MEMORY RENORMALIZES STABILIZER ANNEALING IN ADAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Adam's memory parameters and denominator stabilizer both shape its dynamics, raising a natural question: when can a memory change be compensated by retuning the stabilizer, and what can such compensation preserve? For deterministic full-batch Adam on separable linear data, with sufficiently small fixed positive steps and an exponentially annealed outside-root stabilizer with polynomial modulation, the two moment filters renormalize the schedule into an effective rate r and shape Λ. These coordinates determine all divergent terms of the parameter trajectory, including iterated-logarithmic corrections on degenerate margin geometries. Matching (r, Λ) is necessary and sufficient for the trajectory difference to converge to a finite vector under the prescribed effective clocks, while rate matching alone can leave logarithmic drift. Yet trajectory equivalence need not imply relative-risk equivalence: the residual finite offset can yield a nonunit asymptotic risk ratio. Preserving relative risk at equal effective time additionally requires an amplitude coordinate, whereas equal-update risk is governed by a distinct complete invariant; the two preservation targets can be incompatible. Experiments on actual discrete Adam iterates support these separating predictions across regular, degenerate, and native-softmax geometries. Together, these results give a clock- and observable-dependent classification of equivalence under Adam memory changes.

1 INTRODUCTION

Adam's long-time behavior depends on its stabilizer and step regime. On separable classification problems, a positive-stabilizer model can recover Euclidean max-margin behavior (Wang et al., 2022), whereas unstabilized Adam under diminishing steps can select an ℓ_∞-margin solution (Zhang et al., 2024). These results concern distinct recurrences and limits. Within Adam's update, the moment filters determine the response to a decaying gradient signal, while the stabilizer supplies a competing scale in the denominator. Both therefore affect the normalized update.

This interaction raises a compensation question: can retuning the stabilizer schedule preserve the dynamics when memory changes? Matching the limiting classifier is a natural target, but runs approaching the same direction can still separate in their full parameter vectors. Even a convergent parameter difference can leave a nonunit ratio of vanishing training risks. Equal dynamical time and equal update count can also compare different states. Calibration therefore requires an observable to preserve and a clock on which to compare it.

We study deterministic full-batch Adam (Kingma and Ba, 2015) on separable linear data with both moment filters, sufficiently small fixed positive steps, and an exponentially annealed outside-root stabilizer with polynomial modulation. This two-scale schedule separates two mechanisms: its exponential factor sets the leading equilibrium, while its polynomial factor perturbs the local decay rate by order $1/S$, accumulating a $\log S$ displacement. This annealing regime differs from standard Adam's constant stabilizer. The closest predecessor, memoryless smoothed-sign annealing, identifies a rate-indexed Burg-barrier selector under vanishing steps (Wang et al., 2026). Retaining both filters at a fixed step changes the response to both scales.

Memory renormalizes the raw schedule into $r = k/\sigma$ and $\Lambda = \lambda\chi$, with σ, χ determined by memory, rate and step. These coordinates control the leading motion and logarithmic correction, respectively. RATE matches r; TWO matches both. Figure 1 shows the consequence: RATE leaves logarithmic separation and TWO cancels it.

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

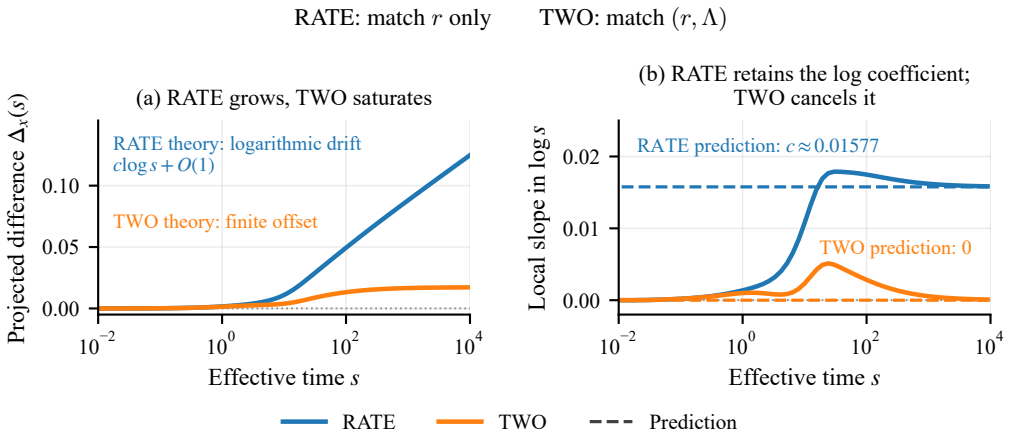

Figure 1: Rate matching alone does not eliminate logarithmic trajectory drift. Left: the signed parameter difference $\Delta_x(s) = \langle v_{\text{pred}}, W_{\text{REF}}(s) - W_x(s) \rangle$, projected onto the unit predicted RATE-drift direction $v_{\text{pred}} = (-0.6, -0.8)$. Right: 63 adjacent log-secants from 64 readouts at geometric midpoints, compared with configuration-based coefficient predictions (dashed). Blue: RATE; orange: TWO. Solid curves are finite-window actual Adam observations on one nonradial instance. Protocol and distinct estimators are given in Section 5.2 and Appendix Q.2.

When the coefficient mismatch has a component orthogonal to the common leading velocity, as in this example, no scalar retiming makes the difference bounded (Corollary 14).

From leading motion to the full trajectory. Removing logarithmic drift is not always enough. On degenerate margin geometries, weak tight rows can generate $\log \log s$ and still slower divergent terms, as higher-order implicit-bias analyses illustrate (Soudry et al., 2018; Nacson et al., 2019). We resolve this finite hierarchy from actual fixed-step Adam updates and prove convergence of the remaining parameter residual. On a fixed learning problem, (r, Λ) determines every divergent coefficient and the hierarchy depth. Matching this pair is therefore necessary and sufficient for full-trajectory equivalence in the prescribed clocks; degenerate geometry adds no independent calibration coordinate within this schedule family.

From trajectory preservation to risk preservation. Even after every divergent term agrees, the finite parameter offset can survive as a multiplicative factor in the vanishing training risk. At equal effective time, preserving a risk ratio of one therefore requires an amplitude coordinate Γ. Equal update count compares different states and leads to a different invariant. The three preservation targets are

$$\begin{array}{ll} \text{Full trajectory, effective time} & (r, \Lambda), \\ \text{Relative risk, effective time} & (r, \Lambda, \Gamma), \\ \text{Relative risk, update count} & (\vartheta, \lambda, K). \end{array}$$

Here $\vartheta = k\eta$, and K is the update-based risk amplitude. AMP adds amplitude matching to TWO; SYNC preserves equal-update risk. These are calibration maps for distinct targets. For some memory changes at nonzero shape, full-trajectory preservation and equal-update risk preservation cannot both hold.

Experiments on actual Adam iterates test memory deformation, trajectory drift and risk preservation, with a further trajectory check on frozen learned representations. Together, these results distinguish full-trajectory equivalence from relative-risk equivalence under the chosen observation clock.

2 Related Work

Implicit bias and memory. Gradient methods on separable data select margin solutions, with geometry and higher-order corrections studied by Soudry et al. (2018); Nacson et al. (2019); Gunasekar et al. (2018); Ji and Telgarsky (2021). In particular, Soudry et al. (2018) already resolve iterated

logarithms with bounded remainder in degenerate cases. Our question is cross-configuration equivalence of full Adam trajectories, including convergence of their difference. Adaptive preconditioning changes selection (Duchi et al., 2011; Qian and Qian, 2019). For Adam, Wang et al. (2022) use a positive under-root stabilizer, a small fixed step and $\beta_2 > \beta_1^4$, while Zhang et al. (2024) remove it and use diminishing steps, $\beta_1 \le \beta_2$ and initially nonzero gradient coordinates to obtain max-ℓ_∞ margin rates. These results provide finite-time directional information under their respective assumptions.

Memoryless annealing (Wang et al., 2026) is closest in schedule and geometry: its Theorem 1 uses exponential loss, diverging cumulative time and square-summable steps, and gives a Burg selector with an $O(S_t^{-1/2})$ normalized-iterate envelope. Our fixed-step, two-filter result instead resolves the full hierarchy and convergent residual for exponential, logistic and native-softmax losses. It does not supply that predecessor's finite-time envelope or a computable common step threshold.

Limits of approximations and extensions. Modified-objective analyses describe finite-step effects (Barrett and Dherin, 2021; Cattaneo and Shigida, 2025). Cattaneo and Shigida (2025) give a memoryless corrected iteration with $O(h^2)$ global error over a fixed horizon $[0, T]$ as $h \to 0$ (Corollary 3.3 under their smoothness and memory-decay assumptions). Our infinite-time, fixed-step comparison instead needs a summable tail remainder after every divergent term is resolved; Appendix K connects the regular-face small-step sensitivity. Nonlinear homogeneous-network results (Lyu and Li, 2020; Wang et al., 2021; Gronich and Vardi, 2026) and per-sample Adam (Baek et al., 2026) address regimes outside our deterministic linear classification.

Calibration and transfer. Learning-rate and batch-size scaling preserve continuous-time or stochastic approximations (Smith et al., 2018; Malladi et al., 2022; Hilton et al., 2022). Search and transfer methods reuse evaluations or parameterizations to improve finite-budget selection (Bergstra and Bengio, 2012; Snoek et al., 2012; Li et al., 2018; Falkner et al., 2018; Swersky et al., 2013; Perrone et al., 2018; Yang et al., 2021). Our maps keep the learning problem fixed and compensate a memory change for a specified observable. Appendix A.2 distinguishes the present fixed-step, long-horizon limit from finite-horizon small-step approximations.

3 Problem Setup

Let $Z \in \mathbb{R}^{n_{\mathrm{row}} \times d}$ be a finite margin matrix with positive weights ω_i and $\gamma_\infty = \max_{\|u\|_\infty \le 1} \min_i z_i^\top u > 0$. Binary rows are $z_i = y_i x_i$; native softmax uses $z_{i,k} = (e_{y_i} - e_k) \otimes x_i$, $k \ne y_i$. We cover the exponential loss and grouped logistic/softmax risk

$$L_{\exp}(w) = \sum_i \omega_i e^{-z_i^\top w}, \qquad R(w) = \sum_i \omega_i \log\left(1 + \sum_{k \ne y_i} e^{-z_{i,k}^\top w}\right).$$

Singleton groups give binary logistic loss. Across runs the feature basis, weights, loss normalization and grouping stay fixed. Softmax trains all coordinates; centering is only an explicitly specified observation.

From finite w_1 and $m_0 = v_0 = 0$, deterministic full-batch Adam uses a fixed $\eta > 0$, $(\beta_1, \beta_2) \in [0, 1)^2$, and $g_t = \nabla R(w_t)$ (or $\nabla L_{\exp}(w_t)$):

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^{\odot 2}, \\ w_{t+1} &= w_t + \eta h_t, \qquad\qquad h_t = -\frac{m_t / C_{1,t}}{\sqrt{v_t / C_{2,t}} + \epsilon_t}. \end{aligned} \tag{1}$$

Operations are coordinatewise; each run consistently uses either $C_{\ell,t} = 1$ or $1 - \beta_\ell^t$. The outside-root stabilizer is

$$S_t = \eta(t-1), \qquad \epsilon_t = \epsilon_0 e^{-k S_t} (1 + S_t / s_0)^\lambda, \qquad \epsilon_0, s_0 > 0. \tag{2}$$

Here k sets exponential decay, λ its polynomial modulation, and ϵ_0, s_0 the amplitude and onset scale. These last two quantities can still affect finite offsets and risk amplitudes after the divergent trajectory has been matched.

Normalizing the gradient and moment scales by the stabilizer exposes the two memory responses. Their exact response to a locally frozen exponential rate gives the factors and effective coordinates

$$\rho_\ell = \beta_\ell e^{\ell k\eta} < 1, \quad A_\ell = \frac{1-\beta_\ell}{1-\rho_\ell}\,(\ell = 1, 2), \quad \sigma = \frac{A_1}{\sqrt{A_2}}, \quad r = \frac{k}{\sigma}, \tag{3}$$

$$d_\sigma = \partial_k \log \sigma, \quad \chi = 1 - k d_\sigma, \quad \Lambda = \lambda\chi, \quad \sigma' = \partial_k \sigma. \tag{4}$$

Derivatives hold memory and step fixed. The ratio σ converts the two filter responses into a velocity scale; r is the resulting effective decay rate. The slower modulation is corrected by χ to give Λ. The same response prescribes the observation clock

$$s = \widehat{S}(S) = \sigma S - \lambda\sigma' \log(1 + S/s_0). \tag{5}$$

The local rate is $k - \lambda/(S + s_0)$. Expanding its filter response $\sigma(k - \lambda/(S + s_0))$ to first order and integrating gives this clock; the proof below controls the actual-update remainder separately. Thus the clock is prescribed by the configuration. Its onset convention also fixes the risk amplitude Γ introduced below. On its positive increasing tail, $W(s)$ is the piecewise-linear interpolant of the actual nodes $(\widehat{S}(S_t), w_t)$, not a fitted time warp. Risk is evaluated on interpolated parameters, not interpolated scalar losses. After n updates the node is w_{n+1} at $S = n\eta$.

We compare bounded full-parameter differences, and relative-risk ratios tending to one either at equal effective time or at equal update count:

$$W_a \sim W_b \iff \sup_{s \geq s_*} \|W_a(s) - W_b(s)\| < \infty \quad \text{for some common late } s_*. \tag{6}$$

An *eligible run* satisfies the data, initialization, loss, bias and schedule conditions above and the data-dependent sufficiently-small-step quantifiers in Theorem 2. Pole legality alone does not establish eligibility.

4 Theory

We first state the complete criterion for comparing two runs. We then derive the memory response and finite trajectory hierarchy that make this criterion possible, before turning to relative risk under two observation clocks.

4.1 Complete trajectory equivalence and calibration

The effective rate and shape completely classify divergent trajectories within the exponential–polynomial schedule family of Eq. (2).

Theorem 1 (Complete calibration and relative drift)**.** *Consider individually eligible runs on the same margin matrix, loss and positive weights. Their steps, memories, bias conventions, finite initializations and schedules may differ. For $D(s) = W_a(s) - W_b(s)$,*

$$(r_a, \Lambda_a) = (r_b, \Lambda_b) \iff D(s) = O(1) \iff D(s) \text{ converges to a finite vector.} \tag{7}$$

Different effective rates produce nonzero linear leading separation; equal rates with different effective shapes produce nonzero logarithmic leading separation. At matched coordinates the limiting finite offset need not be zero.

Matching both coordinates cancels every divergent term of the full parameter difference. Additional scales caused by degenerate geometry introduce no further independent calibration coordinates in this family.

The two-coordinate map (TWO) takes source a to target b by

$$\frac{k_b}{\sigma_b(k_b)} = \frac{k_a}{\sigma_a(k_a)}, \qquad \lambda_b = \lambda_a \frac{\chi_a(k_a)}{\chi_b(k_b)}. \tag{8}$$

Each factor uses its run's own memory and fixed step. On the common legal small-step interval, $\partial_k(k/\sigma) = \chi/\sigma > 0$, so one scalar solve and an explicit shape correction suffice. Local existence, uniqueness and eligibility hold under the small-step quantifiers of Theorem 11. RATE solves only the first equation and retains the source raw shape.

4.2 MEMORY RESPONSE AND THE FINITE TRAJECTORY HIERARCHY

The leading selector and its Hessian are

$$u_r = \operatorname*{arg\,min}_{Zu \geq r\mathbf{1},\ |u|<\mathbf{1}} B(u), \qquad B(u) = \sum_j [-|u_j| - \log(1-|u_j|)], \\ H_r = \operatorname{Diag}(1-|u_r|)^{-2}. \tag{9}$$

Let $p(r)$ be the KL-selected primary dual multiplier (base weights ω), J the union of positive supports of all optimal primary multipliers, and $B_0 = \{i : z_i^\top u_r = r,\ p_i(r) = 0\}$ the weak tight set. Redundant weak rows are allowed. Essential and weak rows are both tight at leading order, but they impose different next-order constraints. Keeping only the tight set would lose this distinction. The logarithmic coefficient is selected by

$$c_{r,\Lambda} = \operatorname*{arg\,min}_c \tfrac{1}{2} c^\top H_r c, \qquad Z_J c = -\Lambda\mathbf{1}, \quad Z_{B_0} c \geq (1-\Lambda)\mathbf{1}. \tag{10}$$

Theorem 2 (Actual two-order selection)**.** *Fix finite separable data, positive weights, $\epsilon_0 > 0$, and a beta pair. On compact sets of finite initial parameters, rates $k \in I \Subset (0, \gamma_\infty)$, and profiles (λ, s_0) with $s_0 > 0$, there exists $\eta_0 > 0$ such that every fixed $0 < \eta < \eta_0$, under either bias convention and any of the three losses, satisfies*

$$h_t \to a_* = \sigma u_r, \qquad S_t(h_t - a_*) \to b = c_{r,\Lambda} - \lambda d_\sigma a_*, \tag{11}$$

$$w_t = \widehat{S}(S_t) u_r + c_{r,\Lambda} \log \widehat{S}(S_t) + o(\log \widehat{S}(S_t)). \tag{12}$$

The normalized reference atoms satisfy

$$\xi_{i,t} = \omega_i e^{-z_i^\top w_t}/\epsilon_t \longrightarrow p_i(r)/\sqrt{A_2},$$

with the analogous pair-row definition for softmax. No rank condition, nonzero aggregate-coordinate condition or positive weak-mass lower bound is required.

From rate to shape. Physical velocity is σu_r, whereas effective-time velocity is u_r. Theorem 2 separates two scales: r selects the leading motion, and Λ selects what remains at logarithmic order. Thus rate matching can align the dominant motion while allowing unbounded parameter separation. The raw update coefficient b and effective-clock coefficient $c_{r,\Lambda}$ differ by a radial memory term; changing clocks changes the full-vector coefficient even when normalization suppresses that term. On a regular face, $c_{r,\Lambda} = -\Lambda u_r'$.

Why the expansion must continue. On a degenerate face, the logarithmic program can itself leave weak tight rows. Their next-order response generates $\log\log s$ and potentially slower iterated logarithms. Resolving only the first two terms would then leave an unbounded residual. Define $L_1(s) = \log s$ and $L_{j+1}(s) = \log L_j(s)$ on a late tail.

Theorem 3 (Finite hierarchy and convergent residual)**.** *For each eligible fixed step and profile, the actual interpolated trajectory has a finite expansion*

$$W(s) = s u_r + c_{r,\Lambda} L_1(s) + \sum_{q=3}^{Q} d^{(q)} L_{q-1}(s) + w_\infty + o(1), \qquad Q < \infty. \tag{13}$$

On the same learning problem, (r, Λ) fixes every divergent selector, the entire essential/weak recursion and its depth. The finite w_∞ may depend on the trajectory, initialization and amplitude. Each nonempty recursive Hessian program is strictly feasible and adds independent constraints; its construction, termination bound and update-level limits are given in Appendix I.1.

The finite hierarchy rules out slower hidden drift, allowing the two-coordinate match to yield a convergent parameter difference. Each stage uses the original Hessian H_r, adds its selected essential constraints and continues on the remaining weak rows. If the secondary program leaves no weak tight inequalities, the hierarchy stops at $Q = 2$ even on a degenerate primary face. Appendix I.2 gives a critical two-row example in which the $\log\log s$ term is necessary for a convergent residual.

Proof idea. Exact normalization by the stabilizer retains both filters and the true logistic/softmax denominators. A coupled KL–RMS argument proves entry and non-exit from finite parameters and zero moments. Eliminating the stable positive subsystem leaves multiplicative weak-mass dynamics in logarithmic time. A forced Hessian program selects the next coefficient; repeating the reduction adds independent equalities and terminates. The resulting update remainder is summable along the actual fixed positive-step iterates, so the terminal parameter residual converges to a finite offset (Appendices B–I).

Relative drift. The expansion makes the separation rates in Theorem 1 explicit:

$$D(s)/s \to u_{r_a} - u_{r_b} \neq 0 \qquad (r_a \neq r_b),$$
$$D(s)/\log s \to c_{r,\Lambda_a} - c_{r,\Lambda_b} \neq 0 \quad (r_a = r_b = r,\ \Lambda_a \neq \Lambda_b).$$

At matched coordinates, $D(s) \to d_\infty = w_{\infty,a} - w_{\infty,b}$. Matching (r, Λ) fixes every divergent term, leaving this difference of finite offsets. Conversely, the leading selector identifies r through $\min_i z_i^\top u_r = r$, and any essential row identifies Λ through $z_j^\top c = -\Lambda$. A coordinate mismatch therefore has a nonzero leading coefficient. The exact angular limit and interpolation proof are in Appendix J.

An explicit witness makes the role of shape concrete. For the row $z = (1/3, 1)$ at $r = 7/12$, compare zero memory with (β, β^2), $0 < \beta < 1$, at a common sufficiently small fixed step. Matching rates while retaining a common raw $\lambda_0 \neq 0$ gives

$$D(s) = -\lambda_0(1-\chi)(3/5, 4/5)\log s + O(1), \qquad 1/2 < \chi < 1.$$

The nonzero coefficient remains despite matched leading motion. Correcting the target shape to λ_0/χ removes it (Proposition 13). This is the example used in the trajectory experiment below.

For common r the offset-free diagnostic connects the classification to finite-window observations:

$$D_\gamma(s) = \frac{D(\gamma s) - D(s)}{\log \gamma} \longrightarrow c_{r,\Lambda_a} - c_{r,\Lambda_b}, \qquad \gamma > 1. \tag{14}$$

For fixed $\gamma > 1$, the multiplicative window cancels the unknown finite offset and isolates the logarithmic mismatch. Its limit is nonzero under shape mismatch and zero under matched coordinates; a single endpoint gap does not distinguish them.

4.3 Relative risk under two observation clocks

Matching all divergent terms leaves a finite parameter offset, $W_a(s) - W_b(s) \to d_\infty$. For a vanishing training risk, this offset can remain visible in the risk ratio. Relative-risk equivalence requires that ratio to tend to one, so we must also resolve the remaining amplitude.

Let $\mu(r) = \mathbf{1}^\top p(r) > 0$ be the mass of the KL-selected primary multiplier. Normalized-atom convergence yields the following risk law for all three losses:

$$\Gamma_x = \frac{\epsilon_{0,x}}{\sqrt{A_{2,x}}(\sigma_x s_{0,x})^{\Lambda_x}}, \tag{15}$$

$$R(W_x(s)) = \mu(r_x)\Gamma_x e^{-r_x s} s^{\Lambda_x}(1 + o(1)). \tag{16}$$

Here R denotes the corresponding training loss, including EXP. The rate r_x controls exponential decay, Λ_x the polynomial correction, and Γ_x the remaining multiplicative scale. The atom limit in Theorem 2 supplies the positive mass $\mu(r)$. Comparing the three asymptotic orders gives the complete criterion.

Theorem 4 (Effective-time relative risk)**.** *For individually eligible runs on the same learning problem in their prescribed clocks, $R(W_a(s))/R(W_b(s)) \to 1$ if and only if $(r_a, \Lambda_a, \Gamma_a) = (r_b, \Lambda_b, \Gamma_b)$. At matched (r, Λ) the ratio tends to Γ_a/Γ_b.*

The amplitude Γ is a risk coordinate, not a third divergent trajectory coordinate. AMP applies TWO and the unique stabilizer amplitude correction in Corollary 19. Amplitude matching gives $Z_J d_\infty = 0$; eliminating the full offset also requires the rank condition in Theorem 20 (Appendix N.3). The amplitude is anchored to the prescribed clock origin, so a fitted time shift would change the comparison.

Equal update count. A different question is whether the risks agree after the same number of optimizer updates. Substituting the physical clock $S = n\eta$ changes which exponential rate and polynomial exponent must match. After n completed updates the relevant state is w_{n+1}, with

$$\vartheta_x = k_x \eta_x, \qquad K_x = \frac{\mu(r_x)\epsilon_{0,x}}{\sqrt{A_{2,x}}} (\eta_x / s_{0,x})^{\lambda_x}, \tag{17}$$

$$R(w_{x,n+1}) = K_x e^{-\vartheta_x n} n^{\lambda_x} (1 + o(1)). \tag{18}$$

The complete equal-update criterion is

$$\frac{R(w_{a,n+1})}{R(w_{b,n+1})} \to 1 \quad \Longleftrightarrow \quad (\vartheta_a, \lambda_a, K_a) = (\vartheta_b, \lambda_b, K_b)$$

(Theorem 22). SYNC preserves this triple and r by changing the learning step and schedule, generally changing Λ. Its full map and joint eligibility for finite memory collections at small ϑ are in Appendix P.

