + \lambda P$ and $|\lambda| \le \zeta$. Setting $D = \mathrm{diag}(f'_R(U), 1)$, we obtain

$$(DT_R)^\top DT_R = D[I - (1-\lambda^2)P]D \preceq D[I - (1-\zeta^2)P]D.$$

To justify the uniform bound, put $J = \mathrm{diag}(f'_R(U)/\theta, 1)$, so $D = D_\theta J = J D_\theta$. The right side is JEJ, whose operator norm is at most $\|J\|^2 \|E\| \le \|E\| = q^2$. This avoids an entrywise comparison of rescaled matrices.

For a nonzero vector x,

$$x^\top E x = \|D_\theta x\|^2 - (1-\zeta^2)\|P D_\theta x\|^2.$$

Equality with $\|x\|^2$ requires $x_1 = 0$ and $Px = 0$. These are mutually exclusive in the case where $c > 0$. The compactness of the unit circle leads to $q < 1$; the square q^2 of the quantity in (18) is the maximal eigenvalue of the symmetric 2×2 matrix. Convexity of K permits integrating the derivative along its segments, proving a uniform q-Lipschitz bound for T_R on K.

Finally,

$$\|T_R(p_*) - p_*\| \le \|C_R(p_*) - C_{R_0}(p_*)\| \le d.$$

On either side of R_0, both $|f_R(U_*) - f_{R_0}(U_*)|$ and $|\log R - \log R_0|$ increase toward the corresponding endpoint. This proves that (19) bounds the entire interval.

A.4 Entry, limiting error, and sign

Let $\gamma = (1-q)h - d > 0$. For $\ell = \|p - p_*\| \ge h$, choose $p_h = p_* + (h/\ell)(p - p_*)$. Applying global nonexpansiveness to p, p_h and local contraction to p_h, p_* gives

$$\|T_R(p) - p_*\| \le \ell - ((1-q)h - d). \tag{25}$$

After a finite time, $\|\varepsilon_k\| < \gamma/2$. As long as an iteration lies outside K, its next distance is shortened by at least $\gamma/2$. However, it will enter the set in finite time because a finite nonnegative distance cannot decrease in this way indefinitely. Inside K, the next distance is at most $qh + d + \gamma/2 = h - \gamma/2 < h$, proving eventual invariance for the perturbed recursion.

Thereafter $\ell_k \le q\ell_{k-1} + d + \|\varepsilon_k\|$, with ℓ_k bounded. Lim sup gives $\limsup \ell_k \le d/(1-q)$. The centered pre-burst approximation together with non-expansiveness of C_R yields

$$|z_2 \cdot w_{t_k - 1} - m_*| \le \sqrt{Q}\big(\|p_{k-1} - p_*\| + d + o(1)\big).$$

Since centering does not contribute to the rare margin, the proof of both estimates in (21) is completed. Under (22), every sufficiently late pre-burst margin has the strict reference sign. Common updates decrease the rare margin, while rare updates increase it. So the positive pre-burst minima ensure positivity for all late times, while negative pre-burst values lead to infinitely many failures. Equation (23) handles the common margin.

For a given $h > 0$ and reference, all displayed constants depend continuously on both endpoints as they converge to R_0. Then $d \to 0$, whereas q converges to a value strictly below one. If $m_* \ne 0$, then both inequalities hold strictly on a sufficiently small nondegenerate interval. Exposure and $R_- > 1$ remain strict. This ends the proof of Theorem 8.

B GEOMETRIC EXTREMA AND BURST-SIZE CONSEQUENCE

For $t_k = \lfloor R^k \rfloor$, the replay ratio tends to R, so the preceding approximation and growth arguments apply. Pick a ball of any fixed radius $h > 0$ centered at the reference point. For the singleton ratio R, local contraction has factor $q_0 < 1$. On this ball $T_{r_k} \to T_R$ uniformly, and $\|T_{r_k}(p_*) - p_*\| \to 0$. Thus, for all sufficiently late ratios, the local contraction constants are bounded by some $q' < 1$, while the forcing tends to zero. The radial entry argument applies with eventual strict slack, and the resulting recurrence gives $p_k \to p_*$. Neither exact ratios nor a prior bound on the second coordinate is needed.

