ON DELAYED LEARNING IN REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS: PROVABLE LAWS OF FEEDBACK AND RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) turns automatically checked outcomes into improvements in reasoning policies. Yet informative updates can coexist with prolonged learning delays, even when the training run eventually reaches its target. To explain this phenomenon, we develop a finite-sample theory of feedback arrival and the training time needed to reach a fixed accuracy in RLVR. Our theory identifies a feedback mechanism that amplifies temporary regressions: when successful answers are rare, a harmful update makes the next informative sample harder to obtain, converting short sequences of regressions into heavy-tailed training costs. The severity of this effect is determined jointly by feedback scarcity and the temporal dependence of conditional progress. For a signed-update model, an exact dependence threshold characterizes when the expected training budget is finite. Stronger persistence can cross this threshold while preserving the stationary distribution of individual updates, producing infinite mean activation time at a fixed initialization and fixed accuracy target. Theoretically, we construct exponentially mixing processes with identical finite-window update distributions but different mean-cost regimes, and establish finite-initialization coupling and full-state bounds. We validate the predicted mechanisms empirically via controlled stochastic experiments and language model experiments.

1 INTRODUCTION

Why can a reasoning model keep making useful updates while spending increasingly many rollouts between them? Reinforcement learning with verifiable rewards (RLVR) obtains supervision by generating solutions and checking their outcomes (Shao et al., 2024; DeepSeek-AI et al., 2025). When a response group receives identical scores, for example, its centered reward signal vanishes. The policy must then generate further solutions to obtain a useful comparison. Thus each update changes both the model's accuracy and the data from which its next update will be learned. Understanding the rollout budget consequently requires following this feedback loop through training.

Consider what happens when an update temporarily reduces success on a difficult task. Besides losing progress, the policy becomes less likely to generate the successful answer needed for a useful comparison. Recovery therefore starts from a state where its own feedback is harder to obtain. If several regressions occur together, their sampling costs compound. This suggests a sharper question: can a policy improve on most informative updates, eventually reach its accuracy target, and still have an infinite expected rollout cost? We answer this question affirmatively in a signed-update model. Moreover, the finite-to-infinite transition can occur while preserving the entire stationary distribution of individual updates. Even the distribution of every update window up to a prescribed length can remain unchanged across the two cost regimes.

To understand this effect, we begin with complete verified solutions and the updates they induce. A fully trainable policy already exhibits the underlying tension: one policy-gradient step raises expected accuracy while increasing the expected cost of the next informative group. We then analyze how successive policy updates accumulate these waiting costs.

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.

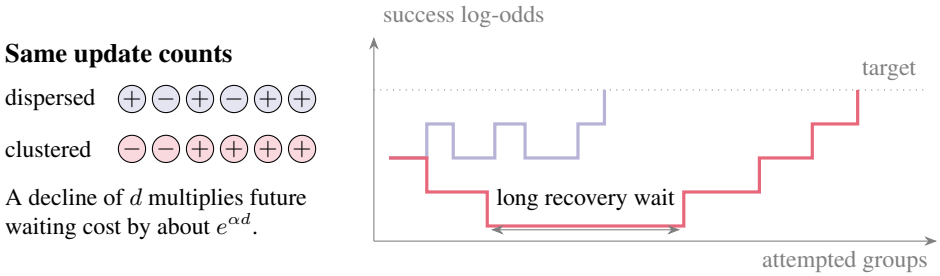


Figure 1: **A regression changes the cost of recovery.** The same six signed updates can be spread out or clustered. At low success probability, a feedback rate proportional to p^{α} makes the next wait exponentially sensitive to a decline in success log-odds. The theorem studies this effect under stationary sign processes with identical update marginals. This schematic uses identical signed increments in both paths, with their ordering determining the displayed recovery costs.