Incompatibility. For an eligible memoryless source and an eligible target (β, β^2), $0 < \beta < 1$, a common nonzero Λ cannot preserve both complete trajectory equivalence and equal-update relative-risk equivalence under the prescribed clocks (Corollary 23). Indeed the latter fixes a common ϑ and raw λ, but

$$\chi_a = 1, \qquad \chi_b = 1 - \frac{\vartheta\rho}{1-\rho^2} < 1, \qquad \rho = \beta e^{\vartheta} < 1.$$

Equal-update equivalence fixes a common raw λ; trajectory equivalence fixes the nonzero product $\lambda\chi$. Both requirements would force the two χ values to agree, contradicting the displayed inequality. Thus the appropriate calibration depends on the declared observable and observation clock.

5 Experiments

We test three finite-window predictions on actual discrete Adam iterates: fixed-step memory deformation, the logarithmic drift left by rate-only matching, and relative-risk preservation under two observation clocks. Protocols appear in Appendices L, M and Q; code and supplementary records are in the anonymous repository.

5.1 Fixed-step memory deformation

Changing memory shifts the physical velocity from u_k to $\sigma u_{k/\sigma}$. We test this response at eight actual steps on each G1/G2/G0 $\times$ C/M curve, for 48 deterministic configurations: two endpoint and six interior steps per curve. The observed $\widehat{a}$ is the mean current-query velocity over $540 \le S \le 600$. We measure deformation in the frozen geometry-specific norm induced by $H_{\mathrm{geom}} = \nabla^2 B(u_k)$. The 32 nonzero cells have deformations of 1.34×10^{-3}–6.77×10^{-2}, with prediction residual at most 3.67×10^{-14}; 16 G0 controls retain their measured near-zero values. Figure 2 tests the response across the prescribed step intervals. C/M are compound groups differing in memory, bias convention and step range.

5.2 Rate-only drift and two-coordinate matching

The motivating example in Figure 1 uses $z = (1/3, 1)$ at $r = 7/12$. RATE matches the effective rate while retaining the source raw shape; TWO also matches the effective shape. Figure 1 shows the resulting parameter difference and its adjacent log-secants against effective time. Writing $D_x = W_{\mathrm{REF}} - W_x$ and $\Delta c_x = c_{\mathrm{REF}} - c_x$, both panels project onto the unit predicted RATE drift, $v_{\mathrm{pred}} = \Delta c_{\mathrm{RATE}} / \|\Delta c_{\mathrm{RATE}}\|$.

At $s = 5000$, the fixed-dilation readout in Eq. (14) has norm 0.015863 for RATE against prediction 0.015770; TWO gives 1.11×10^{-4} against zero. These readouts complement the adjacent secants in the figure: TWO cancels the drift while allowing a finite offset. The separate tail regression is reported in Appendix Q.2. Figure 1 projects onto the full predicted RATE-drift direction, whereas the transverse

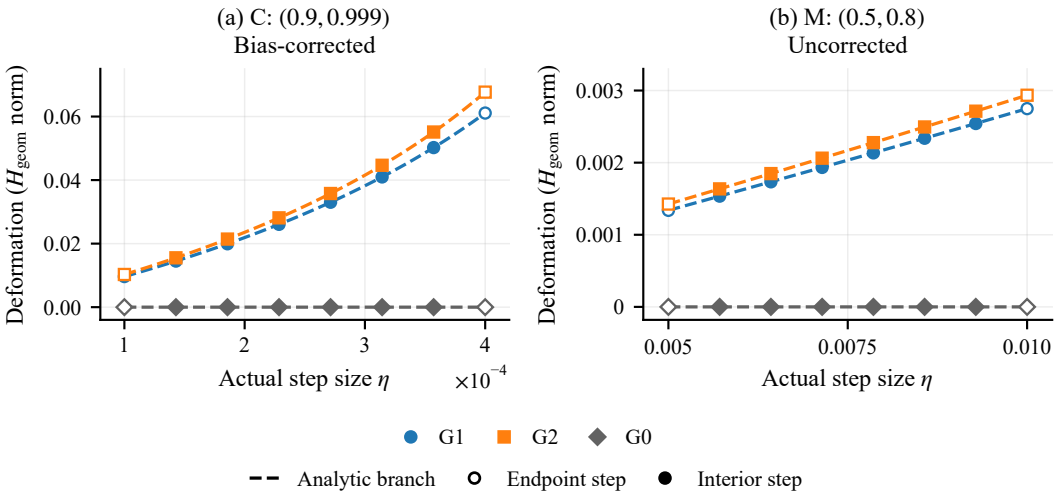

Figure 2: Fixed-step memory response across eight actual steps per geometry–group curve. G1/G2 are nonradial geometries; G0 is the coordinate-row zero-deformation control (Appendix Q.1). Dashed lines show the analytic branch $\|\sigma u_{k/\sigma} - u_k\|_{H_{\text{geom}}}$; markers show $\|\widehat{a} - u_k\|_{H_{\text{geom}}}$, using the uniform dense-tail mean over $540 \le S \le 600$. Panels (a,b) use $(0.9, 0.999)$ with bias correction and $(0.5, 0.8)$ without it, respectively, with different step intervals and linear vertical scales. Each geometry uses its own frozen metric. Hollow markers denote endpoint steps; filled markers denote interior steps. Across 32 G1/G2 cells, $\|\widehat{a} - \sigma u_{k/\sigma}\|_{H_{\text{geom}}} \le 3.67 \times 10^{-14}$; the 16 G0 controls retain their measured near-zero values. Residuals are distinct from the deformation plotted.

readout projects orthogonally to the common leading velocity and isolates the component that survives scalar retiming. A transverse readout of the same trajectories gives $\|\Pi_u G_{\text{RATE}}(5000)\| = 2.83 \times 10^{-3}$ against 2.82×10^{-3} predicted; TWO gives 2.11×10^{-5} against zero. This finite-window readout tests the component in Corollary 14. The complete fits, fixed-$\gamma = 2$ diagnostic, observation windows and linear-clock ablation are in Appendix Q.2.

The critical two-row experiment in Appendix Q.4 exhibits the predicted weak-mass normalization, update coefficient and log-log parameter secant under nonzero memory in a pilot-informed visibility cohort. The default cohort remains unresolved, and nearby side controls remain pre-asymptotic over the evaluated horizon.

5.3 Relative-risk preservation under two clocks

AMP targets relative risk at equal effective time, whereas SYNC targets relative risk at equal update count. Figure 3 compares their parameter drift and both risk targets on the same nonradial P0 instance. With $D_x(s) = W_{\text{REF}}(s) - W_x(s)$, the fixed-dilation readout is $G_{x,2}(s) = [D_x(2s) - D_x(s)]/\log 2$ (Eq. (14)). The signed target/source log-risk ratios are

$$\ell_x^{(s)}(s) = \log \frac{R(W_x(s))}{R(W_{\text{REF}}(s))}, \qquad \ell_x^{(n)}(n) = \log \frac{R(w_{x,n+1})}{R(w_{\text{REF},n+1})}.$$

A zero log-risk ratio means a unit risk ratio and source-risk matching; a negative value means lower target risk, not better calibration. Zero log-drift does not require a zero finite parameter offset. Panels (a,b) use effective time; (c) uses updates.

AMP suppresses the trajectory-drift coefficient and effective-time risk discrepancy, while SYNC brings equal-update risk close to the source. At 64k updates the signed target/source log-risk ratios are -0.529159 for AMP and 1.82×10^{-5} for SYNC. Their different behavior is consistent with the clock-dependent criteria: AMP's equal-update discrepancy can contain both n and $\log n$ terms, while SYNC generally retains a $\log s$ term at common effective time.

The study covers four geometries, including a weak-row instance and native softmax, at 4k/16k/64k updates. It evaluates the memory change $(0, 0) \to (0.6, 0.36)$ (P0) and a nonzero-memory change

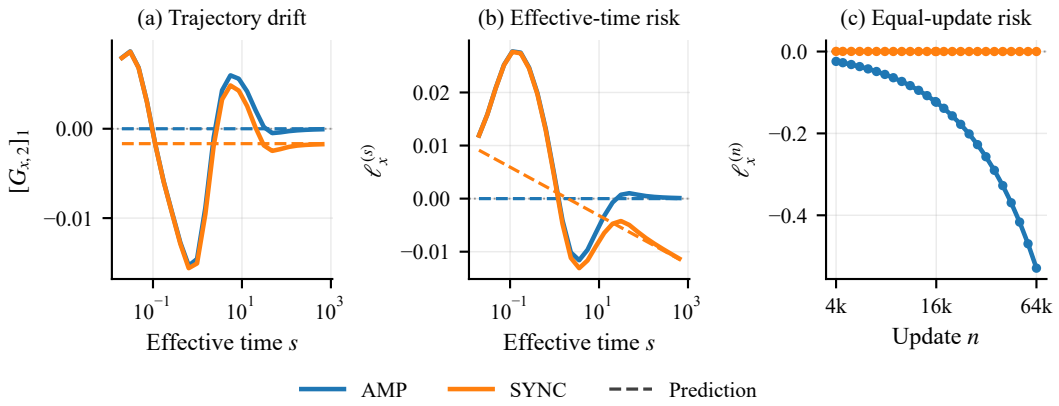

Figure 3: AMP and SYNC on the same frozen nonradial P0 instance. (a) First coordinate of $G_{x,2}$, using reference-minus-target parameters. (b,c) Signed target/source log-risk ratios $\ell_x^{(s)}$ and $\ell_x^{(n)}$. In (b,c), zero log-risk ratio indicates source-risk matching; negative values indicate lower target risk. In (a), zero refers to the displayed log-drift, not to a zero finite parameter offset. All horizontal axes are logarithmic. Blue: AMP; orange: SYNC. Solid curves are observations; dashed curves are configuration-based asymptotic predictions. Panels (a,b) query at effective time s, with (a) also requiring $2s$. Panel (c) uses exact shared saved updates between 4k and 64k; circles mark all 26 observed checkpoints per arm. Protocols and summaries are in Appendix Q.3.

$(0.5, 0.8) \to (0.7, 0.9)$ in both directions (P1/P2). Across the four geometries, AMP's median absolute effective-time prediction residual decreases with budget in both directions. The complete six-arm comparison (REF/RAW/RATE/TWO/AMP/SYNC), signed budget readouts, early deviations, weak-row cases and zero-shape control are summarized in Appendix Q.3; complete row-level tables accompany the source records. Appendix N.5 gives a state-dependent finite-window envelope for the effective-time risk ratio.

As an additional finite-window check, deterministic full-batch Adam trains linear softmax heads on two frozen CIFAR-10 representations with eight preselected memory pairs. TWO reduces the frozen trajectory mismatch relative to RATE in all 16 representation–memory cells, with mean attenuation about 0.306%. These cells share the underlying data; the reduction concerns trajectory mismatch, not accuracy or end-to-end feature learning. Appendix M.2 and Figure 4 provide the protocol and precision checks.

6 Discussion and Limitations

The classification covers deterministic full-batch Adam on separable linear problems with the stated exponential–polynomial schedule family and sufficiently small fixed positive steps. Appendix J.7 shows how richer prescribed profiles change the invariant count and can yield bounded but nonconvergent differences. The same framework suggests three natural extensions: identifying finite-horizon entry regimes, incorporating stochastic gradients, and allowing the learned representation to evolve.

7 Conclusion

Adam memory renormalizes stabilizer annealing into two effective coordinates that completely classify the divergent parameter trajectory in the analyzed regime. Matching these coordinates yields convergence up to a finite offset, but relative-risk equivalence also depends on the observation clock. The resulting criteria distinguish trajectory preservation from risk preservation and identify memory changes for which the targets cannot be met simultaneously. Within this regime, optimizer equivalence is determined by the observable and observation clock chosen for preservation, not by raw hyperparameter agreement alone.

AI USE STATEMENT

Generative AI tools were used to provide ideas and feedback during the development of some mathematical arguments, to aid and polish the writing of this manuscript, and for retrieval and discovery, including identifying relevant related work. All mathematical claims and proofs in the paper were reviewed and verified by the authors. The authors take responsibility for the final content of this work.

REFERENCES

Beomhan Baek, Minhak Song, and Chulhee Yun. Implicit bias of per-sample Adam on separable data: Departure from the full-batch regime. In *International Conference on Learning Representations*, pages 120180–120229, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c3596a5990c2cddb0b0de478df7aca65-Abstract-Conference.html.

David G. T. Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021. URL https://research.google/pubs/implicit-gradient-regularization/.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. URL https://jmlr.org/papers/v13/bergstra12a.html.

Matias D. Cattaneo and Boris Shigida. How memory in optimization algorithms implicitly modifies the loss. In *Advances in Neural Information Processing Systems*, volume 38, pages 156059–156096, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e4cc8ab4a64e99f962f36d07a7723d94-Abstract-Conference.html.

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

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1437–1446, 2018. URL https://proceedings.mlr.press/v80/falkner18a.html.

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.16340.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841, 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 17086–17098, 2022. URL https://arxiv.org/abs/2110.00641.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 772–804, 2021. URL https://proceedings.mlr.press/v132/ji21a.html.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. URL https://jmlr.org/papers/v18/16-558.html.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1906.05890.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pages 7697–7711, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/32ac710102f0620d0f28d5d05a44fe08-Abstract.html.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3420–3428, 2019. URL https://proceedings.mlr.press/v89/nacson19b.html.

Javier Peña, Juan Vera, and Luis Zuluaga. An algorithm to compute the Hoffman constant of a system of linear constraints. *arXiv preprint arXiv:1804.08418v1*, 2018. URL https://arxiv.org/abs/1804.08418v1.

Valerio Perrone, Rodolphe Jenatton, Matthias W. Seeger, and Cédric Archambeau. Scalable hyperparameter transfer learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/14c879f3f5d8ed93a09f6090d77c2cc3-Abstract.html.

Qian Qian and Xiaoyuan Qian. The implicit bias of AdaGrad on separable data. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/3335881e06d4d23091389226225e17c7-Abstract.html.

Jacob Roll. Piecewise linear solution paths for parametric piecewise quadratic programs. *IFAC Proceedings Volumes*, 41(2):2732–2737, 2008. doi: 10.3182/20080706-5-KR-1001.00460. URL https://doi.org/10.3182/20080706-5-KR-1001.00460.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.00489.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

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(70):1–57, 2018. URL https://www.jmlr.org/papers/v19/18-188.html.

Kevin Swersky, Jasper Snoek, and Ryan P. Adams. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/f33ba15effa5c10e873bf3842afb46a6-Abstract.html.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10849–10858, 2021. URL https://proceedings.mlr.press/v139/wang21q.html.

Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? In *Advances in Neural Information Processing Systems*, volume 35, pages 26764–26776, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ab3f6bbe121a8f7a0263a9b393000741-Abstract-Conference.html.

Xiangwu Wang, Chengwei Cao, Yicheng Song, Ran Bi, and Peilin Yu. Stability annealing selects the implicit bias of smoothed sign descent: A rate-indexed barrier path on separable data. *arXiv preprint arXiv:2607.06013*, 2026. URL https://arxiv.org/abs/2607.06013.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, pages 17084–17097, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/8df7c2e3c3c3be098ef7b382bd2c37ba-Abstract.html.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of Adam on separable data. In *Advances in Neural Information Processing Systems*, volume 37, pages 23988–24021, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2ac79356a03fe5e9250e5e77ebc76e6e-Abstract-Conference.html.

A PROOF GUIDE AND CONVENTIONS

The appendix follows the main text's dependencies. Appendices B–I prove selection and the finite hierarchy. Appendix J then proves complete trajectory calibration and the rate-only counterexample. Appendices N and P establish the risk criteria and synchronized controller. The remaining sections document the three main figures, frozen-representation diagnostic and reproducibility record.

A.1 EXPANDED SETUP AND TARGET DEFINITIONS

The grouped risk uses native pair rows $z_{i,k} = (e_{y_i} - e_k) \otimes x_i$, $k \neq y_i$, with the example's positive weight on each pair. A singleton group is binary logistic; EXP uses the ungrouped weighted atom sum. All native softmax parameter coordinates are retained during training. The primary dual solution $p(r)$ is selected by KL relative to the positive weights. The normalized atoms $\xi_{i,t} = \omega_i e^{-z_i^\top w_t}/\epsilon_t$ (pair rows for softmax) converge to $p(r)/\sqrt{A_2}$. The physical barrier is $\Phi(a) = (A_1/A_2)B(a/\sigma)$, with Hessian $H = H_r/A_1$ at a_*. With step and memory fixed, $d_\sigma = \eta[\rho_1/(1-\rho_1) - \rho_2/(1-\rho_2)]$. The schedule amplitude and onset scale can change finite offsets and risk amplitudes even when divergent coordinates agree.

Table 1: Preservation targets for individually eligible runs on the same margin system, loss and weights. Complete invariants refer to the stated canonical model; the last column records implications that do not follow.

Target	Observable	Invariant	Not guaranteed
Trajectory	$W_a(s) - W_b(s)$	(r, Λ)	Unit risk ratio, equal update count at matched effective time, wall-clock equality, validation improvement
Effective-time training risk	$R(W_a(s))/R(W_b(s))$	(r, Λ, Γ)	Equal-update risk or equal wall-clock cost
Equal-update training risk	$R(w_{a,n+1})/R(w_{b,n+1})$	(ϑ, λ, K)	Full trajectory equivalence

A.2 RELATION TO FINITE-HORIZON APPROXIMATIONS

The distinction relevant to the proof is the order of limits. Cattaneo and Shigida (2025) control memory corrections on fixed horizons as the step vanishes; here the step is fixed and the observation horizon grows. After every divergent term is removed, the remaining actual-update error must therefore be summable. Appendix K gives the regular-face small-step derivative of this exact fixed-step branch. The nonlinear and hyperparameter-transfer settings discussed in Section 2 concern different preservation targets.

B NORMALIZED DYNAMICS AND STATIC ESTIMATES

The proof develops the exact normalized recurrence, establishes entry and secondary selection, and then closes the recursive expansion. The complete calibration criterion follows in Appendix J.

Throughout, $D_h(p, x) = \sum_i [p_i \log(p_i/x_i) - p_i + x_i]$, with $0 \log 0 = 0$, and the current atom vector is strictly positive. For differentiable convex f, write $D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$. For $v \neq 0$, $N(v) = v/\|v\|$ and $\Pi_v = I - vv^\top/\|v\|^2$. The notation Π_V^H denotes the H-orthogonal projection onto a subspace V; D^+ denotes the upper right Dini derivative. A superscript $+$ on a state means its next update. Geometric tail factors θ lie in $(0, 1)$. For softmax class blocks $w^{(c)}$, class-centering is $(Cw)^{(c)} = w^{(c)} - n_{\text{cls}}^{-1} \sum_b w^{(b)}$, where n_{cls} is the number of classes. All constants in a fixed-step tail may depend on that step; uniformity in an interval of steps is invoked only where explicitly stated. The parameter w_t is queried before update t. In the queried normalized recurrence below, A_t, ζ_t already include query t. In contrast, an ordinary checkpoint after n completed updates stores w_{n+1}, m_n, v_n, at physical state time $S = n\eta$. Advancing its moments with query $n + 1$ produces the next queried state; these two state conventions must not be interchanged.

Lemma 5 (Exact nonautonomous recurrence)**.** *Define*

$$\xi_{i,k,t} = \omega_i e^{-z_{i,k}^\top w_t}/\epsilon_t, \quad D_{i,t} = 1 + \frac{\epsilon_t}{\omega_i}\sum_k \xi_{i,k,t}, \quad Q_t = \sum_{i,k} \frac{z_{i,k}\xi_{i,k,t}}{D_{i,t}}, \quad Q_t^0 = Z^\top \xi_t,$$

with $D_{i,t} = 1$ for EXP. Put $A_t = -m_t/\epsilon_t$, $\zeta_t = \sqrt{v_t}/\epsilon_t$, and $k_t = k - (\lambda/\eta)\log(1 + \eta/(S_t + s_0))$. Then $Q_t = -g_t/\epsilon_t$ and

$$h_t = \frac{A_t/C_{1,t}}{1 + \zeta_t/\sqrt{C_{2,t}}}, \qquad \xi_{i,t+1} = \xi_{i,t}\exp\{\eta(k_t - z_i^\top h_t)\}, \tag{19}$$

$$\epsilon_{t+1} = \epsilon_t e^{-\eta k_t}, \qquad A_{t+1} = \beta_1 e^{\eta k_t} A_t + (1-\beta_1)Q_{t+1},$$

$$\zeta_{j,t+1} = \left[\beta_2 e^{2\eta k_t}\zeta_{j,t}^2 + (1-\beta_2)Q_{j,t+1}^2\right]^{1/2}. \tag{20}$$

The initial queried moments are $(1-\beta_1)Q_1$ and $\sqrt{1-\beta_2}|Q_1|$. These identities include zero and cancelling aggregate coordinates.

Proof. Differentiate the entire example log loss to get its denominator D_i. Divide each physical moment update by the new ϵ, and use $\epsilon_t/\epsilon_{t+1} = e^{\eta k_t}$. Dividing the two exact reference exponentials gives the atom update. The clock formula follows by taking the logarithm of the ratio in (2). No reference-mass bound or asymptotic approximation is used. □

Uniformly on the declared compact profile/rate family and a sufficiently late tail,

$$\delta_t := k_t - k = -\lambda/(S_t + s_0) + O(\eta/(S_t + s_0)^2),$$

$$|\delta_{t+1} - \delta_t| \le C\eta/(S_t + s_0)^2, \qquad \sum_{t \ge N} \eta\delta_t^2 \le C/(S_N + s_0 - \eta). \tag{21}$$

The continuous local rate is $k(S) = k - \lambda/(S + s_0)$, and k_t is its exact step average. The stabilizer is integrable and exponentially bounded on a tail. Positivity of the local rate is needed only eventually. For any legal frozen K, write

$$s_K(Q) = \frac{A_1(K)Q}{1 + \sqrt{A_2(K)}|Q|}.$$

This map is jointly $C^{1,1}$ in (Q, K) on bounded legal sets. The RMS graph $\sqrt{A_2(K)}|Q|$ is Lipschitz, which is enough for all graph tracking below.

C Static geometry and estimates used for entry

Lemma 6 (Uniform static fiber and selected multiplier)**.** *On a compact interior rate interval and sufficiently small closed step interval, the scaled potential $\Phi(a) = (A_1/A_2)B(a/\sigma)$ subject to $Za \ge k\mathbf{1}$ has unique optimizer $a_* = \sigma u_{k/\sigma}$, uniformly inside its barrier box. Its conjugate is*

$$\Phi^*(Q) = \frac{A_1}{A_2}\sum_j [\sqrt{A_2}|Q_j| - \log(1 + \sqrt{A_2}|Q_j|)], \quad \nabla\Phi^*(Q) = s_k(Q).$$

Every optimal dual multiplier has aggregate $q_* = \nabla\Phi(a_*)$ and mass $m_* = a_*^\top q_*/k$. The full optimal fiber is

$$\mathcal{Y}_{k,\eta} = \{y \ge 0 : Z^\top y = q_*,\ \mathbf{1}^\top y = m_*\}.$$

Its KL-selected member p_* is jointly continuous in (k, η), including at face changes, and equals $p(k/\sigma)/\sqrt{A_2}$. Its positive support is the union of supports of all fiber members.

Proof. Choose a strict feasible point inside the unit box for the largest rate. For sufficiently small steps it remains strictly within every scaled box; poles stay uniformly below one and A_ℓ are uniformly bounded above and away from zero. Comparing the objective at this point and using barrier divergence gives a common positive boundary distance for the optimizer. Strict convexity and the substitution $a = \sigma u$ prove the primal formula. Slater duality gives a nonempty multiplier fiber. Strict convexity of Φ^* forces a common aggregate, while complementarity gives $km_* = a_*^\top q_*$. Conversely, these two

equalities give the optimal dual value, proving the full-fiber characterization. Bounded aggregate and positive k bound its mass; the fiber is compact. Strict convexity of KL with positive base gives its unique minimizer.