In the common gap following burst k, take $s = n/t_k$. The partial-gap expansion holds throughout this interval:

$$m_n = cV_* - \frac{b}{a} \log\big(A + \alpha(s-1)\big) + o(1).$$

The endpoints of the scaled gap tend to $s = 1$ and $s = R$. By the fixed-point identity, the limiting profile decreases from $m_B + Q\kappa \log R$ to m_B. During the burst, the finitely many margin values converge successively to $F^j(m_B)$, for $j = 1, \ldots, B$; these increase strictly. The pre-burst and post-burst endpoint subsequences attain the extrema. The uniform profile covers all remaining late times, completing the proof of Proposition 1.

For the large-fixed-B consequence, start $m_0 = 0$, $m_{j+1} = F(m_j)$, and $Y_j = e^{m_j}$. Then

$$Y_{j+1} - Y_j \le \eta Q e^{\eta Q}, \qquad S_B(0) \le \frac{1}{Q} \log(1 + \eta Q e^{\eta Q} B) = O(\log B).$$

For fixed $L \in (0, \eta/(2\kappa))$, this will be less than $\kappa B L$ if B is large enough. Take $a_j = \lfloor \rho^j \rfloor$, $\rho = e^L$, and start after adjacent gaps exceed B. In the regular case, every a_j will be used; in the burst case, $a_{Bk}, \ldots, a_{Bk} + B - 1$. The number of updates per block for both schedules is B. The burst goes ahead by at most $B - 1$, and regular replay catches up by the endpoint. The geometric proposition proves the opposite outcomes. This assertion is for a fixed finite B, not a growing-burst limit of the learning process.

C UNIFORM JITTER AND EXACT PREFIX COUNTS

These conditions on the signs of (5) are implied by $e > 8/3$ and $\log x > 2(x-1)/(x+1)$ for $x > 1$, where the latter is implied through differentiation of the difference with the derivative equal to $(x-1)^2/[x(x+1)^2] > 0$. The reference assumptions also hold. Choose $e_0 > 0$ so that the ratio intervals $\exp([L - e_0, L + e_0])$ and $\exp([2L - e_0, 2L + e_0])$ lie inside the respective intervals supplied by Theorem 8. Choose $\delta > 0$ with $2\delta < \min(e_0/2, L/2)$.

For large j, uniformly over $|\xi_j| \le \delta$,

$$\log a_j = jL + \xi_j + e_j, \qquad |e_j| \le 2e^{-jL+\delta}.$$

Indeed $a_j = e^{jL+\xi_j}(1 - v_j)$ with $0 \le v_j < e^{-jL+\delta}$, and $|\log(1 - v_j)| \le 2v_j$ once $v_j \le 1/2$. The regular and paired log-ratio errors are bounded by $2\delta + o(1)$, so both eventually fall into the required intervals. Also,

$$e^{(j+1)L+\xi_{j+1}} - e^{jL+\xi_j} \ge e^{jL-\delta}\big(e^{L-2\delta} - 1\big) \longrightarrow \infty$$

uniformly. Choose one k_0 such that all $a_{j+1} - a_j \ge 3$ and both timing conditions are satisfied after $j = 2k_0$. This choice does not depend on initialization or the sequence. Theorem 8 provides the two outcomes for any common finite initialization. Prior to the shared start, both schedules employ only example 1. Within each block, the ordering $a_{2k} < a_{2k} + 1 < a_{2k+1} < a_{2k+2}$ gives (13) and endpoint equality. Since $\log a_j = jL + O(1)$, regular counts are $\log n/L + O(1)$; bounded discrepancy gives the same for paired counts. This ends the proof of Theorem 4.

D COMMON INVARIANT REGION FOR FIXED BURSTS

We prove Theorem 2 under $0 < \eta \le \eta_0/B$. The separate canonical $B = 1, \eta = 1$ case is proved in Appendix E.