This approach connects verified outcomes to policy changes and the cost of obtaining further feedback. We derive the waiting-cost law, identify the temporal information that determines recovery, and retain background learning in finite-budget bounds. Experiments then follow these mechanisms from controlled processes to trainable policies and language-model feedback forecasts. We summarize the resulting theoretical and empirical contributions below:

- **Verified updates and activation laws.** We derive conditional update laws from verified trajectories and prove that a policy-gradient step can improve accuracy while increasing the next feedback cost. Finite-sample laws quantify how signed updates accumulate waiting costs.
- **Dependence beyond local histories.** We prove an exact moment threshold at fixed initialization and target accuracy. We construct exponentially mixing processes with identical update laws through any prescribed finite window, yet finite versus infinite mean activation cost.
- **Progress between feedback events.** We prove finite-budget guarantees that retain optimizer memory and allow adverse blocks. We derive an exact resource-cost criterion and characterize how curvature-matched regularizers change progress between feedback events.
- **Mechanism validation.** We test the predicted cost distributions with controlled interventions and fully trainable neural policies under stochastic gradient descent (SGD) and Adam. We test feedback forecasts with independently sampled language-model observations.

2 Related Work

RLVR objectives and optimization. The feedback loop begins with how verified responses produce an update. Group relative policy optimization (GRPO) estimates advantages within sampled groups (Shao et al., 2024), while DAPO and Dr. GRPO modify filtering or normalization (Yu et al., 2025; Liu et al., 2025). At the objective level, Davis & Recht (2025) identify transformed success probabilities, and Suk & Duan (2025) relate improvement to gradient alignment and step size. Homogeneous-group starvation makes feedback availability equally important (Nie et al., 2026). Building on this distinction, our equal-data comparison fixes the conditional mean update and isolates the additional effect of update variability on training cost.

Curriculum, composition, and stochastic learning. The available signal also evolves as representations improve. Wang et al. (2025) prove easy-to-hard learnability under supervised training, and Huang et al. (2026) show how outcome rewards induce relay and delayed-learning regimes in compositional transformers. Their revised study includes language-model multiplication experiments. Similarly, Tian (2025) analyzes feature emergence and interaction. These mechanisms concern the formation of useful learning directions. We ask how many generated samples are needed to realize progress once updates themselves alter future feedback. This question complements stochastic policy-gradient and feedback-budget analyses (Mei et al., 2021; Wachi et al., 2026), while our full-state bounds retain the representation changes occurring between selected feedback events.

Verification and probability foundations. Verification at inference time selects an output from a fixed proposal, as studied through optimal transport by Mukherjee et al. (2026). During training, however, the verified output also changes the policy supplying subsequent proposals. We analyze the resulting hitting time using classical exponential functionals and Markov-process tools (Davis, 1984; Bertoin & Yor, 2005; Roitershtein, 2007). Finite-word matching also has combinatorial antecedents (Karhumäki et al., 2013). Our constructions combine exact history matching with opposite fixed-target cost regimes, and our coupling bounds connect finite-group learning to its limiting law. Likewise, although Jensen–Shannon penalties are already used in RLVR (Li et al., 2026), we study their effect during feedback waiting periods under a specified optimizer.

Sampling control and optimization hitting times. Sampling controllers intervene directly in this loop. CBS selects or reuses rollouts, DUET allocates tokens, TRACE branches prefixes, and REFT diversifies first tokens (Lu et al., 2026; Hu et al., 2026; Zou et al., 2026; Kim & No, 2026). Accordingly, Appendix G.11 separates their effects on feedback arrival, conditional progress, and generated-data cost. A different hitting-time mechanism arises in stochastic gradient Langevin dynamics, where injected noise assists exploration (Zhang et al., 2017). Here, acquiring the next informative update can itself be expensive. This distinction allows a light-tailed number of informative updates to coexist with heavy-tailed rollout cost at the same accuracy target.