Primal continuity follows by compactness and the feasible competitor $(k'/k)a$ for nearby parameters. Its limit stays in the limiting box; optimality identifies every cluster point. The gradient and mass identities give continuity of (q_*, m_*). For a fixed matrix A, the nonnegative fibers $P_b = \{y \geq 0 : Ay = b\}$ obey $\operatorname{dist}(x, P_b) \leq H_A \|Ax - b\|$ for $x \geq 0$ and feasible b. Here is a short proof of the required fixed-matrix Hoffman bound (Peña et al., 2018). Project x onto P_b, obtaining z. In the inequality representation $C = (A; -A; -I)$, write the normal $x - z = C_I^\top \mu$, $\mu \geq 0$, using a linearly independent subset of active rows, obtained by removing conic dependencies. The finitely many such subsets give $\|\mu\| \leq H_A \|x - z\|$. Then

$$\|x - z\|^2 = \mu^\top (C_I x - d_{b,I}) \leq H_A \|x - z\| \|(Cx - d_b)_+\| = H_A \|x - z\| \|Ax - b\|.$$

Cancellation proves the bound. Apply it to $A = (Z^\top; \mathbf{1}^\top)$. Any limiting selected multiplier can be approximated by feasible multipliers in nearby fibers. Optimality and continuity of $D_h(\cdot, \omega)$, including its zero first coordinates, show that every selected cluster point is the unique limiting KL minimizer. Compactness gives uniform continuity.

The scaled fiber corresponds to the unit one by multiplication by $1/\sqrt{A_2}$. All its members have the same mass, so scaling changes the KL objective only by a positive scalar and a common constant. This proves the selected-split formula. Finally, if another fiber member is positive at a zero coordinate of the KL minimizer, moving toward it gives a negative infinite one-sided entropy derivative there; finite derivatives at the positive coordinates cannot offset it. This contradicts minimality, proving the support assertion. □

Lemma 7 (Gap controls mass forcing and weighted drift)**.** *Let $D(y) = k\mathbf{1}^\top y - \Phi^*(Z^\top y)$, $\Delta(y) = D(p_*) - D(y)$, and $s = Za_* - k\mathbf{1} \geq 0$. For $y \geq 0$,*

$$\Delta(y) = D_{\Phi^*}(Z^\top y, q_*) + s^\top y, \tag{22}$$

$$k(\mathbf{1}^\top y - m_*) = a_*^\top (Z^\top y - q_*) - s^\top y. \tag{23}$$

On a prescribed mass box $\mathbf{1}^\top y \leq R$, uniformly in the compact parameter family,

$$\|Z^\top y - q_*\|^2 + \sum_i y_i [k - z_i^\top s_k(Z^\top y)]^2 \leq C_R \Delta(y).$$

For sufficiently small $|\delta|$, $\delta(\mathbf{1}^\top y - m_*) \leq \Delta(y)/4 + C_R \delta^2$.

Proof. Expand the conjugate Bregman divergence and use complementarity. Its Hessian is $\operatorname{Diag}[A_1/(1 + \sqrt{A_2}|Q_j|)^2]$, uniformly positive on the bounded aggregate hull; this controls the squared aggregate error. Write the row drift as $-s_i + z_i^\top [a_* - s_k(Z^\top y)]$. The weighted square is at most $2 \max_i s_i \, s^\top y + 2R \max_i \|z_i\|^2 A_1^2 \|Z^\top y - q_*\|^2$. Neither an inverse inactive slack nor an inverse selected mass appears. Equation (23), Cauchy–Schwarz and Young's inequality bound the remaining forcing by $C|\delta|\sqrt{\Delta} + |\delta|\Delta/k$; reducing $|\delta|$ proves the assertion. □

Lemma 8 (RMS tracking through cancellation)**.** *At a constant frozen rate with legal factors, put $e^A = A - A_1 Q$, $e^\zeta = \zeta - \sqrt{A_2}|Q|$, and $e = (e^A, e^\zeta)$. The queried map gives*

$$e^{A,+} = \rho_1 e^A + \rho_1 A_1 (Q - Q^+), \qquad |e_j^{\zeta,+}| \leq \sqrt{\rho_2}(|e_j^\zeta| + \sqrt{A_2}|Q_j^+ - Q_j|).$$

On bounded sets, $\|h - s_k(Q)\| \leq C\|e\| + C\theta^t$, where the last term is absent without bias correction. Replacing the rate by $k + \delta$ adds $C\eta|\delta|$ to the graph-error bounds.

Proof. The first identity uses $(1 - \rho_1)A_1 = 1 - \beta_1$. Compare the two Euclidean norms $\|(\sqrt{\rho_2}\zeta_j, \sqrt{1 - \beta_2}Q_j^+)\|$ and $\|(\sqrt{\rho_2 A_2}|Q_j^+|, \sqrt{1 - \beta_2}Q_j^+)\|$. The reverse triangle inequality and $\big||Q_j| - |Q_j^+|\big| \leq |Q_j - Q_j^+|$ give the RMS bound. Direct subtraction in the quotient, whose denominator is at least one, gives $|h_j - s_k(Q)_j| \leq |e_j^A| + A_1|Q_j||e_j^\zeta|$. Bias factors stay above $1 - \beta_\ell$, and their tail discrepancy is exponential. Finally the rate dependence of ρ_1 and $\sqrt{\rho_2}$ is $O(\eta|\delta|)$. At $\beta_2 = 0$ the RMS graph error is exactly zero. No derivative of $\sqrt{v}$ at zero is used. □

A shared normalization estimate. If $X(s)/s \to u \neq 0$ and $(X(s) - Y(s))/a(s) \to v$, where $a(s) > 0$ and $a(s) = o(s)$, then

$$\frac{s}{a(s)}[N(X(s)) - N(Y(s))] \longrightarrow \frac{\Pi_u v}{\|u\|}. \tag{24}$$

Indeed, with $D = X - Y$, homogeneity and the mean-value formula give

$$\frac{s}{a(s)}[N(X) - N(Y)] = \int_0^1 DN\left(\frac{Y + \alpha D}{s}\right) \frac{D}{a(s)}\, d\alpha.$$

The arguments converge uniformly to u, and $DN(u) = \Pi_u/\|u\|$. The same estimate applies after any fixed linear map C with $Cu \neq 0$. Below we use $a(s) = 1$ for finite offsets and $a(s) = \log s$ for logarithmic separation.

D STANDARD-INITIALIZATION PRIMARY CAPTURE

We prove the primary part of Theorem 2 for the whole stated compact family. The nonautonomous argument uses the static lemmas above, not an assumed autonomous convergence conclusion.

1. True reduced-flow risk decays before reference localization. Consider $\bar{w}' = \mathcal{S}(-\nabla R/\epsilon)$, where $\mathcal{S}(q) = q/(1+|q|)$, with the exact continuous schedule. The field has coordinate norm at most one, so solutions are global. For log losses let $A_i = \sum_k e^{-z_{i,k}^\top \bar{w}}$, $G = \sum_i \omega_i A_i/(1 + A_i)$, and write $-\nabla R = Gv$. The vector v is a convex combination of margin rows and $\|v\|_1 \geq \gamma_\infty$. For $l = \epsilon/G$, $x^2/(|x| + l) \geq |x| - l$ gives

$$R' = -G \sum_j \frac{v_j^2}{|v_j| + l} \leq -\gamma_\infty G + d\epsilon.$$

Therefore $R(S) + d\int_S^\infty \epsilon(V)\, dV$ decreases. On the resulting common bounded risk sublevel, $A_i/(1 + A_i) \geq c \log(1 + A_i)$ with a uniform $c > 0$, so $G \geq cR$. Variation of constants proves uniform $R \to 0$ and $\int_0^\infty R\, dS < \infty$. In particular every $A_i \to 0$ uniformly. For EXP take $G = R$.

Let $X = \sum_i \omega_i A_i$, $l_{\mathrm{ref}} = \epsilon/X$, and $v_{\mathrm{ref}} = X^{-1} \sum_{i,k} \omega_i z_{i,k} e^{-z_{i,k}^\top \bar{w}}$. For log losses, $\tau = \max_i A_i \to 0$, $G/X \in [(1 + \tau)^{-1}, 1]$, and $\|v - v_{\mathrm{ref}}\| \leq C\tau$. Differentiating $\log l_{\mathrm{ref}}$, applying the previous scalar inequality and bounding the mixture difference gives

$$(\log l_{\mathrm{ref}})' \geq \gamma_\infty - k(S) - d(1 + \tau) l_{\mathrm{ref}} - C\tau.$$

For EXP use $\tau = 0$. On a common tail, $k(S)$ is uniformly below γ_∞. Choose $l_0, \nu > 0$ so the right side is at least ν whenever $l_{\mathrm{ref}} \leq l_0$. At the starting time the finite parameter box and positive schedule give a common positive lower bound on l_{ref}. It reaches l_0 in uniformly bounded time and cannot cross downward. Hence the reference mass X/ϵ is bounded, including its compact finite prefix. Only now infer $\|Q - Q^0\| \leq C\epsilon$ from the group denominators.

2. Integrable forcing and continuous selection. At $\eta = 0$, fix the selected static (p, q, a) at rate k. Differentiation of KL along the exact log orbit gives its rate perturbation $\delta(S)(\mathbf{1}^\top \xi - \mathbf{1}^\top p)$, where $\delta(S) = -\lambda/(S + s_0)$. The monotone quotient and Lemma 7 absorb it, on the already established mass box and a late tail, into

$$\frac{d}{dS} D_h(p, \xi) \leq -c\Delta + C(\delta(S)^2 + \epsilon^2).$$

The true/reference gradient error pairs with $Q^0 - q$ and costs $C\epsilon^2$. Thus KL is bounded and $\int \Delta < \infty$. Bounded log velocity and derivatives on the mass box make Δ uniformly continuous, so $\Delta \to 0$. Every atom cluster point is optimal. To identify its split, observe the exact identity

$$\log(\xi/\omega) = -Z\bar{w} + [kS - \lambda \log(1 + S/s_0) - \log \epsilon_0]\mathbf{1}. \tag{25}$$

All optimizers have the same aggregate and mass. Consequently $D_h(p, \xi) \leq D_h(z, \xi)$ for each optimizer z, by the KL-minimality of p with base ω. Along $\xi(S_j) \to z$, the right side tends to zero. When $z_i = 0$ its term is simply $\xi_i(S_j)$; other coordinates have positive limits. This uses a fixed first argument, not joint boundary continuity of KL. It follows that $z = p$, and $\xi \to p$.

3. A uniform continuous entry time. On the provisional set $D_h(p,\xi) \le 1$, $\epsilon \le 1$, fix the mass and larger transition boxes. KL bounds mass uniformly because $x - p\log x$ grows linearly as $x \to \infty$, uniformly for bounded p, while $p\log p$ remains bounded at zero. Choose K_ϵ large enough to absorb the $C\epsilon^2$ forcing; on a common tail $k(S) \ge k_{\min}/2$. Fix C_δ from Lemma 7 in this box. Then

$$V(S) = D_h(p,\xi) + K_\epsilon \epsilon^2 + C_\delta \int_S^\infty \delta(U)^2\, dU$$

satisfies $V' \le -c\Delta - c'\epsilon^2$ whenever $V \le 1$. These constants precede the entry time. Pointwise convergence supplies strict entry $V < 1/16$ for each family member. At any finite time atoms are positive, so finite-time solution continuity and Lemma 6 make V continuous in all family parameters. Non-exit preserves each strict entry. An open cover followed by a finite subcover gives one common time S_0 with $V(S_0) < 1/16$. It can be increased to meet the discrete tail reserve below.

4. A discrete inequality including the first exit step. For a fixed $\eta > 0$, use the static objects at (k,η) and the *true* graph error

$$e_t = (A_t - A_1(k)Q_t,\ \zeta_t - \sqrt{A_2(k)}|Q_t|), \qquad E_t = D_h(p_*,\xi_t) + \|e_t\|^2 + K_\epsilon \epsilon_t^2.$$

Current membership in $E_t \le 1$ bounds atoms, moments and both corrected quotients. A larger fixed box contains the full exponential one-step segment, including a proposed exit. Let $d_i = k_t - z_i^\top h_t$. The exact KL increment is

$$D_h(p_*,\xi^+) - D_h(p_*,\xi) = \eta \sum_i (\xi_i - p_{*,i}) d_i + \sum_i \xi_i (e^{\eta d_i} - 1 - \eta d_i).$$

The second term is bounded by $C\eta^2 \sum_i \xi_i d_i^2$ on that segment. Lemmas 7 and 8, together with Young's inequality, imply

$$\begin{aligned} \mathrm{KL}^+ - \mathrm{KL} &\le -c_0\eta\Delta + C_0\eta(\|e\|^2 + \epsilon^2 + \delta_t^2 + \theta^{2t}), \\ \|\xi^+ - \xi\|_1^2 &\le C\eta^2(\Delta + \|e\|^2 + \epsilon^2 + \delta_t^2 + \theta^{2t}). \end{aligned} \tag{26}$$

The group denominators are at least one and $|\epsilon^+ - \epsilon| \le C\eta\epsilon$. Smoothness on the transition box bounds $\|Q^+ - Q\|^2$ by the same right side. The RMS norm estimate and its $O(\eta|\delta_t|)$ rate perturbation, with a fixed beta contraction gap, give after Young's inequality

$$\|e^+\|^2 \le \bar\rho\|e\|^2 + C_1\eta^2(\Delta + \epsilon^2 + \delta_t^2 + \theta^{2t}), \quad \bar\rho < 1.$$

Terms $O(\eta^2\|e\|^2)$ are absorbed in $\bar\rho$. Choose K_ϵ from the provisional boxes and then reduce η_0 to absorb the $C_1\eta^2\Delta$ and $C_0\eta\|e\|^2$ terms. Tail positivity of k_t supplies the negative epsilon term. We obtain

$$E_{t+1} - E_t \le -c\eta\Delta_t - c_e\|e_t\|^2 - c_\epsilon\eta\epsilon_t^2 + C\eta\delta_t^2 + C\eta\theta^{2t}. \tag{27}$$

All estimates use current membership only. The correction term is absent for uncorrected Adam. The case $\beta_2 = 0$ has zero RMS graph error, so no division by β_2 occurs. After fixing these constants, choose S_0 large enough that $C\sum_{S_t \ge S_0} \eta\delta_t^2 < 1/4$ uniformly for the provisional small-step interval.

5. Stopped shadowing supplies actual entry. On $[0, S_0 + 1]$ the schedule has a positive uniform minimum and bounded logarithmic derivatives. Let R_0 bound initial parameters. Reduced solutions have radius at most $R_0 + S_0 + 1$. Choose two larger closed boxes with successive margins one, and stop the discrete process at its first query outside the inner box. On the outer box Q and its first derivatives are bounded by a fixed constant. For small steps, the local moment factors satisfy $\rho_1(k_t), \sqrt{\rho_2(k_t)} \le \theta_0 < 1$. Unrolling from zero moments bounds A, ζ, h at every pre-exit query. Reduce η so the last pre-exit parameter segment stays in the outer box. Then $\|Q_{t+1} - Q_t\| = O(\eta)$ even through the putative exit. RMS tracking, $A_\ell(k_t) = 1 + O(\eta)$, and the initial geometric layer give quotient error $C(\eta + \theta^t)$. Euler local error of the reduced flow is $O(\eta^2)$, hence stopped Gronwall yields

$$d_{t+1}^w \le (1 + C\eta)d_t^w + C\eta^2 + C\eta\theta^t, \qquad \max_{S_t \le S_0+1} d_t^w \le C_{S_0}\eta.$$

Reducing η below the fixed box margin rules out the proposed exit. The parameter and reference-log shadowing bounds therefore hold on the whole finite window; summed initial-layer error is $O(\eta)$.

At $N = 1 + \lceil S_0/\eta \rceil$, static continuity gives $p_{k,\eta} \to p_{k,0}$ uniformly. Finite-window atoms lie in a compact positive box. Uniform continuity of KL there, the continuous entry, and the moment error $O(\eta + \theta^N)$ give $E_N < 1/4$. Shrink the step once more so the corrected tail reserve in (27) is less than $1/4$. A first exit from $E \le 1$, summed *through* that step, would have $E < 3/4$, a contradiction. Only now sum to infinite time. This gives $\sum \eta\Delta_t < \infty$ and $\sum \|e_t\|^2 < \infty$. At fixed η, $\Delta_t, e_t \to 0$. The exact log-leaf identity (25) again identifies the unique selected cluster point p_*. Moment and update limits follow, and Cesàro gives $w_t/S_t \to a_*$. Groupwise Taylor expansion yields both $R(w_t)/\epsilon_t \to \mathbf{1}^\top p_*$ and the same limit for true normalized gradient mass. No secondary rate or parameter expansion was used in this primary argument.

E A COMMON ADMISSIBLE STEP BEFORE LOCAL FACE CONSTANTS

Lemma 9 (Upper-box KL–RMS absorption)**.** *Retain the common mass and transition boxes from Appendix D, a fixed beta pair and the compact rate/profile family of Theorem 2. There is a common positive additional step bound for which every frozen positive subsystem has a nonincreasing KL–RMS energy. On a fixed positive log leaf, near its equilibrium, this energy contracts strictly. The local neighborhood and contraction rate may depend on the face and its positive masses; the step bound does not depend on an inverse positive mass.*

Proof. Freeze omitted mass x and local rate K_0. Write $q = Z_I^\top y + Z_B^\top x$, $q^* = Z_I^\top y^* + Z_B^\top x$, $y^* > 0$, and $h^* = s_{K_0}(q^*)$, with $Z_I h^* = K_0 \mathbf{1}$. The uncorrected frozen map is exactly

$$h = A/(1+\zeta), \qquad y_i^+ = y_i e^{\eta(K_0 - z_i^\top h)},$$
$$A^+ = \rho_1 A + (1-\beta_1) q^+, \qquad \zeta_j^+ = \|(\sqrt{\rho_2}\zeta_j, \sqrt{1-\beta_2} q_j^+)\|_2.$$

Set $u = q - q^*$, $e_A = A - A_1 q$, $e_\zeta = \zeta - \sqrt{A_2}|q|$, $e = (e_A, e_\zeta)$, and $D = D_h(y^*, y)$. On the common enlarged box there are common positive m, L, C_h, M, G with

$$\langle u, s_{K_0}(q) - s_{K_0}(q^*)\rangle \ge m\|u\|^2, \quad \|s_{K_0}(q) - s_{K_0}(q^*)\| \le L\|u\|, \quad \|h - s_{K_0}(q)\| \le C_h\|e\|,$$

and $\sum_i y_i \le M$, $\|z_i\| \le G$. Indeed the coordinate derivative of s_{K_0} is $A_1/(1+\sqrt{A_2}|q_j|)^2$, positive even at zero; the denominators of h are at least one. Put $d_i = K_0 - z_i^\top h = -z_i^\top(h - h^*)$. The exact KL increment is

$$D^+ - D = -\eta\langle u, h - h^*\rangle + \sum_i y_i(e^{\eta d_i} - 1 - \eta d_i).$$

Taylor's integral remainder on the common transition box bounds the sum by $C\eta^2 M G^2 (L\|u\| + C_h\|e\|)^2$. Monotonicity and Young's inequality bound the first term by $-\eta m\|u\|^2/2 + C\eta\|e\|^2$. Thus a common step reduction gives

$$D^+ - D \le -c\eta\|u\|^2 + C_D\eta\|e\|^2. \tag{28}$$

No lower bound on y_i^* has entered.

Apply Lemma 8 to the frozen positive subsystem. It controls both moment errors by their memory contractions and $\|q^+ - q\|$, including zero and cancelling coordinates. Choose a common $\theta < 1$ bounding $\rho_1, \sqrt{\rho_2}$ on a preliminary small step interval. The exact exponential atom map gives $\|q^+ - q\| \le C\eta(\|u\| + \|e\|)$. Young's inequality, with a fixed gap between θ and one, yields common $\rho' < 1, C_E$ after another upper-box step reduction:

$$\|e^+\|^2 \le \rho'\|e\|^2 + C_E\eta^2\|u\|^2.$$

This includes canceled coordinates and zero betas. With graph weight one, $V = D + \|e\|^2$, choose

$$\eta \le \min\{\eta_{\mathrm{pre}}, c/(2C_E), (1-\rho')/(2C_D)\},$$

omitting any vacuous zero-denominator restriction. Then

$$V^+ - V \le -\tfrac{1}{2}c\eta\|u\|^2 - \tfrac{1}{2}(1-\rho')\|e\|^2. \tag{29}$$

All step choices are finished at this point.

Only now fix a positive face and log leaf. Write $y = y^* \exp(z)$, $z \in \operatorname{range} Z_I$, in independent coordinates. If the derivative of $z \mapsto Z_I^\top[y^* \exp(z)]$ annihilates z, pairing with its row-space preimage

gives $z^\top \operatorname{Diag}(y^*)z = 0$, hence $z = 0$. Local invertibility gives $D \le C_{\mathrm{face}}\|u\|^2$. Consequently (29) contracts V on a sufficiently small invariant energy sublevel. Its rate is allowed to approach one when a positive equilibrium mass approaches zero.

Locally, each scalar KL term is a positive smooth factor times $(y_i - y_i^*)^2$; signed square-root coordinates are Lipschitz jointly in state and root. The RMS graph is Lipschitz. Comparing the actual next state with the frozen map and then moving its equilibrium costs $C_{\mathrm{face}}(\|x^+ - x\| + |\delta^+ - \delta|) + Ce^{-a_0 S}$. This proves (33), including true-loss and bias corrections, without a further face-dependent step restriction. □

Choice	Allowed dependencies
Mass/transition boxes; moment gaps	Data, compact initialization/rate/profile family, fixed beta pair; no inverse selected mass
Common η_0	Primary-entry restrictions and (29); no inverse selected mass
Leaf inverse, neighborhood, $q_0 < 1$	Fixed face, step and selected positive masses
Weak separator, tail onset and remainder	Fixed geometry, step and trajectory

The dependency order is common boxes and gaps, common step, actual primary entry, fixed-face local tracking, slow ratios, the one-sided mass bound, and finally the refined update and coefficient programs. In particular no weak-mass power law is used to prove this lemma.

F Forced secondary selection on arbitrary faces

Fix one eligible positive step. Write $a = a_*$, $q = q_*$, $p = p_*$, $H = H_r/A_1$, and $B = B_0$. Strictly inactive rows decay exponentially: $h_t \to a$, $k_t \to k$, and their limiting slack is positive in (19). True-gradient/reference differences and bias correction are also exponentially small on the established mass box.

1. Essential support gives a cone-coercive weak subsystem. Let $T = \ker Z_J$, P its H-orthogonal projection, and

$$L = PH^{-1}Z_B^\top, \qquad K = Z_B PH^{-1}Z_B^\top.$$

There is $t \in T$ with $Z_B t > 0$. To prove this, failure of strict separation would give a nonzero $x \ge 0$ with $Z_B^\top x \in \operatorname{range} Z_J^\top$. Choose signed v_J so $Z_J^\top v_J + Z_B^\top x = 0$. Since $p_J > 0$, a small perturbation $(p_J + \alpha v_J, \alpha x)$ is nonnegative with the same aggregate. All involved rows are tight, so pairing with a shows that its mass also stays fixed. It is dual optimal, contradicting the essential-support property in Lemma 6. Thus a strict separator exists. By $\|Lx\|_H^2 = x^\top Kx$ and compactness of the nonnegative simplex,

$$x^\top Kx \ge \alpha(\mathbf{1}^\top x)^2 \quad (x \ge 0), \quad \alpha > 0. \tag{30}$$

Signed positive definiteness of K is unnecessary. If B is empty, skip the mass argument and use the regular-response proof below.

2. Freeze weak mass and local rate together. Hold input q fixed when defining $j = \partial_k s_k(q)$. Set

$$v = Pj + (I - P)a/k,$$

the minimum of $\frac{1}{2}(v-j)^\top H(v-j)$ under $Z_J v = \mathbf{1}$. For small nonnegative x and small real δ, solve

$$\begin{aligned} Q(x,\delta) &= Z_J^\top y_J(x,\delta) + Z_B^\top x, \\ Z_J s_{k+\delta}(Q(x,\delta)) &= (k+\delta)\mathbf{1}, \qquad \log(y_J(x,\delta)/p_J) \in \operatorname{range} Z_J. \end{aligned} \tag{31}$$

The actual positive atoms lie on this log leaf: the entropy optimality condition for p_J, (25), and $\mathbf{1} = Z_J a/k \in \operatorname{range} Z_J$ show the assertion. Use independent coordinates on that row space. At $(x,\delta) = (0,0)$ the derivative in the positive log coordinates is invertible: its quadratic form uses $\operatorname{Diag}(p_J) > 0$ and $H^{-1} > 0$, and $Z_J^\top \operatorname{Diag}(p_J)z = 0$ for $z \in \operatorname{range} Z_J$ implies $z^\top \operatorname{Diag}(p_J)z = 0$, hence $z = 0$. The $C^{1,1}$ implicit-function theorem applies to s_K, even when q has zero coordinates. Differentiation gives

$$Z_J\, dh = d\delta\, \mathbf{1}, \qquad H(dh - j\, d\delta) - Z_B^\top dx \in \operatorname{range} Z_J^\top.$$

Consequently

$$s_{k+\delta}(Q(x,\delta)) = a + Lx + v\delta + O((\|x\| + |\delta|)^2). \tag{32}$$

The corresponding moment graph $(A_1(k+\delta)Q, \sqrt{A_2(k+\delta)}|Q|)$ is locally Lipschitz. This is a positive-subsystem equilibrium, not a claim that all omitted weak inequalities remain feasible at the perturbed rate.