D.1 Response derivative and geometry of the region

Use $\mu = ab$, $\nu = ac^2/b$, $\gamma = \mu + \nu$, $\chi = b/a$, and $k = \nu/\mu = c^2/b^2$. Unlike a one-step parametrization, μ, ν do not include η; the step size is contained in S_B. Set $s_0 = S_B(0)$, $r = e^{\nu s_0}$, $h = r - 1$, $E = e^{\gamma s_0}$, $A = h/(E-1)$, $Y_* = A + h$, and $P = A^\nu Y_*^\mu$.

Since $\eta Q \le 1/B \le 1$, each derivative of F lies in $(0, 1)$ and

$$0 < -S_B'(m) = \frac{1 - \prod_{j=0}^{B-1} F'(F^j(m))}{Q} \le \frac{B\eta}{4}. \tag{26}$$

Indeed $1 - \prod_j (1 - x_j) \le \sum_j x_j$ for $x_j \in [0, 1]$, and here $x_j \le \eta Q/4$. Also $0 < s_0 \le B\eta/2$, with strict inequality for $B > 1$. The step restriction implies

$$\gamma B\eta \le 1, \qquad H := \chi\gamma B\eta = B\eta Q \le 1, \qquad H \le \frac{2k}{1+k}. \tag{27}$$

The set (10) is convex in (U, V): its product constraint is affine after taking logs, and the other is the epigraph of a positive multiple of $\log(e^{aU}/\alpha + h)$. The two boundary branches meet at (A, Y_*). At $R = r$, $Z = Y_*$, the preburst margin is zero and this point is fixed by (9).

D.2 Both constraints are invariant

The rare response cancels from the product:

$$(X')^\nu (Y')^\mu = (X + R - 1)^\nu Y^\mu / R^{\nu+\mu}.$$

The function $\phi(R) = (R-1)/(R^{1+\mu/\nu} - 1)$ decreases for $R \ge 1$, using its continuous value at one, and $A = \phi(r)$. If $X \le A$ then $X + R - 1 \ge XR^{1+\mu/\nu}$ for $R \le r$, preserving the product bound. If $X \ge A$, use $Y \ge X + h$ and monotonicity to reduce to $X = A$. Thus the first constraint is invariant.

For the difference write $Z = X + R - 1$, $m = \chi \log(Y/Z)$, $s = S_B(m)$, $\ell = -S_B'(m)$ and $N = Ye^{\nu s} - Ze^{-\mu s}$. Throughout the region, $X \le Z \le Y$, $m \ge 0$, $s \le s_0$, and $0 < \ell \le B\eta/4$. Holding X, R fixed gives

$$N_Y = e^{\nu s}(1 - \nu\chi\ell) - \mu\chi(Z/Y)e^{-\mu s}\ell \ge 1 - H/4 > 0.$$

The bracket is positive by (27). It suffices to minimize on $Y = \max\{X + h, P^{1/\mu}X^{-k}\}$.

On $Y = X + h$, $m_X = \chi(1/Y - 1/Z) \le 0$, hence $s_X \ge 0$ and

$$N_X = e^{\nu s} - e^{-\mu s} + s_X(\nu Y e^{\nu s} + \mu Z e^{-\mu s}) > 0.$$

On $Y = P^{1/\mu}X^{-k}$, $Y_X = -kY/X$ and $s_X = \chi\ell(1/Z + k/X)$. Therefore

$$N_X \le (Y/X)\{-ke^{\nu s} + \chi(1+k)\ell[\nu e^{\nu s} + \mu e^{-\mu s}]\} - e^{-\mu s},$$
$$e^{-\nu s}\{\cdots\} \le -k + (1+k)H/4 \le -k/2 < 0.$$

Here $1/Z + k/X \le (1+k)/X$ controls the νY term, and $1 + kZ/X \le (1+k)Y/X$ controls the μ term. The bounds hold over the entire noncompact branch. Both branch minima consequently occur at (A, Y_*).