3 Problem Formulation

The training state. To represent the feedback loop, let Z_n contain the policy parameters, optimizer variables, sampler state, and any required time index. Each transition generates trajectories, evaluates a binary, graded, or noisy verifier, and applies the configured update. We distinguish whether this transition receives a selected source of feedback. For example, $I_n = 1$ can mean that a target prompt obtains nonconstant verified scores. The corresponding event probability and conditional transition laws are

$$\lambda(z) = \mathbb{P}(I_n = 1 \mid Z_n = z), \qquad K_i(z, \cdot) = \mathcal{L}(Z_{n+1} \mid Z_n = z, I_n = i). \tag{1}$$

Conditioning on the event gives the training kernel $P = (1-\lambda)K_0 + \lambda K_1$. Thus λ describes access to feedback, whereas K_1 describes what the optimizer does with it. Meanwhile, K_0 retains momentum, regularization, and learning on other prompts. This decomposition consequently accommodates different architectures and update rules, including updates that reduce the target task's success probability.

What constitutes activation? For a measurable target set A, define activation time as $T_A = \inf\{n \geq 0 : Z_n \in A\}$. We will use a prescribed success probability as the target. Reaching it may require several feedback events, so first consider the simpler time $T_1 = \inf\{n \geq 1 : I_{n-1} = 1\}$. Its survival probability follows by conditioning on each transition without the selected feedback. This first calculation identifies the waiting cost that subsequent updates will change. Time counts attempted groups or optimizer steps, as specified in each comparison.

Proposition 1 (First-feedback survival)**.** *Define* $L_0 f(z) = (1 - \lambda(z)) \int f(z')K_0(z, dz')$. *Then*

$$\mathbb{P}_z(T_1 > n) = L_0^n \mathbf{1}(z) = \mathbb{E}_z^{K_0} \prod_{j=0}^{n-1} (1 - \lambda(Z_j)). \tag{2}$$

If $K_0(z, \cdot)$ is the point mass at z, successive waiting times, conditional on the embedded feedback chain, are independent geometric variables with the corresponding state-dependent rates.

When the state remains fixed, equation (2) reduces to $(1 - \lambda(Z_0))^n$. With shared parameters, however, learning on another task can change this hazard before the target task obtains feedback. The product must then follow that background motion. Even replacing it by an exponential of an averaged cumulative hazard requires control of the logarithmic approximation and its fluctuations. We retain the exact product so that the same formulation covers both cases.

The measurable progress variable. We now specify the downstream target. For a fixed task distribution, write $p(z) \in (0, 1)$ for success probability and $\theta(z) = \operatorname{logit} p(z)$ for success log-odds. Then $T_\star = \inf\{n : p(Z_n) \geq p_\star\}$ measures time to accuracy $p_\star$, at fixed architecture and initialization $\varepsilon = p(Z_0)$. Log-odds provide a useful coordinate because multiplicative changes in rare success become additive increments. Scalar models specify these increments directly, while full-state statements retain interactions among all trainable parameters.

Where feedback scarcity comes from. The arrival rate follows directly from the verifier. At a fixed prompt and checkpoint, let score probabilities be $v_0, \dots, v_m$. For $K \geq 2$ independent complete solutions, mixed scores occur with probability $1 - \sum_j v_j^K$, giving $\lambda_K(p) = 1 - p^K - (1-p)^K \sim Kp$ for binary verification. Moreover, when one score has mass $1 - u$ and $Ku \to 0$, an informative group contains one exceptional-score solution with probability $1 - O(Ku)$. Its conditional trajectory law then determines the optimizer update, with an explicit total-variation error (Appendix B.1). Thus rarity determines how often feedback arrives, while the verified reasoning paths determine its effect. Token dependence within a solution remains unrestricted. For heterogeneous prompts, both quantities must retain the prompt-selection weights, since averaging success probabilities before forming groups changes the feedback law.

4 Main Results

We first connect verified policy updates to subsequent feedback costs, then accumulate these costs across independent and dependent updates. Finally, we examine background motion, batching, and regularization under a common downstream accuracy target in every comparison.