3. Tracking before a weak-mass power law. Let $x_t = \xi_{B,t}$, $M_t = \mathbf{1}^\top x_t$, $\delta_t = k_t - k$. On the fixed positive log leaf, use positive KL relative to $y_J(x_t, \delta_t)$ and squared RMS graph error. For frozen (x, δ), Lemma 9 absorbs the KL–RMS cross terms using only common upper boxes, before any positive-mass lower bound is invoked. Only afterward is positive KL compared locally to aggregate squared distance on the fixed log leaf. The resulting contraction factor and neighborhood may depend on this face; the admissible step is not reduced again. The square root d of the KL–RMS energy is Lipschitz jointly in the positive state and equilibrium: scalar KL can be written locally as a smooth positive factor times the squared difference, so signed square-root coordinates have bounded derivatives. The RMS graph is Lipschitz as well. Replacing the frozen inputs by the next ones therefore gives, after primary capture,

$$d_{t+1} \le q_0 d_t + C\|x_{t+1} - x_t\| + C|\delta_{t+1} - \delta_t| + Ce^{-a_0 S_t}, \qquad q_0 < 1. \tag{33}$$

The exact map uses k_t for this transition; its next comparison uses k_{t+1}. This is why the forcing is a rate *difference*.

Primary selection alone gives $x_{i,t+1}/x_{i,t} \to 1$, so $M_{t+1}/M_t \to 1$ and $\|x_{t+1} - x_t\| = o(M_t)$. For $B_t = M_t + S_t^{-1}$ on a positive-time tail, the ratio tends to one; all other forcing in (33) is $o(B_t)$. Dividing by B_{t+1}, then geometrically summing, yields

$$d_t = o(M_t + S_t^{-1}), \qquad h_t - a = Lx_t - \lambda v/S_t + o(M_t + S_t^{-1}). \tag{34}$$

No lower bound on M_t has been used.

4. A one-sided mass estimate, followed by stronger tracking. The weak-atom log update becomes

$$\log(x_{i,t+1}/x_{i,t}) = \eta[-(Kx_t)_i + \lambda(z_i^\top v - 1)/S_t + o(M_t + S_t^{-1})].$$

After exponentiating and summing against x_t, its Taylor remainder is $O(\eta^2 M_t(M_t + S_t^{-1})^2)$. Cone coercivity absorbs it on a late tail and gives fixed $A, c > 0$ such that

$$M_{t+1} - M_t \le \eta[AM_t/S_t - cM_t^2]. \tag{35}$$

Choose N late enough and $C_0 = \max\{S_N M_N, (A+2)/c\}$, slightly enlarged if necessary. On $0 \le m \le C_0/S$, the comparison map $m \mapsto m + \eta(Am/S - cm^2)$ is increasing: late entry makes $2\eta c \max\{M_N, (A+2)/(cS_N)\} < 1/2$. At the boundary it satisfies

$$C_0/S + \eta(AC_0 - cC_0^2)/S^2 \le C_0/S - 2\eta C_0/S^2 \le C_0/(S+\eta).$$

Induction proves $M_t \le C_0/S_t$. Equation (34) now gives $h_t - a = O(S_t^{-1})$, so the exact weak update yields $\|x_{t+1} - x_t\| = O(\eta/S_t^2)$. Returning to (33) gives $d_t = O(S_t^{-2})$: split its convolution at $t/2$, where the old part is geometric and the recent part has S^{-2} forcing. Thus the actual update, rather than a differentiated parameter ansatz, obeys

$$h_t - a = Lx_t - \lambda v/S_t + O(S_t^{-2}). \tag{36}$$

5. Logarithmic-time optimization includes the zero optimum. Let $\mathfrak{m}_t = S_t x_t$, $\tau_t = \log S_t$, and

$$\mathfrak{d} = (1-\lambda)\mathbf{1} + \lambda Z_B v, \qquad F(\mathfrak{m}) = \mathfrak{d}^\top \mathfrak{m} - \tfrac{1}{2}\mathfrak{m}^\top K \mathfrak{m} \quad (\mathfrak{m} \ge 0).$$

The bounded scaled mass and (36) imply

$$\log(\mathfrak{m}_{i,t+1}/\mathfrak{m}_{i,t}) = (\tau_{t+1} - \tau_t)[\mathfrak{d}_i - (K\mathfrak{m}_t)_i + O(S_t^{-1})].$$

Exponentially interpolate the positive coordinates over each log-time interval. The bounded interpolant has bounded velocity and, almost everywhere,

$$\frac{d\mathfrak{m}_i}{d\tau} = \mathfrak{m}_i[\mathfrak{d}_i - (K\mathfrak{m})_i + e_i(\tau)], \qquad \|e(\tau)\| = O(e^{-\tau}). \tag{37}$$

Interpolation adds $O(\eta/S_t)$, included in this remainder. Apply Lemma 12 with stage time τ, $y = \mathfrak{m}$, $g = \mathfrak{d}$ and the operators H, L, K above. Its hypotheses follow from (30), the boundedness of $S_t x_t$, and (37). The lemma includes a zero maximizer and singular Gram matrices. It gives convergence to an optimal multiplier and hence a unique limiting aggregate $L\mathfrak{m}_*$, which is all that is needed for the next step. The abstract proof of the lemma uses only these hypotheses; its subsequent application to higher Adam scales is separate.

6. The shifted program and the raw parameter. Put $b = -\lambda v + L\mathfrak{m}_*$. Equation (36) gives $S_t(h_t - a) \to b$, and summing the actual updates gives $w_t = S_t a + b \log S_t + o(\log S_t)$. The vector $-\lambda v$ is the H-projection of $-\lambda j$ onto $Z_J b = -\lambda \mathbf{1}$. On its nullspace the remaining program is

$$\min_b \tfrac{1}{2}(b + \lambda j)^\top H(b + \lambda j) \quad \text{s.t. } Z_J b = -\lambda \mathbf{1}, \quad Z_B b \geq (1 - \lambda)\mathbf{1}. \tag{38}$$

Its eliminated weak dual is exactly $\max F(\mathfrak{m})$, up to a constant, with primal map $b = -\lambda v + L\mathfrak{m}$. A large multiple of the strict separator makes the inequalities strict, so strong duality applies even with redundant rows. The positive-definite objective gives a unique primal coefficient. The next section transforms this program into (10), completing the secondary part of Theorem 2.

G Shape response and the critical transition

This geometric result supplies the regular-response expansion and examples used in the trajectory comparison. Its proof is in Appendix H.

G.1 Response to the polynomial exponent

Theorem 10 (Shape response and a transverse obstruction)**.** *For fixed primary geometry (r, H_r, J, B_0), the map $\Lambda \mapsto c_{r,\Lambda}$ is continuous and finite piecewise affine, allowing redundant constraints. If $B_0 \neq \varnothing$, then for every $i \in B_0$, $j \in J$, and every Λ,*

$$\|\Pi_{u_r} c_{r,\Lambda}\| \geq \|z_i - z_j\|^{-1} > 0,$$

where Π_u is Euclidean projection orthogonal to u. The same bound holds in class-centered softmax observation space. On a regular rate neighborhood, with a fixed tight set, positive selected mass on every tight row, and strict inactive slack,

$$w_t = S_t a_* - \lambda a'_*(k) \log(1 + S_t/s_0) + c_\infty + O(S_t^{-1}).$$

A nonzero transverse logarithmic coefficient gives $\log S/S$ direction error; a radial coefficient gives $O(S^{-1})$ direction error.

The transverse bound follows from $(z_i - z_j)^\top c \geq 1$ and $(z_i - z_j)^\top u_r = 0$. For instance, $Z = ((1,0);(1,1))$ gives $u_r = (r, 0)$ and $c = (-\Lambda, 1)$, so the second-coordinate logarithm persists for all polynomial exponents.

H Calibration, regular response and shape geometry

Put $\ell_1 = \eta\rho_1/(1-\rho_1)$ and $\ell_2 = \eta\rho_2/(1-\rho_2)$. Differentiating $\log A_1$ and $\log \sqrt{A_2}$ gives $d_\sigma = \ell_1 - \ell_2$; the second moment's exponent factor two cancels its square-root factor. Differentiating the quotient at fixed input q, coordinatewise, gives

$$j = \partial_k s_k(q) = d_\sigma a + \ell_2 H^{-1} q, \qquad H^{-1} = \operatorname{Diag} \frac{A_1}{(1 + \sqrt{A_2}|q_j|)^2}. \tag{39}$$

At $q_j = 0$ both terms vanish. This fixed-input derivative differs from the derivative of the optimal branch, which also changes its aggregate.

Set $c = b + \lambda d_\sigma a$ in (38). Tightness of all $J \cup B$ rows changes the right sides to $-\lambda(1 - k d_\sigma) = -\Lambda$ and $1 - \Lambda$. The objective becomes

$$\tfrac{1}{2}(c + \lambda \ell_2 H^{-1} q)^\top H (c + \lambda \ell_2 H^{-1} q).$$

Its linear cross term is constant over the equality affine space, since $q = Z_J^\top p_J$ gives $q^\top c = -\Lambda \mathbf{1}^\top p_J$. Also $H = H_r/A_1$ is a scalar multiple of the unit Hessian. The unique minimizer is exactly $c_{r,\Lambda}$, proving the raw-clock coefficient in (11). The full clock satisfies

$$\widehat{S}' = \sigma - \lambda \partial_k \sigma/(S + s_0) > 0 \quad \text{eventually}, \qquad \log \widehat{S} = \log S + \log \sigma + O(\log S/S).$$

Substitution gives (12); the constant multiple $c \log \sigma$ is absorbed by its remainder.

Theorem 11 (Two-scalar schedule calibration)**.** *Fix compact interior target rates, bounded target shapes, and a fixed beta pair. For all sufficiently small fixed steps, there is a common legal interval on which $\chi \geq 1/2$ and*

$$\frac{k}{\sigma(k,\eta)} = r_{\mathrm{tar}}, \qquad \lambda = \frac{\Lambda_{\mathrm{tar}}}{\chi(k,\eta)}$$

have a unique nearby rate root and a bounded shape solution. With a compact positive choice of s_0, the resulting schedules satisfy Theorem 2. In their own full effective clocks, they preserve

$$u_{r_{\mathrm{tar}}}, \qquad c_{r_{\mathrm{tar}},\Lambda_{\mathrm{tar}}},$$

every higher selector, and the hierarchy depth.

Proof of the uniform controller statement. For target rates in a compact interior interval choose a slightly larger compact interval $I \Subset (0, \gamma_\infty)$. At $\eta = 0$, $\sigma = 1$ and $d_\sigma = 0$. Uniform smoothness on I gives a positive step interval with strict pole margins, bounded positive σ, and $\chi \geq 1/2$. The derivative $\partial_k(k/\sigma) = \chi/\sigma > 0$ and the implicit-function theorem give a unique nearby root, uniformly over the target interval. Uniqueness is asserted on this interval, not on an arbitrary domain beyond the poles. Set $\lambda = \Lambda_{\mathrm{tar}}/\chi$, which stays in a compact interval. Choose s_0 in a compact positive interval; it can be chosen large enough for a decreasing schedule from the start. These choices fix a compact controller family *before* invoking the primary theorem's uniform η_0. Intersect its threshold with the pole/root interval. A finite set of beta pairs uses the minimum of the resulting thresholds; uniformity as beta approaches one is not claimed. This proves Theorem 11.

H.1 Regular neighborhoods have a finite parameter offset

When $B = \varnothing$ on a regular neighborhood, the positive equilibrium in (31) depends only on δ. Its quotient $a(K) = \sigma(K)u_{K/\sigma(K)}$ is $C^{1,1}$ by the same positive-leaf implicit-function argument. Redundancy and aggregate cancellation are allowed. The moving graph changes by $O(\eta/S^2)$, so (33) gives $d_t = O(S_t^{-2})$. Hence

$$h_t = a(k_t) + O(S_t^{-2}) = a(k) + a'(k)(k_t - k) + O(S_t^{-2}).$$

There is no extra $1/S$ filtering lag: the instantaneous derivative already includes both memory-factor changes, and its moving-graph lag is $O(S^{-2})$. Use the exact identity $\eta(k_t - k) = -\lambda \log[(S_t + s_0 + \eta)/(S_t + s_0)]$. The increments of

$$w_t - S_t a(k) + \lambda a'(k) \log(1 + S_t/s_0)$$

are $O(\eta/S_t^2)$, summable with tail $O(S_t^{-1})$. They define a finite vector c_∞, proving the regular expansion in Theorem 10. Its nullspace component may depend on initialization, and its beta invariance is not asserted.

Differentiating the unit regular KKT system gives $Z_J u_r' = \mathbf{1}$ and $H_r u_r' \in \operatorname{range} Z_J^\top$. Thus $c_{r,\Lambda} = -\Lambda u_r'$ and $a'(k) = d_\sigma a + \chi u_r'$, agreeing with the arbitrary-face result. Apply (24) to $X = w_t$, $Y = S_t a$ and $a(S_t) = \log S_t$. For a nonzero transverse coefficient this gives

$$\|N(w_t) - N(a)\| \sim \frac{|\lambda| \|\Pi_a a'(k)\|}{\|a\|} \frac{\log S_t}{S_t}.$$

When it vanishes, the proved finite offset gives $O(S_t^{-1})$, which may be exactly zero. For softmax apply the same argument to Cw, Ca.

H.2 Finite affine pieces and a uniform transverse obstruction

The feasible set of (10) is nonempty: start from the $\Lambda = 0$ feasible vector supplied by the strict separator, then subtract $(\Lambda/r)u_r$. For $\Lambda_n \to \Lambda$, translate any feasible competitor by $-((\Lambda_n - \Lambda)/r)u_r$. A common objective bound and positive definiteness bound the minimizers. Any cluster point is feasible and optimal by comparison with these translated competitors. Uniqueness proves continuity.

For each possible set I of binding weak rows, retain a linearly independent subset of the equations for $J \cup I$, giving a full-row-rank matrix A. The constrained minimum-norm solution is

$$H_r^{-1} A^\top (A H_r^{-1} A^\top)^{-1} b(\Lambda),$$

affine in Λ. Dependent rows remain consistency checks. The parameter set on which the candidate satisfies all inequalities and admits nonnegative weak multipliers is the projection of a polyhedron onto a one-dimensional axis, hence an interval, point or empty set. There are finitely many binding subsets. These candidates give a continuous finite piecewise-affine solution path, hence a globally Lipschitz one. This is an application of classical parametric-QP structure (Roll, 2008); the new dynamics selects this specific program.

For any $i \in B, j \in J$, feasibility gives $(z_i - z_j)^\top c \geq 1$ and tightness gives $(z_i - z_j)^\top u_r = 0$. The row difference is nonzero, since otherwise these constraints would be inconsistent. Cauchy–Schwarz proves $\|\Pi_{u_r} c\| \geq 1/\|z_i - z_j\|$, uniformly in Λ. The raw coefficient differs only radially. Equation (24) then gives

$$\lim_{t\to\infty} \frac{S_t}{\log S_t} \|N(w_t) - N(a)\| = \frac{\|\Pi_a b\|}{\|a\|} \geq \frac{1}{\|a\|\|z_i - z_j\|}.$$

Softmax row differences are class-centered, and their pairing with Cu_r is zero. Replace c, u_r, a by Cc, Cu_r, Ca in the same argument. Separability makes $Ca \neq 0$. This boundary concerns the fixed polynomial prefactor family on its original primary face, not all time-varying schedules.

H.3 Exact examples and the weak multiplier at the transition

For the single-row example, $z = (1/3, 1)$, $r = 7/12$, $u = (1/4, 1/2)$, $\nabla B(u) = z$, and $H_r = \mathrm{Diag}(16/9, 4)$. Then $H_r^{-1} z/(z^\top H_r^{-1} z) = (3/5, 4/5)$, giving the stated coefficient. Its transverse norm is $2|\Lambda|/(5\sqrt{5})$. The same direction occurs only at an isolated positive rate along this regular branch: stationarity of cu, $0 < c < 2$, reduces to $[\nabla B(cu)]_1 = [\nabla B(cu)]_2/3$, whose sole positive interior solution is $c = 1$. Thus regularity of a rate does not imply an interval of equivalent directions.

For $Z = ((1, 0); (1, 1))$, the optimizer $(r, 0)$ has multiplier $(r/(1 - r), 0)$. The constraints give $c_1 = -\Lambda$, $c_2 \geq 1$, so $c = (-\Lambda, 1)$. Since $a = (k, 0)$ and $\chi + k d_\sigma = 1$, $b = (-\lambda, 1)$.

For the two-row example, set $z_1 = z$ and add $z_2 = (1, 2/3)$. Both margins equal $7/12$, but the unique primary multiplier is $(1, 0)$. The equality-only solution is feasible precisely when $-2\Lambda/15 \geq 1$. Otherwise the two tight equations give the second line of (45). Both formulas agree at $\Lambda = -15/2$, with value $(9/2, 6)$. The equality-eliminated weak Gram is

$$z_2^\top \left[H_r^{-1} - H_r^{-1} z_1 (z_1^\top H_r^{-1} z_1)^{-1} z_1^\top H_r^{-1}\right] z_2 = 49/180.$$

Therefore the weak unit-Hessian multiplier is $\nu_2 = (180/49)(1 + 2\Lambda/15)$ on the binding side, with the correct nonnegative sign. It is zero at and below the transition. Physical scaled weak mass uses the shifted metric H_r/A_1: its multiplier is ν_2/A_1, with the corresponding additional row scaling for a rescaled dataset. This distinguishes a unit-QP multiplier from an observed $S\xi_B$.

I Critical mass, finite hierarchy, and terminal offset

Throughout this section the step is fixed below the common threshold, $H = H_r/A_1$, $a = a_*$, and L, K, v are from Appendix F. The scaled weak mass is $\mathfrak{m} = S\xi_B$, not the primal coefficient c. Its exponential interpolation obeys (37), with $\mathfrak{d} = (1 - \lambda)\mathbf{1} + \lambda Z_B v$, and the actual query satisfies

$$h_t = a - \lambda v/S_t + L\mathfrak{m}(\log S_t)/S_t + O(S_t^{-2}). \tag{40}$$

These are consequences of the actual three-loss, two-bias-convention recurrence and primary capture, not assumptions about a surrogate algorithm.

I.1 Construction and update-level form of the finite hierarchy

The following gives the construction used in Theorem 3. Let $J_2 \subseteq B_0$ be the union of positive inequality-multiplier supports of (10), and let B_2 be its remaining tight inequalities. Starting from

$$A_3 = J \cup J_2, \qquad B_3 = B_2,$$

solve at each nonempty level

$$d^{(q)} = \arg\min_d \frac{1}{2} d^\top H_r d, \qquad Z_{A_q} d = 0, \qquad Z_{B_q} d \geq \mathbf{1}, \qquad q \geq 3. \tag{41}$$

Add the union E_q of positive inequality-multiplier supports to A_q, remove strictly slack rows, and retain the remaining tight rows as B_{q+1}. Every stage uses the *same original Hessian* H_r; the deeper recursion introduces no new schedule parameter.

Under Theorem 2, every nonempty program is strictly feasible. The recursion terminates with at most $\dim\ker Z_{J\cup J_2}$ nonempty levels $q \geq 3$. Put

$$R_2(S) = S, \qquad R_q(S) = S\prod_{j=1}^{q-2} L_j(S), \qquad q \geq 3.$$

At each nonempty level,

$$R_q(S_t)\left(h_t - a_* - \frac{b}{S_t} - \sum_{j=3}^{q-1}\frac{d^{(j)}}{R_j(S_t)}\right) \longrightarrow d^{(q)}. \tag{42}$$

If $B_2 = \varnothing$, take $Q = 2$: the two-term residual converges even when the primary face is degenerate. The finite-limit statement in the main text follows at actual nodes and extends to the prescribed interpolant by Appendix J.2.

I.2 A WORKED CRITICAL EXAMPLE

Fix one eligible step and a single weak row $B = \{i\}$. Set $\ell_{\mathrm{vec}} = PH^{-1}z_i$, $\kappa = z_i^\top \ell_{\mathrm{vec}} > 0$, and $\theta = 1 + \lambda(z_i^\top v - 1)$. At $\theta = 0$,

$$\xi_{i,t} = \frac{1}{\kappa S_t \log S_t} + O\!\left(\frac{1}{S_t(\log S_t)^2}\right), \tag{43}$$

$$w_t = S_t a - \lambda v \log S_t + \frac{\ell_{\mathrm{vec}}}{\kappa}\log\log S_t + w_\infty + O(1/\log S_t). \tag{44}$$

To prove this, set $\tau_t = \log S_t$ and $\mathfrak{m}(\tau_t) = S_t\xi_{i,t}$. The exact log-time interpolant satisfies $\mathfrak{m}' = \mathfrak{m}(-\kappa\mathfrak{m} + e)$, $|e(\tau)| \leq Ce^{-\tau}$. Its reciprocal $y = 1/\mathfrak{m}$ satisfies $y' = \kappa - ey$. The integrating factor and its inverse are bounded, first giving $y = O(1+\tau)$. Integrating again and using $\int_\tau^\infty se^{-s}ds = O(\tau e^{-\tau})$ gives

$$y(\tau) = \kappa\tau + C_0 + O(\tau e^{-\tau}), \qquad \mathfrak{m} = 1/(\kappa\tau) + O(\tau^{-2}).$$

In particular, at actual paired physical nodes S and $2S$, the unknown constant cancels in the reciprocal readout

$$\frac{(2S\xi_i(2S))^{-1} - (S\xi_i(S))^{-1}}{\log 2} = \kappa + O(\log S/S).$$

Here $\xi_i(S_t) = \xi_{i,t}$; the statement uses the stronger reciprocal remainder above, not a difference of coarse leading equivalents. Substitute into (40) before summation:

$$h_t = a - \lambda v/S_t + \frac{\ell_{\mathrm{vec}}/\kappa}{S_t\log S_t} + O\!\left(\frac{1}{S_t(\log S_t)^2} + S_t^{-2}\right).$$

The remainder is summable in cumulative time with tail $O(1/\log S_t)$. Taylor errors for the three displayed primitives are summable at fixed step. This proves the finite offset, which depends on the trajectory. The finite prefix at $S_1 = 0$ is included in that offset.

The two-row phase example has $z_1 = (1/3, 1)$, $z_2 = (1, 2/3)$ and $r = 7/12$:

$$c_{r,\Lambda} = \begin{cases} -\Lambda(3/5, 4/5), & \Lambda \leq -15/2, \\ (9/7, -3/7) - \Lambda(3/7, 6/7), & \Lambda \geq -15/2. \end{cases} \tag{45}$$

At $\Lambda = -15/2$, this unscaled geometry has $u = (1/4, 1/2)$, $c = (9/2, 6)$ and $\ell_{\mathrm{vec}} = A_1(7/20, -7/60)$, $\kappa = A_1(49/180)$. Thus $\theta = 1 + 2\Lambda/15 = 0$ and, in each run's prescribed clock,

$$W(s) = s(1/4, 1/2) + (9/2, 6)\log s + (9/7, -3/7)\log\log s + w_\infty + O(1/\log s).$$

The clock substitution changes only the finite offset and the stated tail remainder. Subtracting the leading and logarithmic terms leaves an unbounded $\log\log s$ term; subtracting the third term closes the residual. This is an analytic example at the critical profile; the weak-row risk experiment below uses $\Lambda = 0.7$.

I.3 A REDUCTION WITH A CLOSED ERROR CLASS

The following lemma isolates the argument needed at every level. Its clock t is continuous stage time, not the discrete query index.