At this corner set $t = Y_*/Z \in [1, E]$ and denote the difference numerator by $N_0(R)$. Then

$$N_0'(R) = e^{-\mu s}\{\chi\ell[\nu t e^{\gamma s} + \mu] - 1\} < 0,$$

since the positive term in braces is at most

$$(H/4)e^{2\gamma s_0} \le e/4 < 1.$$

As $N_0(r) = rh$, we obtain $N \ge N_0(R) \ge rh \ge Rh$, proving $Y' - X' \ge h$. Both constraints are invariant for every $R \in [1, r]$.

D.3 Discrete transfer, attraction, and all-time margins

The fixed-B growth and gap expansion in Appendix A hold before any transverse boundedness assumption. They give

$$p_k = T_{R_k}(p_{k-1}) + \varepsilon_k, \qquad T_R = G^B \circ C_R, \quad \|\varepsilon_k\| \to 0.$$

Growth bounds U. Every late common step adds $\alpha + O(1/n)$ to e^{au_n} and rare steps reduce it, so $e^{au_n} \le \alpha n + O(\log n)$ and $\limsup X_k \le 1$.

The V increment is $-\beta \log R + cS_B(m) + o(1)$. Uniformly for bounded U and admissible R, it tends to a negative value at large positive V. At large negative V its limiting lower bound is

$$cB\eta - \beta \log R_+ > c(B\eta - s_0) \ge cB\eta/2 > 0.$$

Strict inward drift and bounded increments place the exact centered tail in a compact rectangle.

Take a compact ball containing that tail, its Euclidean projections onto K, and their connecting segments. On it $DC_R = \operatorname{diag}(f_R', 1)$ strictly contracts the horizontal direction. Each rare Jacobian is nonexpansive and strictly contracts the z_2 direction; so does their product, whose eigenspaces are fixed. A vector preserving norm through both parts would be vertical and perpendicular to z_2, hence zero. Compactness over states, ratios and unit vectors gives a joint Lipschitz constant $\lambda < 1$. Invariance and projection imply

$$\operatorname{dist}(p_k, K) \le \lambda \operatorname{dist}(p_{k-1}, K) + \|\varepsilon_k\| \longrightarrow 0.$$

Thus $Y_k - X_k \ge h - o(1)$. The preburst margin equals

$$\chi \log \frac{Y_{k-1}}{X_{k-1} + R_k - 1} + o(1).$$

For every sufficiently small $\epsilon > 0$, eventually $X_{k-1} \le 1 + \epsilon$ and $Y_{k-1} \ge X_{k-1} + r - 1 - \epsilon$. Since $R_+ < r$, the middle fraction below decreases with X_{k-1} when $\epsilon < r - R_+$, and

$$\frac{Y_{k-1}}{X_{k-1} + R_k - 1} \ge \frac{X_{k-1} + r - 1 - \epsilon}{X_{k-1} + R_+ - 1} \ge \frac{r}{R_+ + \epsilon}.$$

Letting $\epsilon \downarrow 0$ gives the preburst liminf at least $\chi \log(r/R_+)$, proving (7). Common steps decrease the rare margin and rare steps increase it, so the same bound holds at every sufficiently late integer time. The common margin is $\log n + O(1) > 0$. Finite prefixes change only the finite starting state.

Finally $s_0 < B\eta$. Given $C > r$, select $r < R < \min\{C, e^{B\eta/\kappa}\}$. Proposition 1 gives a strictly negative limiting preburst margin for geometric bursts, since $S_B(0) < \kappa \log R < B\eta$. This proves the worst-case converse, without a necessity claim for every varying schedule.

E Canonical unit-step invariance beyond the small-step range

All coordinates in this appendix are canonical: $a = b = c = \eta = B = 1$. Set $r_c = \sqrt{e}$, $h_c = r_c - 1$, $A_c = 1/(r_c + 1)$, and $P_c = r_c^2/(r_c + 1)^2 = A_c(A_c + h_c)$. For $X = e^U$, $Y = e^V$, we have

$$Z = X + R - 1, \qquad d = \frac{Z}{Z + Y}, \qquad X' = \frac{Z}{R} e^{-d}, \qquad Y' = \frac{Y}{R} e^{d}. \tag{28}$$

which are precisely the exponential coordinates of T_R. The region (10) is equivalent to $Y \ge \max\{P_c/X, X + h_c\}$; the boundary curves meet at $X = A_c$. Convexity follows because $V \ge \log(e^U + h_c)$ is a convex epigraph and the other constraint is a halfspace.