4.1 Verified policy updates and recovery cost

From complete solutions to policy progress. We first ask how an optimizer step changes access to verified solutions. Let P_z^+ and P_z^- be the current trajectory laws conditioned on passing and failing a fixed binary verifier. For the actual update $z \to z'$, assume mutually absolutely continuous laws and write $L = dP_{z'}/dP_z$. Direct change of measure gives the log-odds increment and its cost factor,

$$D = \log \frac{\mathbb{E}_{P_z^+} L}{\mathbb{E}_{P_z^-} L}, \qquad e^{-\alpha D} = \left(\frac{\mathbb{E}_{P_z^-} L}{\mathbb{E}_{P_z^+} L} \right)^{\alpha}. \tag{3}$$

These expectations use fresh reference trajectories after fixing the update. They measure how probability moves between correct and incorrect solutions, including the effects of normalization, clipping, and optimizer memory. When $\lambda(p) \sim cp^{\alpha}$, the second ratio approximates the multiplier of the next waiting cost at low success probability. It thus links actual RLVR updates to the exponential moments governing the cost of recovery after regressions.

Proposition 2 (Accuracy improvement and feedback cost)**.** *A smooth, fully trainable two-parameter softmax policy over three complete solutions, with one verified correct solution, has the following property. Start at success probability $p = 0.01$, sample two independent solutions, and perform a centered-reward policy-gradient step. Conditional on mixed rewards, both expected success probability and expected success log-odds increase, while the expected waiting cost of the next mixed group also increases. Every success probability in $(0, 1)$ is attainable by this policy.*

The policy in Appendix B.2 gives equiprobable increments $D_- = -0.118868\dots$ and $D_+ = 0.130865\dots$. Thus $\mathbb{E}D > 0$, while $\mathbb{E}e^{-D} = 1.001779\dots > 1$. The exact mixed-group law $\lambda_2(p) = 2p(1-p)$ gives a next-wait cost ratio of $1.001745\dots > 1$: the harmful update makes recovery feedback sufficiently scarce to outweigh the cheaper wait after the favorable update. Filtering homogeneous groups preserves this example's policy path while hiding these waits from an accepted-update plot. The following laws account for their accumulated rollout cost.

The role of trajectory variability. More generally, the conditional score vectors of correct and incorrect solutions determine the exponential moment of a centered update. Appendix B.2 derives this expression for group centering, leave-one-out baselines, and scalar normalization, retaining the finite-step curvature term. Consequently, policies with the same accuracy can have different recovery costs because they distribute probability across different reasoning paths. The appendix also follows this moment through successive optimizer states, identifying how a regression can change the law of the next update as well as its arrival rate.

Verification and resource measurement. This distinction also matters with imperfect verification. Under class-conditional errors, false positives can maintain mixed-group activity while the correctness-directed mean gradient vanishes, provided true-success log-odds gradients remain bounded. Hence activity and actual progress require separate measurements. Increasing group

size likewise also changes both activity and sampling cost. Appendix G.5 shows that a controller can make mean group count finite while mean generated-trajectory cost remains infinite. We therefore compare a common accuracy target using generated data and retain unfinished runs when reporting finite-budget success probabilities and restricted means.

4.2 Signed updates and the cost of obtaining the next feedback

We next accumulate the costs across repeated feedback events. Suppose inactive attempts preserve the scalar state and an informative attempt changes its log-odds by D. If $\lambda(p) \sim cp^\alpha$ near zero, an increase of D multiplies the next expected wait by approximately $e^{-\alpha D}$. Accordingly, a regression and an equally large improvement have unequal effects on future cost. The exponent $\alpha > 0$ records how strongly feedback availability responds to success probability.

Let $D_1, D_2, \dots$ be independent with law ν, let $S_j = \sum_{i=1}^{j} D_i$ with $S_0 = 0$, and let E_j be independent unit exponentials. Define

$$W_{\alpha,\nu} = \sum_{j\geq 0} e^{-\alpha S_j} E_j, \qquad m(r) = \mathbb{E}e^{-r\alpha D}. \tag{4}$$

This is a classical exponential functional (Bertoin & Yor, 2005). The next theorem identifies its relation to a finite-initialization, finite-target training process, including returns to less successful states.