Lemma 12 (Recursive elimination and terminal convergence)**.** *Let H be positive definite on a finite-dimensional space T, $L : \mathbb{R}^b \to T$, and $K = L^\top H L$. Suppose $x^\top K x \geq \alpha(\mathbf{1}^\top x)^2$ for $x \geq 0$, $\alpha > 0$. Let a positive bounded absolutely continuous trajectory satisfy, almost everywhere,*

$$y' = \mathrm{Diag}(y)(g - Ky + e(t)), \qquad \|e(t)\| \leq Ce^{-\gamma t}, \quad \gamma > 0.$$

For the entire maximizer set of $F(p) = g^\top p - p^\top K p/2$ over $p \geq 0$, let E be the union of positive supports, N the strictly slack indices, and B the remaining tight indices. Then $y \to p$ for one maximizer positive on all of E. If B is empty, $L(y - p) = O(e^{-\gamma_1 t})$ for some $\gamma_1 > 0$. Otherwise put $Q = K_{EE}$, let Q^+ invert Q only on its range, and define

$$L_+ = (I - \Pi^H_{\mathrm{span}\, L_E}) L_B, \qquad K_+ = K_{BB} - K_{BE} Q^+ K_{EB} = L_+^\top H L_+.$$

This Gram is cone-coercive. With $x = y_B$,

$$\mathbf{1}^\top x = O(t^{-1}), \qquad (\log x)' = -K_+ x + O(t^{-2}),$$
$$L(y - p) = L_+ x + O(t^{-2}). \tag{46}$$

Thus $n = tx$, $s = \log t$, is positive and bounded and obeys the same input class with growth vector $\mathbf{1}$: $dn/ds = \mathrm{Diag}(n)(\mathbf{1} - K_+ n + O(e^{-s}))$.

Proof. **Optimization and the limiting split.** Cone coercivity gives $F(p) \leq \|g\|_\infty M - \alpha M^2/2$, $M = \mathbf{1}^\top p$, so the maximizing set is nonempty and compact. Concavity at the midpoint of two maximizers implies their difference is in $\ker K = \ker L$; hence all have the same aggregate and KKT gaps. Averaging finitely many maximizers gives one $\bar{p}$ positive on all of E. Along the bounded trajectory,

$$F' = W + O(e^{-\gamma t}), \quad W = \sum_i y_i (g - Ky)_i^2 \geq 0.$$

Integration proves that $F(y)$ has a limit and $\int W < \infty$. The velocity is bounded almost everywhere, so $W(y(t))$ is uniformly continuous and tends to zero. For any maximizer q, including zero,

$$\frac{d}{dt} D_h(q, y) = (y - q)^\top (g - Ky + e) \leq F(y) - F(q) + Ce^{-\gamma t}.$$

A suboptimal limiting value would force the nonnegative KL divergence to decrease at a fixed negative rate. Thus distance to the maximizing set vanishes. The same inequality bounds $D_h(\bar{p}, y)$, so every essential coordinate has a positive tail lower bound and all other coordinates tend to zero.

Let the columns of U span $\ker Q$. Positive semidefiniteness gives $L_E U = 0$, $K_{\mathrm{all},E} U = 0$, and KKT gives $U^\top g_E = 0$. Consequently the log leaf $l = U^\top \log y_E$ satisfies $l' = U^\top e_E$ and $l - l_\infty = O(e^{-\gamma t})$. Two cluster points p, p' would have the same aggregate and leaf, hence

$$(p_E - p'_E)^\top (\log p_E - \log p'_E) = 0, \qquad p_E - p'_E \in \ker Q.$$

Strict monotonicity of log proves $p = p'$. This split may depend on the integrated perturbation; it need not be the KL minimizer with original base ω. If E is empty this argument is omitted and $p = 0$. Strict slack now gives $y_N + |y'_N| = O(e^{-g_0 t})$ for some $g_0 > 0$.

Independent core coordinates and tracking. Near $(0, l_\infty)$, define a positive root $z(x, l)$ by

$$Qz + K_{EB} x = g_E, \qquad U^\top \log z = l. \tag{47}$$

The first equations are imposed only on $\mathrm{range}\, Q$ and are consistent there. A vector in the combined derivative kernel satisfies $Qu = 0$ and $U^\top \mathrm{Diag}(1/p_E) u = 0$; pairing with $u \in \mathrm{range}\, U$ proves $u = 0$. The implicit-function theorem therefore gives a smooth positive root with bounded local derivatives. Actual convergence supplies entry into this chart. On a frozen leaf and at fixed x, KL to this root has derivative $-(y_E - z)^\top Q (y_E - z)$. This controls KL locally: otherwise normalized differences would give a nonzero vector in the same combined kernel. KL is uniformly equivalent to squared Euclidean distance in the positive chart. For $d = \sqrt{D_h(z, y_E)}$, root motion and forcing give

$$D^+ d \leq -cd + C\|x'\| + Ce^{-g_1 t}, \qquad c, g_1 > 0. \tag{48}$$

Indeed $\partial_z D_h = \log(z/y_E) = O(d)$, $\|z'\| \le C(\|x'\| + \|l'\|)$, and the remaining vector-field forcing is $O(e + y_N)$. At zero distance use the integrated inequality, or regularize by $\sqrt{D_h + \varepsilon^2}$ and let $\varepsilon \downarrow 0$. For empty E, set $d = 0$.

Although the root split is nonlinear, its aggregate response is exact: $L_E(z - p_E) = -L_E Q^+ K_{EB} x$. Kernel terms vanish under L_E. The projection formula for L_+ follows, including dependent columns, and gives

$$L(y - p) = L_+ x + O(d + e^{-g_1 t}), \qquad (\log x)' = -K_+ x + O(d + e^{-g_1 t}). \tag{49}$$

Coercivity and the one-sided bootstrap. If $x \ge 0$, $x \ne 0$, and $L_+ x = 0$, write $L_B x = L_E v$. Perturb p by $-\varepsilon v$ on E and $+\varepsilon x$ on B. For sufficiently small positive ε, positivity on E preserves feasibility. The aggregate is unchanged; since the altered rows are tight, the linear objective is unchanged too. This constructs a maximizer positive on a B row, a contradiction. Compactness on the nonnegative unit simplex therefore proves

$$x^\top K_+ x \ge \alpha_+ (\mathbf{1}^\top x)^2, \qquad \alpha_+ > 0. \tag{50}$$

This includes empty E. The strict alternative gives a separator of the projected rows, and hence strict feasibility of the next primal.

For nonempty B, put $M = \mathbf{1}^\top x > 0$. Convergence and zero limiting brackets first give $\max_i |(\log x_i)'| \to 0$, $|M'|/M \to 0$, and $\|x'\| = o(M)$. For arbitrarily small $\varepsilon > 0$, $M(t) \ge M(t_\varepsilon) e^{-\varepsilon(t - t_\varepsilon)}$. Divide the integrated tracking inequality by $M(t)$. On a late interval $u \le t$, $M(u)/M(t) \le e^{\varepsilon(t-u)}$. Choosing $\varepsilon < \min(c, g_1)/2$ makes the old part exponentially negligible and bounds the recent convolution of $o(M)$ by $o(M(t))$. Thus $d = o(M)$ and $e^{-g_1 t} = o(M)$, before assuming a power law. Equations (49)–(50) give $M' \le -\alpha_+ M^2/2$ on a tail, hence $M = O(t^{-1})$. Now $\|x'\| = O(t^{-2})$; splitting the stable convolution at $t/2$ gives $d = O(t^{-2})$. This proves both errors in (46). Rescaling multiplies the multiplicative error by t, producing $O(e^{-s})$. All deeper changes use this same interpolant, without a new discrete interpolation approximation.

If B is empty, the tracking inequality has only exponential forcing. Its convolution is $O((1 + t)e^{-\min(c,g_1)t})$, which is $O(e^{-\gamma_1 t})$ after decreasing the positive exponent. The aggregate identity proves terminal convergence. If E is also empty, strict slack proves it directly; an initially empty system is vacuous. □

I.4 Finite termination, actual updates, and the terminal offset

Apply Lemma 12 to the actual interpolant in (40), with $t_2 = \log S$, $y_2 = S\xi_{B_0}$, $L_2 = L$, and $g_2 = \mathfrak{d}$. Its limiting aggregate gives $b = -\lambda v + L_2 p_2$. The next input set is $B_3 = B_2$, with equalities $A_3 = J \cup J_2$. Let $P^H_{A_q}$ project onto $T_q = \ker Z_{A_q}$. Define

$$L_q = P^H_{A_q} H^{-1} Z^\top_{B_q}, \qquad K_q = L_q^\top H L_q, \qquad \langle z, L_{q,i} \rangle_H = z_i^\top z \quad (z \in T_q).$$

Thus eliminating the essential columns restricts to $T_{q+1} = \ker Z_{A_q \cup E_q}$ in the *same* original metric. Cone coercivity supplies strict feasibility of (41). Strict convexity gives a unique primal; strong duality identifies it with $L_q p_q$, where p_q maximizes $\mathbf{1}^\top p - p^\top K_q p/2$. Multiplying the metric by A_1 does not change this primal minimizer.

For nonempty B_q, $q \ge 3$, this dual objective is positive at εe_i for small $\varepsilon > 0$. Its optimizer is nonzero, so $E_q \ne \varnothing$. For $i \in E_q$, complementarity gives $z_i^\top d^{(q)} = 1$, while $Z_{A_q} d^{(q)} = 0$. At least one independent equality is therefore added at each such level. Dimension strictly decreases; when it is zero, strict feasibility excludes any remaining weak rows. This proves the stated finite bound, including redundant constraints and singular Grams.

Here $\log^{(j)}$ denotes j-fold composition. Put $t_q = \log^{(q-1)} S$, so $dt_q/dS = 1/R_q$ and $R_{q+1} = R_q t_q$. At every nonterminal level the lemma gives

$$y_{q+1} = t_q (y_q)_{B_{q+1}}, \qquad L_q y_q = L_q p_q + L_{q+1} y_{q+1}/t_q + O(t_q^{-2}).$$

The next state is bounded and positive, with exponentially decaying forcing in its new clock. At the terminal level, $L_Q(y_Q - p_Q) = O(e^{-\gamma t_Q})$. Substitution into the actual queried-update interface

gives

$$h_t = a_* + b/S_t + \sum_{q=3}^{Q} d^{(q)}/R_q(S_t) + E(S_t),$$
$$\|E(S)\| \le C/S^2 + \sum_{q=2}^{Q-1} \frac{C}{R_q(S)t_q^2} + \frac{Ce^{-\gamma t_Q}}{R_Q(S)}. \tag{51}$$

Constants may depend on the fixed trajectory. The sum is empty at $Q = 2$. Multiplying by R_q after subtracting preceding terms proves (42): for an earlier error at level $l < q$, $R_q/(R_l t_l^2) = \prod_{i=l+1}^{q-1} t_i/t_l \to 0$. Later terms vanish at this scale too. No parameter expansion has been differentiated.

Each error in (51) is integrable in physical time: $dS/R_q = dt_q$, so its integral is bounded by a constant times

$$S^{-1} + \sum_{q=2}^{Q-1} t_q^{-1} + e^{-\gamma t_Q}.$$

These envelopes are eventually decreasing, so their fixed-step sums also converge absolutely. Primitive discretization errors converge because $|(1/R_q)'|$ is integrable on a late tail. Summing $w_{t+1} - w_t = \eta h_t$ therefore constructs a finite offset and an $o(1)$ residual after all terms. The finite initial prefix is included in the offset. The identities $a_* = \sigma u_r$ and $b = c_{r,\Lambda} - \lambda d_\sigma a_*$ replace the first two terms by the full clock with only a constant plus $O(\log S/S)$ difference. For every $q \ge 3$, $\log^{(q-1)} \widehat{S} - \log^{(q-1)} S \to 0$. This proves (13), including the terminal secondary case on degenerate primary faces.

Finally, matching (r, Λ) fixes H_r, J, J_2, B_3. Since $H = H_r/A_1$, every physical reduced Gram is A_1 times its unit Gram. At every higher level the growth is $\mathbf{1}$, so the entire physical optimal fiber is the unit fiber divided by A_1. Its essential and tight/slack sets are unchanged, and $L_{q,\mathrm{phys}}(\nu/A_1) = L_{q,\mathrm{unit}}\nu$. Induction preserves the depth and every selector. No extra factor accumulates per time change. Entry times, physical masses and offsets may differ.

J Complete calibration: proofs and sharp rate-only obstruction

This section derives trajectory comparisons from Theorems 3 and 11. All limits fix the compared configurations first. The claims use the specific effective clock (5), not an optimized time change. They do not assume an observed trajectory has entered its tail.

Proposition 13 (A nonzero logarithmic mismatch)**.** *For the row $z = (1/3, 1)$ at $r = 7/12$, let $u = (1/4, 1/2)$ and $v = (3/5, 4/5)$. Compare zero memory with (β, β^2), $0 < \beta < 1$, at a common sufficiently small fixed positive step. Calibrate their rates, but retain the same raw exponent $\lambda_0 \neq 0$. Writing $\rho = \beta e^{k\eta}$ gives $\chi = 1 - k\eta\rho/(1-\rho^2) \in (1/2, 1)$ and $\Delta\Lambda = \lambda_0(1-\chi) \neq 0$. With $D = W_0 - W_\beta$,*

$$D(s) = -\Delta\Lambda v \log s + O(1), \qquad \frac{s}{\log s}\|N(W_0(s)) - N(W_\beta(s))\| \longrightarrow \frac{8}{25}|\Delta\Lambda| > 0. \tag{52}$$

Choosing $\lambda_\beta = \lambda_0/\chi$ instead makes $D(s)$ converge.

Relative direction at matched coordinates. Under Theorem 1, matched coordinates give $d_\infty = w_{\infty,a} - w_{\infty,b}$ and

$$s\big[N(W_a(s)) - N(W_b(s))\big] \longrightarrow \frac{\Pi_{u_r} d_\infty}{\|u_r\|}. \tag{53}$$

Thus relative direction error is $O(s^{-1})$. The same directional observation can be class-centered in softmax; the full-parameter iff remains the one stated in the main theorem.

J.1 A TRANSVERSE OBSTRUCTION TO SCALAR RETIMING

Corollary 14 (Transverse logarithmic drift survives retiming)**.** *Consider two individually eligible runs on the same learning problem with common r, and put $u = u_r \neq 0$, $\Delta c = c_a - c_b$. For every eventually nondecreasing map τ into the second run's legal clock tail with $\tau(s) \to \infty$ and $\tau(s)/s \to 1$,*

$$\frac{\Pi_u[W_a(s) - W_b(\tau(s))]}{\log s} \longrightarrow \Pi_u \Delta c.$$

If $\Pi_u \Delta c \neq 0$, no eventually nondecreasing retiming into that tail, with $\tau(s) \to \infty$, makes the parameter difference bounded.

Proof. The finite hierarchy and its parameter interpolation give $W_x(t) = tu + c_x \log t + o(\log t)$; every deeper iterated logarithm is $o(\log t)$. Projection removes $(s - \tau(s))u$ exactly. Since $\log \tau(s)/\log s \to 1$, the projected expansion gives the stated limit, including its composed remainder. This argument is for each fixed map τ and does not assert a uniform convergence rate over maps.

For the second assertion, suppose the retimed difference is bounded. The leading laws give $\|W_a(s)\|/s \to \|u\|$ and $\|W_b(\tau(s))\|/\tau(s) \to \|u\|$. Bounded difference forces the ratio of these diverging norms to tend to one, hence $\tau(s)/s \to 1$. The first assertion then contradicts boundedness when $\Pi_u \Delta c \neq 0$. No remainder has been differentiated. □

For the nonradial example above, $\Delta c = -\Delta\Lambda v$ and $\Pi_u v = (4/25, -2/25)$, so every nonzero shape mismatch has a transverse obstruction. In contrast, the row $z = (1, 1)$ has $u = (r/2, r/2)$ and $c = -\Lambda(1/2, 1/2)$. Here $\Delta c = \alpha u$, and $\tau(s) = s + \alpha \log s$ cancels the logarithmic mismatch, leaving a convergent finite difference by the regular expansion. Thus the obstruction concerns nonzero transverse mismatch; it does not replace the prescribed-clock classification with a classification under arbitrary time changes.

J.2 CLOCK ALIGNMENT AT FIXED POSITIVE STEPS

For one admissible run put $s_t = \widehat{S}(S_t)$, and define

$$P(s) = su_r + c_{r,\Lambda} \log s + \sum_{q=3}^{Q} d^{(q)} \log^{(q-1)} s.$$

All iterated logarithms below are restricted to a sufficiently late tail. The full hierarchy theorem says $w_t = P(s_t) + w_\infty + e_t$, $e_t \to 0$. Exactly,

$$s_{t+1} - s_t = \sigma\eta - \lambda(\partial_k \sigma) \log\left(1 + \frac{\eta}{S_t + s_0}\right) \longrightarrow \sigma\eta > 0.$$

Choose a tail on which the mesh is positive and bounded above and below. For $s = (1-\alpha)s_t + \alpha s_{t+1}$, $0 \leq \alpha \leq 1$, set $W(s) = (1-\alpha)w_t + \alpha w_{t+1}$. For this fixed finite hierarchy, $P''(s) = O(s^{-2})$: the leading linear term has zero second derivative, and every iterated logarithm has second derivative bounded by a constant times s^{-2} on a late tail. The error of its linear interpolant is consequently

$$\|(1-\alpha)P(s_t) + \alpha P(s_{t+1}) - P(s)\| \leq C(s_{t+1} - s_t)^2 \sup_{v \in [s_t, s_{t+1}]} v^{-2} = O(s^{-2}).$$

The interpolated residual is at most $\max\{\|e_t\|, \|e_{t+1}\|\} \to 0$. Thus

$$W(s) = P(s) + w_\infty + o(1). \tag{54}$$

This step is essential for the finite-limit claim: rounding an inverse clock to its nearest query can retain a bounded grid-phase oscillation of size $O(\eta)$, which need not vanish as $s \to \infty$ at fixed step. The interpolation is for observation only. At every query it equals the unaltered algorithm output.

J.3 Recovering the effective parameters from the selectors

For $r > 0$, the Burg minimizer is nonzero and obeys

$$\min_i z_i^\top u_r = r. \tag{55}$$

Indeed, if this minimum were $m > r$, then $(r/m)u_r$ would remain feasible and inside the barrier box. The scalar function $-x - \log(1-x)$ is strictly increasing on $(0,1)$, so contraction along a nonzero ray strictly decreases the objective, a contradiction. This is the usual radial tightness property of the Burg program (Wang et al., 2026); we use it here to identify the effective rate. In particular, writing $G = \max_i \|z_i\| > 0$,

$$|r_a - r_b| \le G\|u_{r_a} - u_{r_b}\|. \tag{56}$$

This follows from the Lipschitz property of the minimum of the finitely many linear functions $z_i^\top u$.

The essential primary set J is nonempty. Otherwise the selected multiplier would vanish, and KKT stationarity would give $\nabla B(u_r) = 0$. Since $\nabla B(u)_j = u_j/(1-|u_j|)$, this would force $u_r = 0$, contradicting $r > 0$. At a fixed rate, the same J applies to all shapes. The equalities in (10) therefore give, for any $j \in J$,

$$z_j^\top [c_{r,\Lambda_a} - c_{r,\Lambda_b}] = -(\Lambda_a - \Lambda_b), \tag{57}$$

$$\|c_{r,\Lambda_a} - c_{r,\Lambda_b}\| \ge \frac{|\Lambda_a - \Lambda_b|}{\|z_j\|}. \tag{58}$$

Strict separability excludes zero margin rows. These facts require neither an independent full set of active rows nor a differentiable primary path through a face change. They identify the parameters from the full vectors, not from their normalized directions.

J.4 Proof of the complete-invariant theorem

Apply (54) to each run. If their (r, Λ) agree, the hierarchy theorem identifies the same primal selectors, whole-fiber sets, and depth in the two clocks. Subtracting all terms gives

$$D(s) = w_{\infty,a} - w_{\infty,b} + o(1).$$

Thus matching implies a finite limit, and a finite limit implies boundedness. Conversely, suppose $D(s) = O(1)$. Since each correction to the leading term is $O(\log s)$, division by s gives $u_{r_a} = u_{r_b}$. Equation (55) then gives $r_a = r_b =: r$. Dividing the remaining difference by $\log s$ gives $c_{r,\Lambda_a} = c_{r,\Lambda_b}$, since all higher iterated logarithms are $o(\log s)$. Equation (57) and $J \neq \varnothing$ give $\Lambda_a = \Lambda_b$. This proves all equivalences in (7).

The same division argument proves the two mismatch limits. The nonzero limiting vectors have quantitative lower bounds

$$\|u_{r_a} - u_{r_b}\| \ge |r_a - r_b|/G, \qquad \|c_{r,\Lambda_a} - c_{r,\Lambda_b}\| \ge |\Lambda_a - \Lambda_b|/\|z_j\| \quad (j \in J).$$

Hence different rates force linear full-parameter separation, and equal rates with different shapes force logarithmic full-parameter separation, in these fixed clocks. The latter may be radial, so it need not force a nonzero angular coefficient. Neither lower bound supplies a common onset as the mismatch tends to zero.

Apply (24) with $a(s) = 1$ and $D(s) \to d_\infty$ to obtain (53). If $\Pi_{u_r} d_\infty = 0$, the directional difference is $o(s^{-1})$. For any fixed linear observation map C with $Cu_r \neq 0$, the same proof yields

$$s[N(CW_a(s)) - N(CW_b(s))] \longrightarrow \frac{\Pi_{Cu_r} C d_\infty}{\|Cu_r\|}. \tag{59}$$

Class-centering for native softmax has this property by separability. The converse in (7) concerns full parameters; we do not infer it from equality of observed limiting directions.

At equal rates but different shapes, use (24) with $a(s) = \log s$ and $v = c_{r,\Lambda_a} - c_{r,\Lambda_b}$ to obtain

$$\frac{s}{\log s}[N(W_a(s)) - N(W_b(s))] \longrightarrow \frac{\Pi_{u_r}[c_{r,\Lambda_a} - c_{r,\Lambda_b}]}{\|u_r\|}. \tag{60}$$

The projected numerator can vanish. For example, the one-row geometry $z = (1,1)$ has $u_r = (r/2, r/2)$ and $c_{r,\Lambda} = (-\Lambda/2, -\Lambda/2)$ for $0 < r < 2$. All primary directions agree even when the complete invariants differ.

Proof of the offset-free diagnostic. For equal rates, pad the shorter of the two finite hierarchies with zero coefficients. Their difference is

$$D(s) = \Delta c \log s + \sum_{q=3}^{Q_{\max}} \Delta d^{(q)} \log^{(q-1)} s + \Delta w_\infty + o(1).$$

For fixed $\gamma > 1$, $\log(\gamma s) - \log s = \log \gamma$, whereas $\log^{(q-1)}(\gamma s) - \log^{(q-1)} s \to 0$ for every $q \geq 3$. The constant cancels, and the difference of the two $o(1)$ terms still tends to zero. This proves (14), even if the shapes have different hierarchy depths. No derivative of an asymptotic residual is taken. The claim fixes $\gamma > 1$; it does not cover a shrinking window $\gamma \to 1$ jointly with time. No monotonic decay or computable finite-budget tolerance is implied.

J.5 Exact proof of the rate-only counterexample

Take $z = (1/3, 1)$, $r = 7/12$, and positive loss weight one. Appendix H.3 computes the unique primary selector $u = (1/4, 1/2)$, its positive multiplier, and the secondary selector $c_{r,\Lambda} = -\Lambda v$, where $v = (3/5, 4/5)$. The single active row is independent and has a positive multiplier, so its KKT branch is regular. Appendix H.1 therefore supplies the finite-offset expansion without further iterated logarithms. In particular,

$$W(s) = su - \Lambda v \log s + w_\infty + o(1). \tag{61}$$

For the fixed pair $(0, 0)$, $\sigma = \chi = 1$, so rate calibration sets $k_0 = r$. For (β, β^2), choose the unique nearby root $k_\beta/\sigma(k_\beta, \eta) = r$ from Theorem 11. A common sufficiently small positive step works for these two fixed memory pairs and both compared shape choices; all initial parameters are finite and all moments start at zero. With $\rho = \beta e^{k_\beta \eta} \in (0, 1)$,

$$\rho_2 = \rho^2, \qquad d_\sigma = \eta \left(\frac{\rho}{1-\rho} - \frac{\rho^2}{1-\rho^2} \right) = \frac{\eta \rho}{1-\rho^2} > 0.$$

After decreasing the common step if needed, $1/2 < \chi = 1 - k_\beta d_\sigma < 1$. This is an analytic inequality, not a numerical evaluation of an unknown global threshold. Keeping $\lambda = \lambda_0 \neq 0$ in both runs gives $\Lambda_0 = \lambda_0$, $\Lambda_\beta = \lambda_0 \chi$, and $\Delta\Lambda = \lambda_0(1 - \chi) \neq 0$. Subtracting (61) proves $D(s) = -\Delta\Lambda v \log s + O(1)$. Finally,

$$\Pi_u v = (4/25, -2/25), \qquad \|u\| = \sqrt{5}/4, \qquad \frac{\|\Pi_u v\|}{\|u\|} = \frac{8}{25}.$$

Equation (60) proves the claimed nonzero angular limit. Choosing $\lambda_\beta = \lambda_0/\chi$ matches both invariants, so Theorem 1 instead applies. This construction works for EXP or logistic loss, under either allowed bias convention. Both filters of the nonzero-memory run are nonzero; the comparison is not limited to a first-moment-only method.