E.1 A common invariant region

We prove invariance for $1 \le R \le r_c$. First $X'Y' = Y(X + R - 1)/R^2$. When $X \le A_c$, use $Y \ge P_c/X$ and

$$X + R - 1 - XR^2 = (R - 1)\big(1 - X(R + 1)\big) \ge 0.$$

When $X \ge A_c$, use $Y \ge X + h_c$ and monotonicity of $(X + h_c)(X + R - 1)$ in X, followed by the same inequality at A_c. Both cases give $X'Y' \ge P_c$.

For the second constraint take $N = Ye^d - Ze^{-d}$, whence $Y' - X' = N/R$. Throughout the region $Y \geq X + h_c \geq Z$ and $0 < d \leq 1/2$. Taking X, R to be fixed results in

$$N_Y = e^d\big(1 - d(1-d)\big) - e^{-d}d^2 > 0.$$

One only needs to find the minimum of N at the lower boundary. On $Y = X + h_c$, $X \geq A_c$, differentiation along the boundary gives

$$d_X = \frac{h_c - (R-1)}{(Z+Y)^2} \geq 0, \qquad N_X = e^d - e^{-d} + d_X(Ye^d + Ze^{-d}) > 0.$$

So the minimum of this function at this branch occurs at A_c. On $Y = P_c/X$, $0 < X \leq A_c$, we have $X \leq Z \leq Y$ and

$$d_X = \frac{Y(X+Z)}{X(Z+Y)^2}.$$

Consequently

$$\begin{aligned} N_X &= \frac{Y}{X}e^d\left(-1 + \frac{Y(X+Z)}{(Z+Y)^2}\right) + e^{-d}\left(-1 + \frac{YZ(X+Z)}{X(Z+Y)^2}\right) \\ &\leq -\frac{Y}{2X}e^d - e^{-d} + \frac{Y}{2X}e^{-d} < 0. \end{aligned}$$

Here $Y(X+Z) \leq 2YZ \leq (Z+Y)^2/2$, $YZ/(Z+Y)^2 \leq 1/4$, and $X + Z \leq 2Y$. Hence the minimum on this branch is also at A_c.

Write $N_0(R) = N(A_c, A_c + h_c, R)$. Direct differentiation gives

$$N_0'(R) = e^{-d}\left((1-d)^2e^{2d} + d(1-d) - 1\right) \leq \frac{1}{4},$$

using $(1-d)e^d \leq 1$ and $d(1-d) \leq 1/4$. At $R = r_c$ we have $Z = Y$, $d = 1/2$, and $N_0(r_c) = r_c h_c$. Therefore

$$N_0(R) \geq r_c h_c - \tfrac{1}{4}(r_c - R) \geq Rh_c,$$

since $h_c > 1/4$. This proves $Y' - X' \geq h_c$ and completes invariance.

The discrete approximation, compactness, contraction, and margin argument for Appendix D holds the same way for $a = b = c = \eta = 1$, where $\eta Q = 2 < 8$, $\chi = \mu = \nu = 1$, and (10) corresponds to the region just proved. This yields (7) with $R_c = \sqrt{e}$. The same appendix proves the geometric worst-case converse. Thus the canonical $B = 1$ case extends the admissible step range without changing the threshold formula.

F Proofs for the coupled multi-example family

All constants in this appendix concern (11). Write $Q = Q_*$, $\kappa = \kappa_*$, $z = z_*$, $\beta = (\sum_i b_i/a_i)/c$, $\alpha_i = \eta a_i^2/d$, and $S_B = S_{B,*}$. Common coordinates are $(u_1, \ldots, u_d)$ and the final coordinate is v. Orthogonal components in a larger ambient space are invariant and irrelevant to these margins.