Theorem 3 (Signed activation law)**.** *Assume $\lambda(p)/(cp^\alpha) \to 1$ as $p \downarrow 0$ and $\lambda(p) \geq c_- p^\alpha$ for $0 < p < p_\star$. If active increments are independent draws from ν, $\mathbb{E}|D| < \infty$, and $\mathbb{E}D > 0$, then, as $\varepsilon \downarrow 0$,*

$$c\varepsilon^\alpha T_\star \longrightarrow W_{\alpha,\nu} < \infty \quad \textit{almost surely under a coupling}. \tag{5}$$

If $m(r) < 1$ for some real $r \geq 1$, convergence also holds in L^r under that coupling. In particular,

$$\mathbb{E}W_{\alpha,\nu} = \begin{cases} (1-m(1))^{-1}, & m(1) < 1, \\ \infty, & m(1) \geq 1. \end{cases} \tag{6}$$

The state-dependent extension and its explicit finite-initialization rate are stated in Appendix F.3.

The mean increment therefore determines eventual progress in the independent event chain, whereas its negative exponential moment determines the cost of waiting. The stated integrability conditions allow negative and unbounded increments. Theorem 16 further permits bounded increments whose law changes with the policy, with an explicit finite-initialization error. Does the obstruction persist at fixed initialization? Theorem 4 answers this for unit-sign updates, with independence recovered at $\rho = 0$ and its exact mean given in Appendix F.4.

The exact finite-group specialization. For a Bernoulli policy, verifier $R(a) = a$, group size $K \geq 2$, and one on-policy step using the within-group mean and standard deviation with variance divisor K,

$$N \sim \mathrm{Bin}(K, p), \quad \Delta\theta = g_K(N) = \eta\sqrt{\frac{N}{K}(1-\frac{N}{K})}, \quad \lambda_K(p) = 1 - p^K - (1-p)^K. \tag{7}$$

Homogeneous groups give zero update. Conditional on a rare informative group, the active-increment law concentrates on the singleton jump $\delta = \eta\sqrt{K-1}/K$. The state-dependent positive-increment bound in Appendix C yields

$$K\varepsilon T_\star \Rightarrow W_\delta = \sum_{j\geq 0} e^{-j\delta} E_j, \qquad \mathbb{E}W_\delta = \frac{1}{1-e^{-\delta}}, \qquad K\varepsilon \bar{T}_\star \longrightarrow \frac{1}{\delta}, \tag{8}$$

where $\bar{T}_\star$ uses the exact finite-group mean at every step. At $\delta = 1$, the two limiting mean costs are 1.582 and 1. Small jumps recover the deterministic limit $\delta W_\delta \to 1$ in mean square. Appendix D separately analyzes population normalization, which changes the mean field itself. These comparisons keep distinct two operations that are often conflated when interpreting a learning curve.

4.3 A DEPENDENCE THRESHOLD AT FIXED ACCURACY

The independent law isolates the cost of one update distribution. To see what changes when regressions cluster, we next keep that distribution fixed and vary dependence alone. Start from $0 < \varepsilon < p_\star < 1$. At feedback event j, let $C_j \in \{+1, -1\}$ follow the stationary Markov chain

$$R_\rho = \rho I + (1-\rho)\mathbf{1}\pi^\top, \qquad \pi = (q, 1-q)^\top, \quad 1/2 < q < 1, \quad 0 \le \rho < 1. \tag{9}$$

An event changes log-odds by δC_j, where $\delta > 0$. Inactive attempts preserve both variables. Assume $c_- p^\alpha \le \lambda(p) \le c_+ p^\alpha$ for $p < p_\star$, with positive constants. The sign distribution at each deterministic event index is $(q, 1-q)$ for every ρ. In particular, the event-count drift is always $\delta(2q-1) > 0$. Persistence changes only the dependence across those event indices.