An exactly dyadic margin row. Rescale the example by $3/4$. Then $z' = (1/4, 3/4)$ and $r' = 7/16$ have the same u, H, while $c'_{r',\Lambda} = -\Lambda(4/5, 16/15)$. Indeed the secondary constraint is the old one with right side multiplied by $4/3$; its minimum-norm solution scales by the same factor. The angular coefficient in (52) is consequently $(32/75)|\Delta\Lambda|$ after recomputing the controller at r'. This gives a rational example whose input row and target rate have exact binary encodings. The nonzero-memory controller still solves a nonlinear scalar equation; dyadic data do not make its rate root exactly representable.

J.6 Local independence and the size of the correction

For fixed (η, β_1, β_2), define $\mathcal{C}(k, \lambda) = (k/\sigma(k), \lambda\chi(k))$. On a legal interval with $\chi > 0$,

$$D\mathcal{C} = \begin{pmatrix} \chi/\sigma & 0 \\ \lambda \partial_k \chi & \chi \end{pmatrix}, \qquad \det D\mathcal{C} = \chi^2/\sigma > 0.$$

The inverse function theorem makes these independent local schedule coordinates. The controller theorem supplies the nearby admissible inverse; this determinant alone does not prove global invertibility past a pole. The concrete counterexample proves that the specific rate-only rule cannot

universally preserve the next-order geometry. It is not an information or topological lower bound on arbitrary scalar parametrizations.

One can also read the size of the correction analytically. For a fixed memory pair put $M_\beta = \beta_1/(1-\beta_1) - \beta_2/(1-\beta_2)$. Taylor expansion of the explicit EMA factors at $\eta = 0$, uniformly for k on a compact interior interval, gives

$$\sigma = 1 + kM_\beta\eta + O(\eta^2), \qquad d_\sigma = M_\beta\eta + O(\eta^2), \qquad \chi = 1 - kM_\beta\eta + O(\eta^2).$$

Consequently the nearby inverse controller at fixed (r, Λ) is

$$k = r + r^2 M_\beta\eta + O(\eta^2), \qquad \lambda = \Lambda\,[1 + rM_\beta\eta] + O(\eta^2).$$

These expansions follow by substituting the first expression into $r = k/\sigma$, then expanding $1/\chi$. The constants depend on the fixed memory pair; no uniform expansion as a beta tends to one is asserted. Implementations requiring exact calibration must solve the original scalar equations rather than equating a truncated expansion to their exact roots.

J.7 Why the classification depends on the schedule family

The canonical schedule fixes which divergent scales can vary independently. A conditional regular-face extension makes this dependence explicit. Retain the deterministic full-batch data, loss, bias convention and fixed eligible step, but write

$$\epsilon_t = \epsilon_0 e^{-kS_t + F_t}, \qquad \delta_t = (F_t - F_{t+1})/\eta \to 0, \quad \sum_t \eta\delta_t^2 < \infty, \quad \sum_t |\delta_{t+1} - \delta_t| < \infty.$$

Assume a regular rate neighborhood: the tight set is fixed, every tight row has a positive selected multiplier, and inactive slacks are strict. In addition, assume that the normalized trajectory is captured on a late tail in the stable positive-leaf neighborhood of this branch. This local capture is an extra hypothesis for F_t, not a consequence of the canonical from-start theorem. Work where $\chi > 0$.

Conditional normal form. Let $a(K) = \sigma(K)u_{K/\sigma(K)}$ and $v_r = u_r'$. Regular KKT differentiation (Corollary 15) gives $Z_J v_r = \mathbf{1}$, $v_r \neq 0$, and $a'(k) = \sigma' u_r + \chi v_r$. Then

$$\begin{aligned} w_t &= S_t a(k) - a'(k)F_t + b_\infty + o(1), \\ s_t &= \sigma S_t - \sigma' F_t, \qquad W(s) = su_r - v_r\Psi(s) + b_\infty + o(1), \end{aligned} \tag{62}$$

where W and Ψ linearly interpolate (s_t, w_t) and $(s_t, \chi F_t)$, respectively. These nodes and profiles are prescribed by the configuration.

Proof. The exact normalized recurrence uses the ratio $\epsilon_t/\epsilon_{t+1} = e^{\eta(k+\delta_t)}$. On the captured positive leaf, the local contraction used in Appendix H.1 is uniform in a small rate neighborhood. Comparing first to the equilibrium at $k + \delta_t$, then to that at $k + \delta_{t+1}$, yields a graph distance satisfying

$$d_{t+1} \le qd_t + C|\delta_{t+1} - \delta_t| + Ce^{-cS_t}, \qquad \|h_t - a(k+\delta_t)\| \le Cd_t + Ce^{-cS_t}, \quad q < 1.$$

Strict inactive slacks give exponential inactive-atom decay. Telescoping F_t and $\delta_t \to 0$ give $F_t = o(S_t)$, so true-loss and bias-correction terms are also exponentially small on this captured tail. Summing the contraction inequality gives $\sum_t d_t < \infty$. The branch is $C^{1,1}$, hence $a(k+\delta_t) = a(k) + a'(k)\delta_t + O(\delta_t^2)$. The increment of $w_t - S_t a(k) + a'(k)F_t$ is $\eta[h_t - a(k) - a'(k)\delta_t]$, which is absolutely summable. This proves the finite offset in (62).

The branch derivative gives its effective-clock form. Since $s_t/S_t \to \sigma > 0$ and $s_{t+1} - s_t = \eta(\sigma + \sigma'\delta_t) \to \eta\sigma > 0$, the nodes increase eventually. Parameter interpolation preserves the affine term and the defined Ψ exactly, and preserves convergence of the residual. For two captured regular runs at the same r,

$$W_a - W_b = -v_r(\Psi_a - \Psi_b) + \Delta b_\infty + o(1).$$

Because $v_r \neq 0$, boundedness or convergence of this difference is equivalent to the corresponding property of $\Psi_a - \Psi_b$. If rates differ, $\Psi = o(s)$ and (55) exclude bounded difference.

Several logarithmic amplitudes. Set $\mathcal{L}_1(S) = \log(1 + S/s_0)$, $\mathcal{L}_{j+1}(S) = \log(1 + \mathcal{L}_j(S))$ and $F(S) = \sum_{j=1}^{m} \lambda_j \mathcal{L}_j(S)$ for fixed finite m. Here $F' = O(S^{-1})$ and $F'' = O(S^{-2})$; exact cell-averaged slopes satisfy the required square-summability and total-variation bounds. Since $s = \sigma S + O(\log S)$, $\mathcal{L}_1(S) - \mathcal{L}_1(s) \to -\log \sigma$ and $\mathcal{L}_j(S) - \mathcal{L}_j(s) \to 0$ for $j \geq 2$. The first constant changes only the offset. Thus the matching quantities are $(r, \chi\lambda_1, \ldots, \chi\lambda_m)$: subtracting the two expansions proves sufficiency, and division by the first unmatched divergent scale proves necessity. For $m = 1$ these reduce to the canonical rate and shape.

Bounded difference without convergence. For $F(S) = \sin \log(1 + S/s_0)$, the same derivative bounds verify the profile conditions. Against a captured $F = 0$ run at the same r,

$$\Psi(s) = \chi \sin[\log s - \log(\sigma s_0)] + o(1).$$

The nonzero vector v_r therefore gives a bounded, nonconvergent parameter difference. Both stabilizers are eventually decreasing because $k - F'(S) > 0$ on a tail. This conditional example explains why the canonical bounded-if-and-only-if-convergent statement is specific to its schedule family.

K GEOMETRIC MEANING OF THE MEMORY CORRECTION

This section connects the regular-face shape coefficient to the small-step sensitivity of the exact fixed-step branch. It is a static consequence of the geometric programs, not a variable-step trajectory theorem.

Corollary 15 (Rate sensitivity and the first memory deformation)**.** *Suppose the Burg selector is on a regular rate neighborhood: its tight set J is locally constant, all tight rows admit a strictly positive multiplier, and the other rows have strict slack. Put $M = Z_J$, $H_r = \nabla^2 B(u_r)$, and*

$$P_r = I - H_r^{-1} M^\top (M H_r^{-1} M^\top)^\dagger M.$$

Then the selector is continuously differentiable and

$$u_r' = \frac{(I - P_r)u_r}{r}, \qquad c_{r,\Lambda} = -\Lambda u_r'. \tag{63}$$

Now hold the raw rate k and memory fixed and let the step tend to zero. For $d_\ell = \beta_\ell/(1 - \beta_\ell)$, the physical branch $a_*(k, \eta) = \sigma(k, \eta) u_{k/\sigma(k,\eta)}$ satisfies

$$\begin{aligned} a_*(k, \eta) &= u_k + \eta v_k + o(\eta), \\ v_k &= k(d_1 - d_2) P_k u_k = -k^3 (d_1 - d_2) \frac{d}{dk} \left(\frac{u_k}{k} \right). \end{aligned} \tag{64}$$

In particular $v_k = 0$ exactly when $\beta_1 = \beta_2$ or $P_k u_k = 0$. This is a pointwise first-order statement, not equality of the fixed-step branches to all orders.

Proof. Choose independent rows of M spanning its row space, denoted by $\widetilde{M}$. Writing $M = C\widetilde{M}$, the active equality at a positive rate gives $C\mathbf{1} = \mathbf{1}$; the reduced equalities therefore imply all active equalities at nearby rates. The equality-constrained KKT Jacobian

$$\begin{pmatrix} H_r & -\widetilde{M}^\top \\ \widetilde{M} & 0 \end{pmatrix}$$

is invertible: a null vector (a, b) has $\widetilde{M}a = 0$ and $a^\top H_r a = 0$, hence $a = b = 0$. The implicit-function theorem gives a differentiable branch. A positive full multiplier remains positive locally by a continuous lift through $M^\top$; strict inactive slacks persist. Convex KKT optimality identifies this branch with u_r.

Differentiating gives $Mu_r' = \mathbf{1}$ and $H_r u_r' \in \operatorname{range}(M^\top)$. Thus $P_r u_r' = 0$. Both u_r' and $(I - P_r)u_r/r$ belong to the H_r-orthogonal complement of $\ker M$ and solve $Mx = \mathbf{1}$; they coincide. On a regular face the secondary program has only the equality $Mc = -\Lambda\mathbf{1}$, so its minimum-energy solution is $-\Lambda u_r'$.

Expanding the exact factors at zero step gives

$$A_1 = 1 + kd_1\eta + O(\eta^2), \quad A_2 = 1 + 2kd_2\eta + O(\eta^2), \quad \sigma = 1 + k(d_1 - d_2)\eta + O(\eta^2).$$

The chain rule for $\sigma u_{k/\sigma}$ yields $v_k = k(d_1 - d_2)(u_k - ku_k')$. The first identity then gives $u_k - ku_k' = P_k u_k$, proving both expressions for v_k. Since $k > 0$ and $\beta \mapsto \beta/(1-\beta)$ is injective, the stated zero condition follows. □

The two appearances of rate sensitivity concern different comparisons: $-\Lambda u_r'$ is a logarithmic trajectory coefficient at one eligible fixed step, whereas v_k differentiates a family of stationary branches with respect to the step. No derivative of a finite-time trajectory, exchange of the time and step limits, or computable global entry threshold is asserted.

K.1 Direction equivalence is a weaker comparison

The complete calibration theorem concerns full parameter vectors, not merely their unit directions. The following geometric criterion makes the distinction explicit without adding a new dynamical assumption.

Proposition 16 (Rates with the same limiting direction)**.** *Fix $r_0 \in (0, \gamma_\infty)$ and write $v_0 = u_{r_0}/r_0$ and $A_0 = \{i : z_i^\top v_0 = 1\}$. For $r \in (0, \gamma_\infty)$,*

$$N(u_r) = N(u_{r_0}) \iff \begin{array}{c} r\|v_0\|_\infty < 1 \quad \textit{and} \\ \nabla B(rv_0) \in \operatorname{cone}\{z_i : i \in A_0\}. \end{array} \tag{65}$$

When the condition holds, $u_r = rv_0$. Thus distinct eligible rates can preserve the limiting direction while failing full-trajectory equivalence.

Proof. By radial tightness (55), equality of directions forces $u_r = rv_0$, whose active set is A_0. The open-box condition and convex KKT stationarity give the right-hand side of (65). Conversely that condition supplies nonnegative active multipliers for the feasible point rv_0; strict convexity makes it the unique minimizer. If $r \neq r_0$, Theorem 1 gives $[W_r(s) - W_{r_0}(s)]/s \to (r - r_0)v_0 \neq 0$. For example, with the single row $z = (1, 1)$, symmetry and the active margin constraint give $u_r = (r/2, r/2)$ for every $0 < r < 2$. The normalized direction is identical throughout this interval, although the full trajectories at distinct rates separate linearly in the prescribed clocks. □

L Reproducibility record

The anonymous artifact contains execution scripts, frozen protocols, locked figure inputs, row-level tables and manifests for the reported results. Large raw trajectories are indexed by the accompanying manifests; the manuscript gives the definitions and numerical summaries needed to interpret each figure.

M Calibration obstruction and learned-feature evidence

M.1 Actual rate-only obstruction

The configuration, tail regression and clock comparisons for the RATE/TWO obstruction are collected in Appendix Q.2. The next subsection tests the same calibration principle on frozen learned representations.

M.2 Learned representations and frozen comparison protocol

We use the fixed 45,000/5,000 split of CIFAR-10's official training set (Krizhevsky, 2009). A CIFAR-stem ResNet-18 (He et al., 2016) with a bias-free $512 \to 10$ head is trained with SGD and momentum. Its checkpoint is selected by the predeclared training-margin rule under deterministic evaluation. The encoder is frozen before the Adam comparisons. Features, labels, separating head and encoder checkpoint are hash-identified. H3/H4 are representation seeds 3/4, with independently checked

minimum pairwise training margins 5.413694 and 5.896547 and no nonpositive margin. They share the underlying data and are not independently collected datasets.

The seed-0 representation and development memory pool are used for adaptive search over step, schedule, initialization and analysis. All final choices are frozen before H3/H4 outcomes and eight new memory pairs are run. A midpoint-grid rule selects four settings inside and four outside the development rule $\beta_1 \le 0.9$. These eight deliberately stratified settings are not a random sample of ordinary Adam hyperparameters.

Optimizer and stopping time. The source memory is $(0.9, 0.999)$; the step is half of $0.009471374649\ldots$, with the source effective rate and shape targets, $\epsilon_0 = 10^{-3}$, $s_0 = 300$ and initial head 0.03 times the separating witness. Moments are uncorrected and start at zero. All arms use the same features and initial head, deterministic full-batch mean cross-entropy, and no weight decay, label smoothing, dropout or augmentation. RATE matches r while retaining the source raw shape; TWO matches (r, Λ). Each arm stops at its first integer update T satisfying $\widehat{S}((T-1)\eta) < 400 \le \widehat{S}(T\eta)$, about $84{,}400$ updates. This aligns effective time, not wall-clock cost. Static separability does not certify the selected step against the theorem's existential threshold.

For the class-centered difference

$$D_x(s) = C[W_{\mathrm{REF}}(s) - W_x(s)],$$

we use the offset-free diagnostic motivated by (14),

$$Q_x(s) = \left\| \frac{D_x(\gamma s) - D_x(s)}{\log \gamma} \right\|_F, \qquad A = 1 - \frac{M_{\mathrm{TWO}}}{M_{\mathrm{RATE}}}, \tag{66}$$

where $\gamma = 1.25$ and M is the median over the frozen tail windows. Positive A means that TWO leaves less residual trajectory mismatch than RATE under this diagnostic.

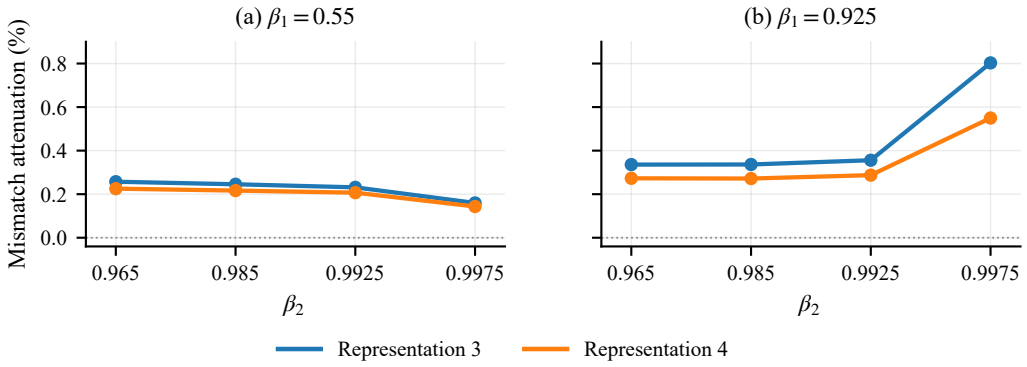

Figure 4: Trajectory-mismatch reduction on two frozen representations (seeds 3 and 4). The horizontal axis lists discrete memory settings at equal spacing; lines connect these settings. Float64 replay preserves every sign, with mean 0.3059% versus 0.3062% in float32.

Frozen measurement. At 64 logarithmically spaced legal left endpoints, piecewise-linear interpolation uses each arm's own clock and no extrapolation. Class-centering is performed separately at each window. Equation (66) uses $\gamma = 1.25$ and the last $\max\{3, \lceil 0.1 \times 64 \rceil\} = 7$ windows. Their median Q_x is M_x; $A = 1 - M_{\mathrm{TWO}}/M_{\mathrm{RATE}}$ is multiplied by 100 when reported as percent attenuation.

Numerical precision. Float64 re-integration preserves the sign in all 16 held-out cells and gives mean attenuation 0.3059%, compared with 0.3062% in float32. Repeat and storage-density checks are provided in the artifact.

Uncertainty and interpretation. Memory-pair bootstrap averages each pair over H3/H4 before resampling. The original mean's 95% interval is approximately $[0.2231, 0.4220]\%$; the float64

Table 2: Every held-out cell: percent attenuation of the frozen trajectory diagnostic. Float64 trajectories are separately integrated.

β_1	β_2	H3 float32	H3 float64	H4 float32	H4 float64
0.55	0.965	0.25726	0.25740	0.22534	0.22425
0.55	0.985	0.24552	0.24361	0.21650	0.21483
0.55	0.9925	0.23134	0.23042	0.20692	0.20517
0.55	0.9975	0.15974	0.15823	0.14334	0.14339
0.925	0.965	0.33584	0.33709	0.27293	0.27081
0.925	0.985	0.33634	0.33715	0.27182	0.27090
0.925	0.9925	0.35593	0.35931	0.28741	0.28330
0.925	0.9975	0.80293	0.80772	0.54973	0.55032

interval is $[0.2224, 0.4224]\%$. With two representation seeds, eight selected memory pairs and shared references, these intervals describe uncertainty conditional on this study design. On the original trajectories, trimmed-mean dilation gives mean attenuation 0.2902% and a Huber fit of $\|D(s)\|$ against $\log s$ gives 0.0224%; all signs remain positive, but the latter measures a different quantity. The frozen memory rule does not transfer: inside/outside means are $0.2107\%/0.4016\%$. These experiments support a small finite-window reduction in trajectory mismatch on frozen learned representations.

N Risk-aware calibration: complete statements and proofs

The risk statements below inherit the actual primary-selection and finite-hierarchy theorems proved earlier. Controlled diagnostics appear in Appendix Q.3.

N.1 Inherited hypotheses and the observation convention

Fix a finite separable margin system, its grouping into examples, and strictly positive example weights. The loss R is either weighted exponential loss or the covered grouped logistic/softmax loss. For a run x, write

$$\begin{aligned} S_{x,n} &= \eta_x n, & \epsilon_x(S) &= \epsilon_{0,x} e^{-k_x S}(1 + S/s_{0,x})^{\lambda_x}, \\ \rho_{\ell,x} &= \beta_{\ell,x} e^{\ell k_x \eta_x} < 1, & A_{\ell,x} &= \frac{1-\beta_{\ell,x}}{1-\rho_{\ell,x}}, \\ \sigma_x &= A_{1,x}/\sqrt{A_{2,x}}, & r_x &= k_x/\sigma_x, \\ \chi_x &= 1 - k_x \partial_k \log \sigma_x, & \Lambda_x &= \lambda_x \chi_x, \\ s_x(S) &= \sigma_x S - \lambda_x \sigma_x' \log(1 + S/s_{0,x}). \end{aligned} \tag{67}$$

Here $w_{x,n+1}$ is the parameter after n completed updates, and $\sigma_x' = \partial_k \sigma_x$ with memory and step fixed. Each compared run is individually eligible under the manuscript's actual-selection and finite-hierarchy theorems; no new uniform step bound is asserted. $W_x(s)$ is the linear interpolation of *parameters* at the actual nodes $(s_x(S_{x,n}), w_{x,n+1})$ on a common increasing tail.

The inherited primary theorem gives, for the normalized reference atoms,

$$\xi_{i,x,n} := \frac{\omega_i e^{-z_i^\top w_{x,n+1}}}{\epsilon_x(S_{x,n})} \longrightarrow p_i(r_x)/\sqrt{A_{2,x}}. \tag{68}$$

For grouped losses, i here denotes an expanded margin row and ω_i its example weight. The support J of the KL-selected multiplier $p(r)$ is the union of positive supports over the optimal dual fiber (Lemma 6). Its mass is

$$\mu(r) := \mathbf{1}^\top p(r) = \frac{u_r^\top \nabla B(u_r)}{r} > 0. \tag{69}$$

The equality follows from stationarity $Z^\top p = \nabla B(u_r)$ and complementarity; positivity follows since $u_r \neq 0$ and $u_r^\top \nabla B(u_r) = \sum_j u_{r,j}^2/(1 - |u_{r,j}|) > 0$.

From reference mass to risk. At any parameter state and any positive normalizer ϵ, let $m = \sum_i \omega_i e^{-z_i^\top w}/\epsilon$ and, for grouped loss, $\tau = \max_j \sum_{i\in j} e^{-z_i^\top w}$. Then

$$\frac{\epsilon m}{1+\tau} \le R(w) \le \epsilon m. \tag{70}$$

This follows by summing $X/(1+X) \le \log(1+X) \le X$ over the positive-weight groups. For exponential loss set $\tau = 0$; both bounds are equalities. We use this bound for both asymptotic and finite-node results.

N.2 The missing risk amplitude and a complete relative-risk criterion

Define the positive, clock-anchored amplitude

$$\Gamma_x := \frac{\epsilon_{0,x}}{\sqrt{A_{2,x}}(\sigma_x s_{0,x})^{\Lambda_x}}. \tag{71}$$

This training-risk amplitude is defined relative to the prescribed clock origin.

Lemma 17 (Interpolation preserves normalized atom limits)**.** *Let $S_x(s)$ be the tail inverse of $s_x(S)$. The atoms $\xi_{i,x}(s) = \omega_i e^{-z_i^\top W_x(s)}/\epsilon_x(S_x(s))$ converge to the same limits as in* (68)*. Moreover,*

$$\epsilon_x(S_x(s)) = \sqrt{A_{2,x}}\Gamma_x e^{-r_x s} s^{\Lambda_x} \left(\frac{\sigma_x(S_x(s)+s_{0,x})}{s}\right)^{\Lambda_x}. \tag{72}$$

The last factor tends to one.

Proof. Drop x. Since $r\sigma' = k\partial_k \log\sigma = 1-\chi$, the clock identity gives exactly $-kS + \lambda\log(1 + S/s_0) = -rs + \Lambda\log(1+S/s_0)$. This proves (72). Also $s/S \to \sigma$. On the increasing tail, $S'(s)$ is bounded and $S''(s) = O(S^{-2})$ by differentiating the explicit clock twice. Thus $F(s) := \log\epsilon(S(s)) = -rs + \Lambda\log(1+S(s)/s_0)$ obeys $F''(s) = O(s^{-2})$. The effective node spacing tends to $\sigma\eta > 0$ and is bounded above. For $s = (1-a)s_n + as_{n+1}$, $0 \le a \le 1$,

$$\log\xi_i(s) = (1-a)\log\xi_{i,n} + a\log\xi_{i,n+1} + (1-a)F(s_n) + aF(s_{n+1}) - F(s).$$

The last term is $O(s_n^{-2})$, uniformly within the cell. Atoms whose limit is positive therefore converge by taking logarithms. If both endpoint atoms tend to zero, their geometric interpolation is bounded above by their maximum, times a factor tending to one, and also tends to zero. This covers weak and inactive rows without taking a logarithm of a zero limit. □

Theorem 18 (Complete relative-training-risk calibration)**.** *For every eligible run,*

$$R(W_x(s)) = \mu(r_x)\Gamma_x e^{-r_x s} s^{\Lambda_x}(1+o(1)). \tag{73}$$

For two eligible runs on the same margin system, loss and weights,

$$\frac{R(W_a(s))}{R(W_b(s))} \longrightarrow 1 \quad\Longleftrightarrow\quad (r_a,\Lambda_a,\Gamma_a) = (r_b,\Lambda_b,\Gamma_b). \tag{74}$$

More generally, their risk ratio is bounded above and away from zero on a tail if and only if $(r_a,\Lambda_a) = (r_b,\Lambda_b)$. In that case its limit is Γ_a/Γ_b.