F.1 Centered maps and geometric reference

For $R > 1$, define

$$\begin{aligned} f_{i,R}(U) &= \frac{1}{a_i}\log\left(\frac{e^{a_i U}}{R} + \alpha_i(1 - R^{-1})\right), \\ C_R(p) &= (f_{1,R}(U_1), \ldots, f_{d,R}(U_d), V - \beta\log R), \\ G(p) &= p + \frac{\eta z}{1 + e^{z\cdot p}}, \qquad T_R = G^B \circ C_R. \end{aligned} \tag{29}$$

The scalar response has the strict monotonicity and limits proved in Appendix A. Set $D_* = \kappa \log R_0$ and let $S_B(m_*) = D_*$. Define

$$A_i = \frac{\alpha_i(R_0 - 1)}{R_0 e^{a_i b_i D_*} - 1}, \quad U_{i,*} = \frac{\log A_i}{a_i}, \quad V_* = \frac{m_* + \sum_i b_i U_{i,*} + QD_*}{c}. \tag{30}$$

Then $0 < A_i < \alpha_i$ and $f_{i,R_0}(U_{i,*}) = U_{i,*} + b_i D_*$. Since $\beta \log R_0 = cD_*$, the preburst margin of $C_{R_0}(p_*)$ is m_*. The burst displacement is $D_* z$, giving $T_{R_0}(p_*) = p_*$.

F.2 Explicit invariant-ball constants

For the h, R_-, R_+ in Theorem 3, put

$$\begin{aligned}
\theta_i &= \frac{e^{a_i(U_{i,*}+h)}}{e^{a_i(U_{i,*}+h)} + \alpha_i(R_- - 1)},\\
g(R) &= c(V_* - \beta \log R) - \sum_i b_i f_{i,R}(U_{i,*}),\\
H &= \max\{|g(R_-)|, |g(R_+)|\} + \sqrt{Q}h + B\eta Q,\\
s_{\min} &= \frac{e^H}{(1+e^H)^2}, \qquad \zeta = \max\{|1-\eta Q s_{\min}|, |1-\eta Q/4|\}^B,\\
P &= zz^\top/Q, \quad D_\theta = \operatorname{diag}(\theta_1, \dots, \theta_d, 1),\\
q &= \sqrt{\lambda_{\max}(D_\theta[I-(1-\zeta^2)P]D_\theta)},\\
\Delta &= \max_{R\in\{R_-,R_+\}} \left[\sum_i (f_{i,R}(U_{i,*}) - f_{i,R_0}(U_{i,*}))^2 + \beta^2(\log R - \log R_0)^2\right]^{1/2}.
\end{aligned} \tag{31}$$

These are finite formulas and one scalar root; none uses a realized trajectory.

F.3 Growth and approximation without transverse boundedness

Let $y_i = e^{a_i u_i}$. A common i-update adds

$$y_i\left(e^{\eta a_i^2/(1+y_i)} - 1\right) \le \eta a_i^2 e^{\eta a_i^2};$$

other common updates leave y_i unchanged, and a rare update multiplies it by at least $e^{-\eta a_i b_i}$. Hence $y_i \le C_i n$ and $u_i \ge -b_i \eta B k + O(1)$ before burst k. There are $O(\log t_k)$ previous rare updates. In the common gap of order t_k before burst k, each example appears a fraction $1/d$ of the time up to an $O(1)$ count error. At most $O(k)$ of its updates raise u_i to zero if necessary; once $y_i \ge 1$, each adds at least $\eta a_i^2/2$. The remaining order-t_k updates give a lower bound proportional to t_k, which loses at most the fixed factor $e^{-B\eta a_i b_i}$ during the burst. Bounded successive time ratios then imply

$$y_{i,n} = \Theta(n), \qquad a_i u_{i,n} = \log n + O(1)$$

at all late integer times, without any bound on v.