Proof. Lemma 17 gives $m \to \mu(r)/\sqrt{A_2}$ and $\tau \to 0$ (with $\tau = 0$ for EXP). Equation (70) therefore yields $R/\epsilon \to \mu(r)/\sqrt{A_2}$. Together with (72), this proves (73). Taking logarithms of its positive ratio gives

$$\log\frac{R(W_a(s))}{R(W_b(s))} = -(r_a - r_b)s + (\Lambda_a - \Lambda_b)\log s + \log\frac{\mu(r_a)\Gamma_a}{\mu(r_b)\Gamma_b} + o(1).$$

Boundedness of this log ratio forces equality of rates and then shapes. At a common rate, $\mu(r)$ cancels. The remaining constant vanishes exactly when the amplitudes match. This proves both claims. □

Relative risk and parameter interpolation. All eligible risks tend to zero, so $|R(W_a)-R(W_b)| \to 0$ by itself would be vacuous. The relative criterion distinguishes a genuine constant multiplicative mismatch. Likewise, linearly interpolating saved *loss* values is not equivalent to evaluating loss on $W(s)$. Even for $R = e^{-rs}$, a fixed mesh of width h introduces the persistent factor $(1-a)e^{rah} + ae^{-r(1-a)h}$ at a fractional node. The proof and any verification must use parameter interpolation or actual nodes with the appropriate clock, not silently interpolate losses.

Corollary 19 (Constructive amplitude compensation)**.** *After the original source-to-target map has matched (r, Λ), the unique positive target amplitude that also matches Γ is*

$$\epsilon_{0,b} = \epsilon_{0,a}\sqrt{\frac{A_{2,b}}{A_{2,a}}}\left(\frac{\sigma_b s_{0,b}}{\sigma_a s_{0,a}}\right)^{\Lambda}. \tag{75}$$

It does not require a target classifier, target trajectory, or evaluation of $\mu(r)$. For fixed memory, step and s_0, where $\chi > 0$, the map $(k, \lambda, \log\epsilon_0) \mapsto (r, \Lambda, \log\Gamma)$ has Jacobian determinant $\chi^2/\sigma > 0$.

Proof. Solve equality of the two positive expressions in (71). The first two coordinates have determinant χ^2/σ by Appendix J.6 and do not depend on ϵ_0. Since $\partial_{\log\epsilon_0}\log\Gamma = 1$, the full Jacobian has the same determinant. Local invertibility holds within the legal pole interval; the resulting run must satisfy the inherited eligibility conditions. □

An arbitrarily large mismatch inside one trajectory class. Fix the same memory, step, k, λ, s_0, and multiply only ϵ_0 by any fixed $M > 0$. For eligible runs the original trajectory class is unchanged, but the new run's training risk divided by the old run's tends to M. No bound on this factor follows from two-coordinate equality. For $z = 1$, zero memory and $0 < r = k < 1$, the selected multiplier is $p = r/(1-r)$ and, for $\lambda = 0$, $W(s) = rs - \log(p\epsilon_0) + o(1)$. This gives an explicit full-rank example. A step change or an altered amplitude must be checked for eligibility; no uniform statement in $M \to \infty$ is used.

Clock-origin dependence. If one replaces the fixed coordinate by $\widetilde{s} = s + c$ and describes the same path as $W(\widetilde{s} - c)$, its amplitude becomes $e^{rc}\Gamma$. Thus Γ is complete only for the prescribed clock origins. This is not an invariant modulo arbitrary post-hoc time shifts.

N.3 What amplitude matching removes from the finite offset

Theorem 20 (Essential-margin and observable offset cancellation)**.** *Suppose (r, Λ) agree and let $d_\infty = \lim_{s\to\infty}(W_a(s) - W_b(s))$, as supplied by the inherited hierarchy. Then*

$$Z_J d_\infty = \log(\Gamma_b/\Gamma_a)\mathbf{1}. \tag{76}$$

Consequently, amplitude matching implies $d_\infty \in \ker Z_J$. If $\ker Z_J \subseteq \ker \mathcal{O}$ for a fixed linear observation map $\mathcal{O}$, then matching all three coordinates gives $\mathcal{O}(W_a - W_b) \to 0$. In particular $W_a - W_b \to 0$ when Z_J has full column rank. For softmax this rank statement should instead be checked on the class-centered observation space.

Proof. For $i \in J$, the atom limit is positive, so Lemma 17 and (72) imply

$$\omega_i e^{-z_i^\top W_x(s)} = p_i(r)\Gamma_x e^{-rs}s^\Lambda(1 + o(1)).$$

The same-data, same-weight assumption fixes $p_i(r)$ in the two runs. Dividing and taking logarithms gives $z_i^\top(W_a - W_b) \to \log(\Gamma_b/\Gamma_a)$. The existing finite-limit result proves (76); the linear observation statements follow immediately. A remaining component in $\ker Z_J$ is not claimed to vanish. For example $z = (1, 0)$ leaves its second parameter unobserved and unchanged by zero-moment Adam, so two different initial second coordinates persist at matched risk amplitude. □

N.4 Equal-update risk and the clock incompatibility

Proposition 21 (Effective time to a small training-loss threshold)**.** *Let $T_x(\ell)$ be the first effective time on a fixed common tail at which $R(W_x(s)) \le \ell$, using the continuous parameter interpolant. As $\ell \downarrow 0$, with $L = \log(1/\ell)$,*

$$T_x(\ell) = \frac{L + \Lambda_x \log L + \log(\mu(r_x)\Gamma_x) - \Lambda_x \log r_x}{r_x} + o(1). \tag{77}$$

At matched (r, Λ), $T_a(\ell) - T_b(\ell) = r^{-1} \log(\Gamma_a/\Gamma_b) + o(1)$.

Proof. The negative log of (73) is $H_x(s) = r_x s - \Lambda_x \log s - \log(\mu(r_x)\Gamma_x) + o(1)$. The displayed leading function is strictly increasing on a tail with slope at least $r_x/2$. Its inverse therefore sandwiches first threshold crossings of H_x within $o(1)$ of its own root; continuity and positivity on every compact prefix exclude an earlier crossing for all sufficiently small ℓ. Substitution first gives $s \sim L/r_x$, then (77). No monotonicity assumption on the entire finite training history is required. □

Theorem 22 (Risk equivalence at equal update count is a different class)**.** *For actual states after n completed updates, define*

$$\vartheta_x = k_x \eta_x, \qquad K_x = \frac{\mu(r_x)\epsilon_{0,x}}{\sqrt{A_{2,x}}} (\eta_x/s_{0,x})^{\lambda_x}. \tag{78}$$

Then

$$R(w_{x,n+1}) = K_x e^{-\vartheta_x n} n^{\lambda_x}(1 + o(1)), \quad \frac{R(w_{a,n+1})}{R(w_{b,n+1})} \to 1 \iff (\vartheta_a, \lambda_a, K_a) = (\vartheta_b, \lambda_b, K_b). \tag{79}$$

Let $N_x(\ell)$ be the first integer count of completed updates at which $R(w_{x,n+1}) \le \ell$. With $L = \log(1/\ell)$ and $\ell \downarrow 0$,

$$N_x(\ell) = \frac{L + \lambda_x \log L + \log K_x - \lambda_x \log \vartheta_x}{\vartheta_x} + O(1), \qquad \frac{N_b(\ell)}{N_a(\ell)} \to \frac{\vartheta_a}{\vartheta_b}. \tag{80}$$

In particular, matched (r, Λ, Γ) does not imply matched update cost: $\vartheta_x = r\eta_x\sigma_x$ may differ.

Proof. Apply the inherited node limit $R(w_{n+1})/\epsilon(\eta n) \to \mu(r)/\sqrt{A_2}$ and expand $(1 + \eta n/s_0)^{\lambda}$. The same successive-scale argument as in Theorem 18 proves the equivalence. Consecutive risks have ratio tending to $e^{-\vartheta_x} < 1$, so they are eventually strictly decreasing. Inverting the negative-log expansion gives (80); integer overshoot gives $O(1)$, not $o(1)$. □

Corollary 23 (An obstruction to matching both clocks)**.** *Two-coordinate trajectory equivalence at nonzero common Λ and relative-risk equivalence at equal update count together require, at a common $\vartheta > 0$,*

$$\chi_a(\vartheta) = \chi_b(\vartheta), \qquad \chi_x(\vartheta) = 1 - \vartheta \left(\frac{\beta_{1,x} e^{\vartheta}}{1 - \beta_{1,x} e^{\vartheta}} - \frac{\beta_{2,x} e^{2\vartheta}}{1 - \beta_{2,x} e^{2\vartheta}} \right). \tag{81}$$

In particular, a memoryless source and a target (β, β^2) with $0 < \beta < 1$ cannot meet both requirements at nonzero Λ, for any individually eligible choices of their positive steps, rates and amplitudes.

Proof. Equal-update relative-risk equivalence forces a common ϑ and raw exponent λ. Trajectory equivalence forces common $\Lambda = \lambda\chi$; nonzero Λ forces $\lambda \neq 0$. This proves the necessary condition. For the stated target let $\rho = \beta e^{\vartheta} < 1$. Its factor is $\chi_b = 1 - \vartheta\rho/(1-\rho^2) < 1 = \chi_a$. The contradiction is independent of ϵ_0 and s_0. The claim does not cover $\Lambda = 0$, permit a freely fitted time warp, or rule out approximate finite-budget transfer. □

N.5 A FINITE-WINDOW TRAINING-RISK ENVELOPE

This pointwise bound uses an evaluated normalized state or rigorous bounds on that state. It decomposes finite-window error into unequilibrated mass, the grouped-loss correction and the finite schedule/clock factor. For one evaluated parameter $W_x(s)$ let $m_x(s) = \sum_i \xi_{i,x}(s)$ and $m_x^* = \mu(r_x)/\sqrt{A_{2,x}}$. For a grouped loss let $\tau_x(s) = \max_j \sum_{i \in j} e^{-z_i^\top W_x(s)}$. Set $\tau_x = 0$ for EXP.

Proposition 24 (Normalized mass controls finite-window log risk)**.** *For $\delta_x = |m_x/m_x^* - 1| < 1$, define*

$$E_x(s) = -\log(1 - \delta_x) + \log(1 + \tau_x) + |\Lambda_x| \left| \log \frac{\sigma_x(S_x(s) + s_{0,x})}{s} \right|. \tag{82}$$

Then

$$\left|\log \frac{R(W_x(s))}{\mu(r_x)\Gamma_x e^{-r_x s} s^{\Lambda_x}}\right| \le E_x(s).$$

At matched (r, Λ) *this yields*

$$\left|\log \frac{R(W_a(s))}{R(W_b(s))} - \log \frac{\Gamma_a}{\Gamma_b}\right| \le E_a(s) + E_b(s).$$

Consequently amplitude matching and a verified sum $E_a + E_b \le \varepsilon$ *imply the finite-window ratio lies in* $[e^{-\varepsilon}, e^{\varepsilon}]$.

Proof. Apply (70) with $m = m_x$ and $\tau = \tau_x$. The normalized mass contributes at most $-\log(1-\delta_x)$ to the absolute log error; the grouped-loss factor contributes at most $\log(1+\tau_x)$. Equation (72) contributes the remaining term. Subtract the single-run log identities to prove the comparison. □

The corresponding equal-update bound is Theorem 26. Evaluations on observed C states and complete tables are provided in the artifact. The bounds are generally conservative, with tighter values at some evaluated windows; they serve as a posterior finite-window diagnostic.

O A CONSTRUCTIVE ALTERNATIVE AT EQUAL UPDATE BUDGET

When equal-update risk is the preservation target, we can match its invariants while retaining the primary selector. The resulting controller changes the target step and remains subject to the source and target runs' eligibility conditions.

Theorem 25 (Equal-update risk calibration preserving the primary selector)**.** *Fix an eligible source run* a *and write* $\vartheta = k_a\eta_a > 0$, $r = r_a$, *and* $\lambda = \lambda_a$. *For a target memory pair define*

$$A_{\ell,b}(\vartheta) = \frac{1-\beta_{\ell,b}}{1-\beta_{\ell,b}e^{\ell\vartheta}}, \qquad \sigma_b(\vartheta) = A_{1,b}/\sqrt{A_{2,b}},$$

assuming both target poles are below one. For any positive $s_{0,b}$ *set*

$$k_b = r\sigma_b, \qquad \eta_b = \frac{\vartheta}{r\sigma_b}, \qquad \lambda_b = \lambda_a,$$

$$\epsilon_{0,b} = \epsilon_{0,a}\sqrt{\frac{A_{2,b}}{A_{2,a}}}\left(\frac{\eta_a s_{0,b}}{\eta_b s_{0,a}}\right)^{\lambda_a}. \tag{83}$$

If the resulting target run is eligible under the inherited selection hypotheses, it has the same primary selector u_r, *and*

$$\frac{R(w_{a,n+1})}{R(w_{b,n+1})} \longrightarrow 1 \qquad (n \to \infty). \tag{84}$$

The formula requires no target classifier, selected multiplier, or target trajectory. It requires freedom to change the target learning step; it is not a same-learning-rate rule. Among choices preserving r, ϑ, *the raw exponent and the chosen* $s_{0,b}$, *it fixes* k_b, η_b *and* $\epsilon_{0,b}$ *uniquely.*

Proof. The proposed choices give $k_b\eta_b = \vartheta$, so the target memory factors evaluated from its actual (k_b, η_b) are exactly the displayed ones. They also give $k_b/\sigma_b = r_a$. Equality of the fixed margin system and r fixes u_r and $\mu(r)$. Finally,

$$\frac{\epsilon_{0,b}}{\sqrt{A_{2,b}}}(\eta_b/s_{0,b})^{\lambda_a} = \frac{\epsilon_{0,a}}{\sqrt{A_{2,a}}}(\eta_a/s_{0,a})^{\lambda_a},$$

so $K_b = K_a$ in (78). Theorem 22 proves (84). Fixing ϑ first fixes the memory factors; preserving r then fixes k_b and η_b, and the positive amplitude equation fixes $\epsilon_{0,b}$. □

Choosing the preservation target. The original controller preserves (r, Λ) and the amplitude correction in Corollary 19 additionally preserves risk in effective time. The budget-aligned controller instead preserves $(r, \vartheta, \lambda, K)$, with K obtained without evaluating $\mu(r)$ because the primary rate is shared. It generally changes $\Lambda = \lambda\chi$, so effective-clock parameter differences can retain logarithmic drift even while equal-update training risks have ratio one. This is consistent with Corollary 23 and shows why full-trajectory preservation is not a necessary target for every useful tuning transfer.

P EXACT UPDATE SYNCHRONIZATION AND JOINT ELIGIBILITY

We specialize the controller of Appendix O by synchronizing the onset in update units. This yields an exact finite-node normalization and a jointly eligible small-step family. Let a denote the source. Write $\vartheta = k_a\eta_a > 0$ and $r = r_a > 0$. For every fixed memory pair x set

$$A_{\ell,x}(\vartheta) = \frac{1-\beta_{\ell,x}}{1-\beta_{\ell,x}e^{\ell\vartheta}}, \quad \ell = 1,2, \qquad \sigma_x = A_{1,x}/\sqrt{A_{2,x}}, \tag{85}$$

$$k_x = r\sigma_x, \qquad \eta_x = \frac{\vartheta}{r\sigma_x}. \tag{86}$$

All denominators in (85) must be positive. Choose

$$\lambda_x = \lambda_a, \qquad s_{0,x} = \frac{\eta_x}{\eta_a}s_{0,a}, \qquad \epsilon_{0,x} = \epsilon_{0,a}\sqrt{\frac{A_{2,x}}{A_{2,a}}}. \tag{87}$$

This specialization of the budget controller synchronizes the polynomial onset in *update units*: $s_{0,x}/\eta_x = s_{0,a}/\eta_a =: n_0$.

Theorem 26 (Exact schedule normalization and a finite-node risk envelope)**.** *The choices* (86)–(87) *give, for every integer $n \geq 0$,*

$$\frac{\epsilon_x(\eta_x n)}{\sqrt{A_{2,x}}} = \frac{\epsilon_{0,a}}{\sqrt{A_{2,a}}}e^{-\vartheta n}(1+n/n_0)^{\lambda_a} =: E_n. \tag{88}$$

Thus there is no remaining polynomial-onset or effective-clock conversion error in this identity. Let $X_x(n)$ be the weighted exponential reference loss at the actual parameter $w_{x,n+1}$, even when training uses grouped logistic loss, and define

$$m_x(n) = X_x(n)/\epsilon_x(\eta_x n), \qquad a_x(n) = \sqrt{A_{2,x}}m_x(n)/\mu(r).$$

Here $\mu(r) > 0$ is the inherited primary multiplier mass on the fixed learning problem. For exponential training loss $R_x(n)$,

$$\frac{R_a(n)}{R_b(n)} = \frac{a_a(n)}{a_b(n)}. \tag{89}$$

For grouped logistic/softmax let $\tau_x(n) = \max_j \sum_{i\in j} e^{-z_i^\top w_{x,n+1}}$. Then

$$\frac{a_a/a_b}{1+\tau_a} \leq \frac{R_a}{R_b} \leq (1+\tau_b)\frac{a_a}{a_b}. \tag{90}$$

If $\delta_x = |a_x - 1| < 1$, a valid finite-node bound is

$$\left|\log\frac{R_a}{R_b}\right| \leq -\log(1-\delta_a) - \log(1-\delta_b) + \log(1+\tau_a) + \log(1+\tau_b), \tag{91}$$

with $\tau_a = \tau_b = 0$ for exponential loss.

Proof. The selected parameters give $k_x\eta_x = \vartheta$ exactly. Substitution in the schedule proves (88), including $n = 0$. Consequently $X_x(n) = E_n\mu(r)a_x(n)$, proving (89). Applying (70) to each run and dividing proves (90). Finally $|\log a_x| \leq -\log(1-\delta_x)$ gives (91).

□

The envelope uses the evaluated target mass. Under eligibility, $a_x(n) \to 1$ and $\tau_x(n) \to 0$, recovering equal-update relative-risk calibration. At finite nodes, the mass ratio measures the discrepancy that remains after schedule normalization.

Lemma 27 (Compact-amplitude extension of inherited admissibility)**.** *In the inherited selection and hierarchy setting, the common sufficiently small step can also be chosen uniformly for $\epsilon_0 \in [e_-, e_+] \Subset (0,\infty)$, with the original compact initialization, interior rate and profile sets and a fixed memory pair. For a finite set of memory pairs, take the minimum of their step thresholds.*

Proof. We extend the compact-parameter argument by checking amplitude dependence in entry, shadowing and absorption. First, on every fixed time interval the schedule has a uniform positive minimum and bounded derivatives when the amplitude lies in $[e_-, e_+]$; its exponential tail and squared-tail sums have uniform upper bounds. Initial normalized atoms are bounded above and away from zero on the finite initialization box. Second, the continuous primary-capture argument uses an upper reference-mass box, the integrable schedule forcing and finite-time continuity. The mass-box and forcing constants can be chosen using e_+; initial positive bounds use e_-. The augmented KL energy is continuous also in amplitude at finite times. The same strict-entry, non-exit and finite-subcover argument therefore gives a common finite entry time on the enlarged compact parameter set. Third, stopped finite-window shadowing uses only the just established positive schedule minimum, finite boxes and derivative bounds. It supplies one uniform sufficiently small step and discrete entry for all these amplitudes. The KL–RMS non-exit reserve uses the uniform squared tails; its absorption constants involve upper boxes and memory gaps, not inverse individual equilibrium masses. Finally the subsequent fixed-face hierarchy is invoked after entry. Its face-dependent tail constants may vary, but its admissible step is the already chosen common one. None of these arguments uses a limit as $\epsilon_0 \downarrow 0$ or $\epsilon_0 \uparrow \infty$. These are the entry and absorption arguments of Appendices D and E, now applied to the enlarged compact parameter set. □

Theorem 28 (A nonempty jointly eligible family)**.** *Fix finite separable data, positive weights, a compact initialization set, a finite collection of memory pairs, r in a compact subset of $(0, \gamma_\infty)$, bounded λ_a, and $\epsilon_{0,a}, s_{0,a}$ in compact positive intervals. For all sufficiently small $\vartheta > 0$,* (86)–(87) *define individually eligible runs for every pair. They preserve the primary selector and have training risk ratio tending to one at equal update count.*

Proof. For the finite memory set, the factors in (85) are smooth on a common interval around zero, and $A_{\ell,x}, \sigma_x \to 1$ uniformly. Hence all poles remain below one; $k_x = r\sigma_x$ lies in a slightly larger compact interior rate interval and $\eta_x = O(\vartheta) \to 0$ uniformly. Moreover $s_{0,x} = (\sigma_a/\sigma_x)s_{0,a}$ and $\epsilon_{0,x} = \epsilon_{0,a}\sqrt{A_{2,x}/A_{2,a}}$ remain in compact positive intervals. Apply Lemma 27 to these fixed compact sets, then shrink ϑ so that every η_x lies below the corresponding common threshold. The inherited node mass limit and Theorem 26 finish the proof. □

Q MEMORY DEFORMATION, CLOCK READOUTS, AND MULTI-HORIZON RISK EVIDENCE

This appendix gives the detailed protocols and readouts for the three experiment families in Section 5: fixed-step memory response, clock-dependent trajectory drift, and multi-budget risk calibration.

Q.1 FIXED-STEP MEMORY-RESPONSE DIAGNOSTIC

The grid contains eight distinct steps for each G1/G2/G0 $\times$ C/M curve: 32 nonzero configurations and 16 G0 controls. Group C uses $(0.9, 0.999)$ with bias correction and $\eta \in [10^{-4}, 4 \times 10^{-4}]$; group M uses $(0.5, 0.8)$ without correction and $\eta \in [0.005, 0.01]$. These compound groups vary all three choices together; Table 3 summarizes the six curves.

Measurement definition. Let $N = 600/\eta$ and $n = t - 1$, so $S_t = n\eta$ and w_t is the state after n updates. The statistic is the uniform dense-tail mean

$$\widehat{a} = \frac{1}{N/10 + 1} \sum_{n=9N/10}^{N} h_{n+1}, \qquad E_{\mathrm{mem}} = \|\widehat{a} - \sigma u_{k/\sigma}\|_{H_{\mathrm{geom}}}, \quad E_{\mathrm{naive}} = \|\widehat{a} - u_k\|_{H_{\mathrm{geom}}}.$$

Both endpoints are included; the final current-query velocity is read before any further parameter update. Compensated summation uses every query, not sparse checkpoints. Here $\|v\|_{H_{\mathrm{geom}}} = \sqrt{v^\top H_{\mathrm{geom}} v}$ has no relative-error denominator. The frozen diagonal $H_{\mathrm{geom}} = \nabla^2 B(u_k)$ matrices for G1, G2 and G0 are $(16/9, 4)$, $(9/4, 4)$ and $(144/25, 144/25)$: barrier Hessians at u_k, not at $u_{k/\sigma}$ and not divided by A_1. G1 has rows $(1/3, 1), (1/3, 1), (1, 1)$, and G2 replaces $1/3$ by $1/2$; their weights are $(0.37, 0.63, 0.5)$ and raw rates $7/12, 2/3$. G0 has the two coordinate rows, weights $(0.37, 0.63)$ and raw rate $7/12$. Duplicate rows are included. All cells start at $w_1 = (3/5, -7/20)$ with zero moments and $\epsilon(S) = e^{-kS}$.

Coverage. Each curve contains two endpoint and six interior configurations. Figure 2 compares their measured means with the analytic branch. The artifact provides the complete 48-row table, step sizes and averaging windows.

Table 3: Memory-response coverage by geometry and compound group. C: $(0.9, 0.999)$, bias-corrected; M: $(0.5, 0.8)$, uncorrected. End./int. counts endpoint and interior configurations. Deformation and residual are in the frozen H_{geom} norm; the full-precision table includes all 48 configurations.

Geom.	Group	η interval	Steps	End./int.	Deformation range	Max. E_{mem}
G1	C	$[10^{-4}, 4 \times 10^{-4}]$	8	2/6	9.63e-03–6.11e-02	1.79e-14
G1	M	$[0.005, 0.01]$	8	2/6	1.34e-03–2.75e-03	2.22e-16
G2	C	$[10^{-4}, 4 \times 10^{-4}]$	8	2/6	1.03e-02–6.77e-02	3.66e-14
G2	M	$[0.005, 0.01]$	8	2/6	1.43e-03–2.93e-03	1.39e-16
G0	C	$[10^{-4}, 4 \times 10^{-4}]$	8	2/6	0.00e+00–6.93e-15	6.93e-15
G0	M	$[0.005, 0.01]$	8	2/6	0.00e+00–3.77e-16	3.77e-16

Q.2 TRAJECTORY AND CLOCK PROTOCOL

The L0/L1 study includes an equal-coordinate regular control and the nonradial row $z = (1/3, 1)$ at $r = 7/12$, with $u = (1/4, 1/2)$ and $v = (3/5, 4/5)$. The control compares source memory $(0.9, 0.999)$ with three shifts; the nonradial example compares $(0, 0)$ with $(0.5, 0.25)$. Each block has reference, uncalibrated, rate-only and two-coordinate arms. Sixteen bias-corrected arms run for 10^6 actual updates; sixteen uncorrected arms check the alternative convention rather than adding independent scientific replications.

For the nonradial example the reference effective shape is 4. RATE matches r within 10^{-11} but leaves $\Delta\Lambda = 0.0157703612$. The predicted logarithmic coefficient $(-0.00946222, -0.01261629)$ is fitted as $(-0.00978622, -0.01324183)$, with relative error 0.0447 and cosine 0.99998. The TWO

fitted norm is 0.0010171, below the prespecified zero-coefficient diagnostic tolerance 0.0055196. Terminal discrepancies are 22.5983, 0.124891 and 0.0172067 for uncalibrated, RATE and TWO arms. The fixed nine-horizon diagnostic resolves its effective-clock criterion from 10,000 updates on this geometry.

The measurements below distinguish adjacent secants, tail regression and fixed-dilation readouts on the same trajectories.

Figure 1 measurement. The frozen 64-point readout defines $D_x = W_{\mathrm{REF}} - W_x$ and $\Delta_x = \langle v_{\mathrm{pred}}, D_x \rangle$, with $v_{\mathrm{pred}} = \Delta c_{\mathrm{RATE}}/\|\Delta c_{\mathrm{RATE}}\| = (-0.6, -0.8)$. The right panel uses 63 adjacent secants

$$L_{x,i} = \frac{\Delta_x(s_{i+1}) - \Delta_x(s_i)}{\log s_{i+1} - \log s_i}, \qquad s_{\mathrm{display}} = \sqrt{s_i s_{i+1}}.$$

Logarithms in the estimator are natural. Every adjacent pair is used once; there are no extrapolated endpoint slopes, regression or smoothing. The plotted logarithmic axis does not change this denominator. The coefficient fit reported above is an intercept-plus-$\log s$ OLS fit to the final 25 of 64 vector readouts (zero-based start index 39). Its 4.47% error is not a local-slope error. On this geometric grid the secant is a projected fixed-dilation operation with ratio s_{i+1}/s_i, distinct from the retrospective $\gamma = 2$ query. Its coefficient annotation refers to the fixed-dilation limit in Appendix J; no derivative of an $o(1)$ remainder or shrinking-window limit is used.

For the fixed-dilation readouts below, parameters are interpolated between adjacent actual saved nodes inside common clock support. These readouts use no extrapolation, fitted shift or coefficient regression. The endpoint comparison is in Table 4. Direction agreement suppresses information about parameter scale and slower divergent terms. To expose that information, we use the retrospective readouts

$$D_x(s) = W_{\mathrm{REF}}(s) - W_x(s), \qquad G_x(s) = \frac{D_x(2s) - D_x(s)}{\log 2}. \tag{92}$$

At matched r, the full-clock prediction for G_x is $c_{\mathrm{REF}} - c_x$: nonzero for RATE in the canonical nonradial example and zero for TWO. A finite nonzero D_x alone does not establish unbounded separation; the predicted logarithmic coefficient and its G_x readout are the relevant comparison.

In the canonical block at $s = 5000$, RATE has normalized direction gap 5.63×10^{-6} but $\|D_{\mathrm{RATE}}\| = 0.113905$. Its $\|G_{\mathrm{RATE}}\| = 0.015863$ is close to the frozen coefficient norm 0.015770. TWO has direction gap 6.80×10^{-7}, $\|D_{\mathrm{TWO}}\| = 0.017130$ and $\|G_{\mathrm{TWO}}\| = 1.11 \times 10^{-4}$ against a zero coefficient (Figure 5).

The observation clock also changes the coefficient being tested. The full clock $s = \sigma S - \lambda\sigma' \log(1 + S/s_0)$ is fixed by the configuration. The linear-clock ablation $s_{\mathrm{lin}} = \sigma S$ omits its logarithmic term; it compares the same trajectories using a different observable. Its prediction is $b_{\mathrm{REF}} - b_x$, with $b_x = c_x - \lambda_x d_{\sigma,x} a_x$, rather than $c_{\mathrm{REF}} - c_x$. At the largest query the linear-clock RATE and TWO norms are 0.002850 and 0.015126, against predictions 0.002850 and 0.015173 (Figure 5). Neither clock is fitted to the outcome, and their coefficients cannot be interchanged. The effective-clock diagnostic meets its criterion from 10k updates; the shorter 1k/3k and physical-clock windows remain unresolved.

Table 4: RATE/TWO readouts on the nonradial example at $s = 5000$ (and at the same numerical linear-clock query). Predictions are numerically evaluated vector-coefficient norms, not fitted values; the full-clock TWO value is roundoff around the analytic zero.

Arm	Clock	Direction gap	$\|D\|$	$\|G\|$	Prediction
RATE	full	5.632e-06	1.139e-01	1.586e-02	1.577e-02
TWO	full	6.801e-07	1.713e-02	1.113e-04	4.783e-15
RATE	linear	5.593e-06	2.457e-02	2.850e-03	2.850e-03
TWO	linear	6.406e-07	7.713e-02	1.513e-02	1.517e-02

Transverse readout. With $u = (1/4, 1/2)$ in the nonradial block, project the same $G_x(s)$ onto $u^{\perp}$. At $s = 5000$, full-clock RATE has transverse norm 2.83×10^{-3} versus 2.82×10^{-3} predicted, with residual norm 1.09×10^{-5}. TWO has norm 2.11×10^{-5} against the zero analytic transverse

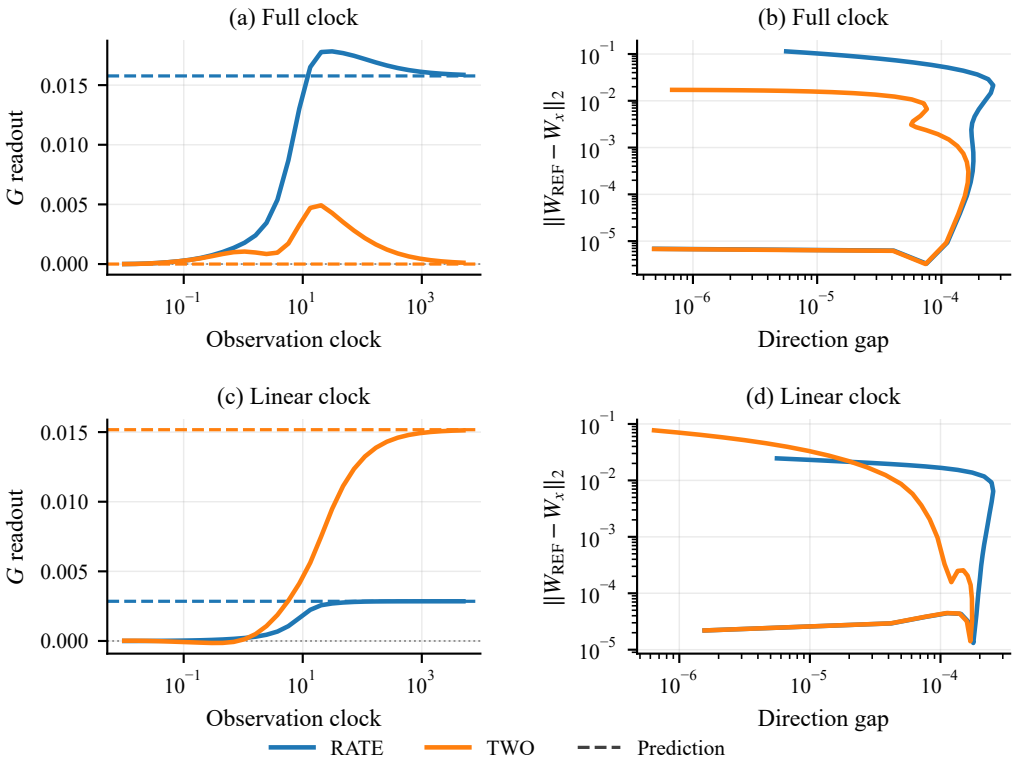

Figure 5: The same saved RATE/TWO trajectories under full (top row) and linear (bottom row) clocks. Blue: RATE; orange: TWO. Solid curves are observations. (a,c) Fixed-dilation readouts, projected onto the predicted coefficient when nonzero and using the norm otherwise; dashed lines are analytic predictions. (b,d) Direction gap versus full parameter difference.

prediction. Under the linear clock the observed norms are 2.82×10^{-3} and 5.68×10^{-6}, respectively. The full- and linear-clock coefficient predictions have the same transverse projection at matched r, since their difference is radial. The analysis covers all 512 supported query rows, both clocks and the equal-coordinate control. This projection is distinct from Figure 1's prediction-axis readout and tests the finite-window consequence of Corollary 14 on the same trajectories.

Q.3 Risk protocol and multi-budget summaries

C evaluates one-dimensional EXP, nonradial two-dimensional EXP, weak-row EXP and native grouped softmax. P0 changes memory from $(0, 0)$ to $(0.6, 0.36)$; P1 changes $(0.5, 0.8)$ to $(0.7, 0.9)$, and P2 reverses this change. Each configuration is integrated to 64k updates, with readouts at 4k, 16k and 64k.

Effective-time risk is evaluated on linearly interpolated parameters within common clock support, whereas equal-update risk uses the exact saved N-update states. The artifact contains the complete signed predictions, observations, query clocks and all six arms.

Geometry and query definitions. The comparison uses four data geometries: 1D has margin row (1) and weight 1; 2D has rows $(1, 0.2), (0.25, 1), (0.8, 0.7)$ with weights $(1, 0.8, 1.2)$; Weak has rows $(1/3, 1), (1, 2/3)$ and weights $(1, 0.7)$; Softmax has features $(1, 0), (0, 1), (-1, -1)$, labels $0, 1, 2$, weights $(1, 0.8, 1.2)$ and all six parameter coordinates. Each starts with zero parameters and moments, without bias correction. The source step is 0.02, amplitude 0.3 and onset 2; $r = 0.15$, except $7/12$ for Weak. P0 compares $(0, 0)$ to $(0.6, 0.36)$, P1 compares $(0.5, 0.8)$ to $(0.7, 0.9)$ and P2 reverses P1. These 64k schedules fix effective shape $\Lambda = 0.7$ (and raw source shape $\lambda = \Lambda/\chi$), with an additional 1D/P0 $\Lambda = 0$ control. For budget N, each arm's effective-time query is

$s = \min\{s_{\mathrm{REF}}(\eta_{\mathrm{REF}}N), s_x(\eta_x N)\}$; loss is recomputed on interpolated parameters. Equal-update queries use the actual N-update states. Table 5 is generated directly from the frozen CSV rows. Complete signed values, all six arms, query clocks and prediction residuals are provided in the artifact.

Risk observables and query support. We write $\ell_x^{(s)}(s) = \log[R(W_x(s))/R(W_{\mathrm{REF}}(s))]$ and $\ell_x^{(n)}(n) = \log[R(w_{x,n+1})/R(w_{\mathrm{REF},n+1})]$. The prediction residual is $e_x = \ell_{x,\mathrm{obs}} - \ell_{x,\mathrm{pred}}$; a median $|e_x|$ summarizes prediction error, not necessarily source-risk mismatch. Figure 3(b) uses the same saved left endpoints as (a), ending at $s = 639.96$; this is not the largest common risk-only query. Panel (a) additionally evaluates $2s$, within REF and target support. The equal-update panel uses every shared saved integer-update checkpoint between 4k and 64k, marking all 26 observed checkpoints per arm. The saved plan and trajectory CSV retain consecutive actual-node brackets; risk is evaluated on the interpolated parameter. At the last common query ($s = 639.96$), AMP/SYNC give $-5.73 \times 10^{-5}/-1.73 \times 10^{-3}$ for $[G_{x,2}]_1$ and $1.04 \times 10^{-4}/-1.13 \times 10^{-2}$ for the effective-time log-risk ratio. These figure readouts are distinct from the budget-specific effective-time queries in the tables. The equal-update readouts use exact checkpoint losses without interpolation.

The comparison uses the full signed predictions from the frozen configurations. Writing R_x for the common risk evaluated along arm x, Eqs. (16) and (18) give

$$\log \frac{R_x(W_x(s))}{R_{\mathrm{REF}}(W_{\mathrm{REF}}(s))} = -(r_x - r_{\mathrm{REF}})s + (\Lambda_x - \Lambda_{\mathrm{REF}}) \log s + \log \frac{\mu(r_x)\Gamma_x}{\mu(r_{\mathrm{REF}})\Gamma_{\mathrm{REF}}} + o(1), \tag{93}$$

$$\log \frac{R_x(w_{x,n+1})}{R_{\mathrm{REF}}(w_{\mathrm{REF},n+1})} = -(\vartheta_x - \vartheta_{\mathrm{REF}})n + (\lambda_x - \lambda_{\mathrm{REF}}) \log n + \log(K_x/K_{\mathrm{REF}}) + o(1). \tag{94}$$

At common (r, Λ), TWO permits a nonzero log amplitude ratio, while AMP targets zero. SYNC instead targets zero at equal update count. AMP at equal updates can retain both the n and $\log n$ terms; SYNC at common effective time generally retains a $\log s$ term through its changed Λ. Their nonzero finite endpoints are therefore not generally nonzero constant limits.

Figure 3 places parameter drift and both risk targets on the same nonradial P0 instance. At the last gamma-two query ($s = 639.96$), AMP has $\|G\| = 7.8 \times 10^{-5}$ against zero; SYNC has $G = (-0.001733, -0.001624)$ against $(-0.001669, -0.001564)$. At 64k equal updates, their signed log-risk ratios are -0.529159 and 1.82×10^{-5}, respectively. These observations distinguish what AMP and SYNC preserve; matching source risk is a different objective from reducing loss or improving accuracy.

Across the four P1 geometries, AMP's median absolute effective-time prediction residual $|e_x|$ falls from 1.08×10^{-3} to 4.38×10^{-4} to 1.50×10^{-4}; P2 gives the same rounded absolute summaries; the tables retain its signed observations and prediction residuals separately. At 64k, TWO's median observed/predicted log ratios are $-0.007446/-0.007295$ for P1 and $+0.007446/+0.007295$ for P2. Figure 6 retains all four geometries in both orientations, including the weak-row case. SYNC's median absolute equal-update residual at 64k is 9.67×10^{-6} for P1 and 9.56×10^{-6} for P2. For P0, including its separately listed zero-shape control, the corresponding median is 1.82×10^{-5}. The accompanying complete signed tables retain RAW/RATE, early deviations, large weak-row readouts and the zero control. The observations are consistent with the finite-window predictions on these synthetic examples.

Q.4 A CRITICAL LOG-LOG FINITE-WINDOW TEST

The analytic example in Appendix I.2 uses $z_1 = (1/3, 1)$, $z_2 = (1, 2/3)$, unit EXP weights, $r = 7/12$ and $\Lambda = -15/2$. Its selectors are $u = (1/4, 1/2)$, $c = (9/2, 6)$ and $d = (9/7, -3/7)$, with physical weak-mass coefficient $\kappa = A_1(49/180)$. We test this example with zero initial parameters and moments, uncorrected Adam, memory pairs $(0, 0)$ and $(0.5, 0.8)$, and $\eta = 0.02$. Each schedule solves $k/\sigma = r$ and $\lambda = \Lambda/\chi$ before execution. Sixteen configurations cover critical and neighboring shapes $-8.5, -6.5$, two stabilizer settings, and critical half-step checks at $\eta = 0.01$. All 192 prescribed nodes through eight million updates and 80 fixed-window secants are evaluated.

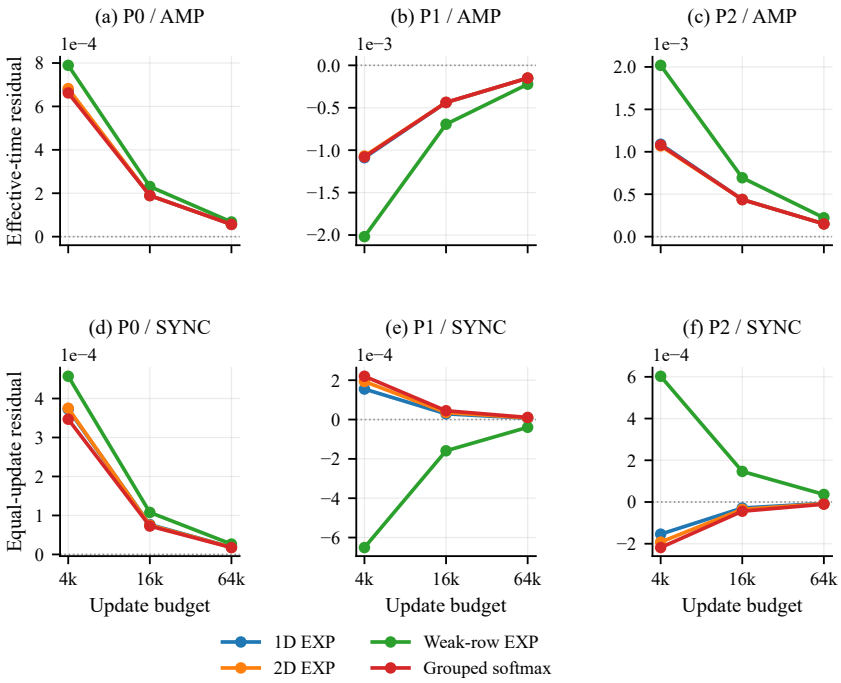

Figure 6: P0 and nonzero-memory P1/P2 comparisons at 4k, 16k and 64k on all four frozen geometries. Residuals compare observed signed log-risk ratios with the full theoretical prediction in the relevant clock. The reverse memory orientation is analyzed separately. Full signed values and all arms are provided in the artifact.

Readouts. At physical nodes $S = \eta n$ use the prescribed full clock s and define

$$R_2 = W(s) - su - c \log s, \qquad R_3 = R_2 - d \log\log s.$$

The R2 secant divides its vector change by the change in $\log\log s$; the R3 readout is its uncentered vector change, leaving the unknown terminal offset free. The other readouts are $\kappa S \log S\, \xi_{\mathrm{weak}}$, the actual-update vector $S \log S(h - a - b/S)$, and the reciprocal difference in Appendix I.2. Thus the update and weak-mass measurements use physical time, while parameter secants use effective time.

Finite-window readouts. A visibility setting $\epsilon_0 = 10^9$, $s_0 = 2$ was selected after a pilot, with the formal grid and stopping rule fixed before the reported endpoints. At $S = 160000$, the nonzero-memory critical arm has normalized weak mass 0.992080. Its endpoint update vector is $(1.277519, -0.425140)$ and its 4m–8m R_2 secant is $(1.271735, -0.426887)$, compared with $d = (1.285714, -0.428571)$. The corresponding R_3 change is $(-0.000836, 0.000101)$, and the reciprocal readout is 0.276537 against $\kappa = 0.275337$. The half-step arm gives the same qualitative readout.

Under the default $\epsilon_0 = s_0 = 1$ setting, the critical coefficient remains pre-asymptotic at eight million updates. The nearby $\Lambda = -8.5$ and -6.5 controls also remain transitional over the evaluated horizon. Complete default and control readouts accompany the artifact.

Numerical precision. Parameters were constructed by an 80-digit static solve and executed in long-double arithmetic. A separately implemented MP50 recurrence reproduces the selected eight-million-step endpoint within 7.32×10^{-12} in parameter coordinates and 5.71×10^{-12} relatively in weak mass. Complete precision checks accompany the artifact.

Table 5: Multi-budget risk summary for all 13 geometry/pair configurations. Under "TWO effective-time ratio," obs. and pred. are signed target/source log-risk ratios. Under "Calibration residual," AMP $|e_s|$ and SYNC $|e_n|$ are absolute prediction residuals in effective time and update count. For AMP and SYNC, the predicted log-risk ratios have magnitude at most 6.13×10^{-12}, so residuals equal source-risk mismatch to that accuracy. "(0)" denotes the zero-shape control; no budget or weak-row result is omitted.

Geometry/pair	Budget	TWO effective-time ratio		Calibration residual	
		obs.	pred.	AMP $\|e_s\|$	SYNC $\|e_n\|$
1D P0 (0)	4k	-1.70e-03	-1.70e-03	2.58e-06	2.57e-06
	16k	-1.70e-03	-1.70e-03	1.01e-12	2.83e-14
	64k	-1.70e-03	-1.70e-03	3.79e-12	2.78e-14
1D P0	4k	-3.01e-03	-3.68e-03	6.75e-04	3.73e-04
	16k	-3.49e-03	-3.68e-03	1.90e-04	7.67e-05
	64k	-3.62e-03	-3.68e-03	5.66e-05	1.85e-05
1D P1	4k	-8.38e-03	-7.30e-03	1.09e-03	1.55e-04
	16k	-7.73e-03	-7.30e-03	4.38e-04	2.91e-05
	64k	-7.45e-03	-7.30e-03	1.50e-04	6.92e-06
1D P2	4k	8.38e-03	7.30e-03	1.09e-03	1.54e-04
	16k	7.73e-03	7.30e-03	4.38e-04	2.88e-05
	64k	7.45e-03	7.30e-03	1.50e-04	6.85e-06
2D P0	4k	-2.98e-03	-3.68e-03	6.82e-04	3.75e-04
	16k	-3.49e-03	-3.68e-03	1.90e-04	7.50e-05
	64k	-3.62e-03	-3.68e-03	5.66e-05	1.82e-05
2D P1	4k	-8.33e-03	-7.30e-03	1.07e-03	1.93e-04
	16k	-7.73e-03	-7.30e-03	4.38e-04	3.57e-05
	64k	-7.45e-03	-7.30e-03	1.50e-04	8.55e-06
2D P2	4k	8.33e-03	7.30e-03	1.07e-03	1.92e-04
	16k	7.73e-03	7.30e-03	4.38e-04	3.54e-05
	64k	7.45e-03	7.30e-03	1.50e-04	8.46e-06
Weak P0	4k	-1.38e-02	-1.46e-02	7.89e-04	4.57e-04
	16k	-1.44e-02	-1.46e-02	2.31e-04	1.08e-04
	64k	-1.45e-02	-1.46e-02	6.76e-05	2.65e-05
Weak P1	4k	-3.02e-02	-2.82e-02	2.02e-03	6.51e-04
	16k	-2.89e-02	-2.82e-02	6.94e-04	1.59e-04
	64k	-2.84e-02	-2.82e-02	2.23e-04	3.94e-05
Weak P2	4k	3.02e-02	2.82e-02	2.02e-03	6.03e-04
	16k	2.89e-02	2.82e-02	6.94e-04	1.47e-04
	64k	2.84e-02	2.82e-02	2.23e-04	3.63e-05
Softmax P0	4k	-3.02e-03	-3.68e-03	6.62e-04	3.47e-04
	16k	-3.49e-03	-3.68e-03	1.89e-04	7.27e-05
	64k	-3.62e-03	-3.68e-03	5.66e-05	1.76e-05
Softmax P1	4k	-8.37e-03	-7.30e-03	1.08e-03	2.20e-04
	16k	-7.73e-03	-7.30e-03	4.37e-04	4.48e-05
	64k	-7.45e-03	-7.30e-03	1.50e-04	1.08e-05
Softmax P2	4k	8.37e-03	7.30e-03	1.08e-03	2.18e-04
	16k	7.73e-03	7.30e-03	4.37e-04	4.43e-05
	64k	7.45e-03	7.30e-03	1.50e-04	1.07e-05