Each late i-update adds $\eta a_i^2 + O(1/n)$ to y_i. Between the postburst time $t_{k-1} + B - 1$ and preburst time $t_k - 1$, there are exactly $t_k - t_{k-1} - B$ common updates. The resumed cycle allocates $(t_k - t_{k-1} - B)/d + O(1)$ to each coordinate. Summing errors over the bounded-ratio gap therefore yields

$$y_{i,t_k-1} = y_{i,t_{k-1}+B-1} + \alpha_i(t_k - t_{k-1}) + O(1).$$

After division by t_k, the leading expression is at least $\alpha_i(1 - R_-^{-1}) > 0$, so taking logs is legitimate. Centering by $\ell = (1/a_1, \dots, 1/a_d, \beta)$ commutes with every rare step because $z \cdot \ell = 0$. Finally $DG = I - \eta\sigma zz^\top$, $0 < \sigma \le 1/4$, is globally nonexpansive when $\eta Q < 8$. For

$$p_k = w_{t_k+B-1} - \ell \log t_k$$

we obtain $p_k = T_{R_k}(p_{k-1}) + \varepsilon_k$, with $\|\varepsilon_k\| \to 0$.

F.4 Contraction, global entry, and sign

Each C_R is globally nonexpansive. On the ball $\|p - p_*\| \le h$, its common derivatives are bounded by $\theta_i < 1$. As $A_i < \alpha_i$, each $f_{i,R}(U_{i,*})$ increases with R, and $g(R)$ decreases. Nonexpansiveness and the bound $B\eta Q$ on the total rare-margin increase give the curvature lower bound $s_{\min}$ for every rare Jacobian on the ball's image. Those Jacobians share the projection P, so their product is $I - P + \lambda P$ with $|\lambda| \le \zeta < 1$.

Let $D = \operatorname{diag}(f'_{1,R}, \dots, f'_{d,R}, 1) = D_\theta J = J D_\theta$, where $\|J\| \le 1$. Then

$$(DT_R)^\top DT_R \preceq J D_\theta [I - (1-\zeta^2)P] D_\theta J.$$

Its norm is at most q^2. Norm equality at one in the middle matrix would require all common coordinates of a vector to vanish and the remaining vector to be perpendicular to z. Since $c > 0$, this vector must be zero, proving $q < 1$. Integrating derivatives over segments in the ball gives q-contraction. Endpoint monotonicity of all common shifts and the log shift gives $\|T_R(p_*) - p_*\| \le \Delta$.

For $\rho = \|p - p_*\| \ge h$, radial projection onto the ball and global nonexpansiveness imply

$$\|T_R(p) - p_*\| \le \rho - [(1-q)h - \Delta].$$

The positive slack in (12) eventually exceeds twice the approximation error. Every finite state therefore enters the ball in finite time and stays there. The limsup recurrence yields

$$\limsup_k \|p_k - p_*\| \le \frac{\Delta}{1-q}, \qquad \limsup_k |z \cdot w_{t_k - 1} - m_*| \le \sqrt{Q}\left(\frac{\Delta}{1-q} + \Delta\right). \tag{32}$$

The second bound follows by comparing $C_{R_k}(p_{k-1})$ with $C_{R_0}(p_*)$. Common steps decrease the rare margin, while rare steps increase it. Consequently the preburst margins control all-time retention; common margins already grow as $\log n + O(1)$. The proof is global in initialization despite using a local invariant ball.

When both endpoints equal R_0, $\Delta = 0$ and $q < 1$ for each fixed $h > 0$. Continuity gives a nondegenerate interval meeting both strict tests whenever $m_* \neq 0$. For a geometric schedule, $R_k \to R_0$. Choose a small interval satisfying the first test, obtain a bounded tail, and write the recursion as $p_k = T_{R_0}(p_{k-1}) + o(1)$ by uniform continuity on that tail. The same radial argument with zero forcing gives $p_k \to p_*$. Preburst and postburst margins thus converge to m_* and $m_* + QD_*$, respectively. Their monotonicity between endpoints proves the asserted liminf and limsup, including the equality-root case without deciding its eventual sign. Finite prefixes and common-cycle phases alter only bounded count errors and the finite initial state.

AI USAGE DISCLOSURE

Generative AI tools were used only to polish the writing for readability.
The authors reviewed all edits and take responsibility for the final content.