Theorem 4 (Correlation-dependent activation threshold)**.** *Here $T_\star$ counts attempted rollout groups, while feedback-event count has an exponential tail. For the preceding process, define*

$$\xi_\rho = \frac{1}{\alpha\delta} \log \frac{q + \rho(1-q)}{1 - q + \rho q}. \tag{10}$$

At every fixed initialization and fixed target, $T_\star < \infty$ almost surely and

$$\mathbb{E}T_\star^r < \infty \iff 0 < r < \xi_\rho, \qquad \lim_{t\to\infty} \frac{\log \mathbb{P}(T_\star > t)}{\log t} = -\xi_\rho. \tag{11}$$

The classification includes divergence at $r = \xi_\rho$. For $\alpha = \delta = 1$ and $q = 0.8$, the finite-mean boundary is $\rho_c \approx 0.1298$, although all stationary one-update marginals agree.

The threshold is obtained from $M_r = R_\rho \operatorname{diag}(e^{-r\alpha\delta}, e^{r\alpha\delta})$, with states ordered as $(+1, -1)$: its spectral radius is below one exactly for $0 < r < \xi_\rho$. A negative excursion of m log-odds steps has probability of order $e^{-\alpha\delta\xi_\rho m}$ and makes feedback cost of order $e^{\alpha\delta m}$. Eliminating m gives the tail exponent. Appendix G proves these bounds directly for the stopped, discrete process and also gives its exact finite mean when $\lambda(p) = p$. The argument preserves a fixed accuracy target throughout.

Consequently, the fraction of favorable updates leaves their recovery cost undetermined. Here ρ measures persistence of actual progress signs at feedback indices. Its value belongs to this transition model and requires its own measurement. In a neural policy, shared representations, sampler history, and optimizer memory jointly determine these transitions. Appendix B.2 makes that dependence explicit through successive verified updates. Section 4.6 then isolates a different effect by matching conditional mean updates at every policy while varying their distribution.

4.4 IDENTICAL SHORT HISTORIES CAN HIDE DIFFERENT COST REGIMES

The threshold suggests examining recent progress signs. How much can such a history establish when the memory structure is unknown? The next result shows that even the complete distribution of every short update window can agree across processes with opposite mean-cost regimes.

Theorem 5 (Finite histories leave mean activation cost unidentified)**.** *Fix $0 < \varepsilon < p_\star < 1$, $\alpha, \delta > 0$, and any integer $L \ge 1$. There exist two stationary exponentially mixing sign processes $C_j \in \{-1, +1\}$ such that their laws of $(C_1, \dots, C_k)$ agree for every $k \le 2L+1$, with common positive mean $(4L+1)/(12L+1)$. Under the same rate $c_- p^\alpha \le \lambda(p) \le c_+ p^\alpha$, logit updates δC_j, and inactive steps that preserve the state, both reach $p_\star$ almost surely and need light-tailed numbers of feedback events. Nevertheless their attempted-group costs satisfy*

$$\mathbb{E}T_\star^{\mathrm{f}} < \infty, \qquad \mathbb{E}T_\star^{\mathrm{s}} = \infty.$$

The first distinguishing word has length $2L+2$, under finite-state irreducible aperiodic kernels.

To construct the pair, we concatenate favorable and unfavorable blocks whose boundaries look identical within a short window. Repeated unfavorable blocks nevertheless lower the probability of obtaining recovery feedback. Their frequency can therefore change total cost while preserving each specified window law (Appendix G.7). Thus qualitative mixing alone supplies no budget guarantee. Quantitative restrictions on memory or adverse excursions are needed, while waiting timestamps provide information beyond the matched update windows.

4.5 A FINITE-BUDGET CONDITION WITH FULL-STATE BACKGROUND MOTION

The preceding laws hold the policy fixed between selected feedback events. In RLVR, however, other prompts, regularization, and optimizer memory can continue to move it. To retain this motion, we return to the full training state and define

$$V(z) = e^{-\alpha\theta(z)}, \qquad \rho_i(z) = \mathbb{E}[e^{-\alpha(\theta(Z_{n+1})-\theta(z))} \mid Z_n = z, I_n = i]. \tag{12}$$

These quantities concern the actual post-update success probability, so cross-task interactions and all trainable layers enter the same conditional measurement of progress.

Theorem 6 (Activation with signed feedback and background motion)**.** *Assume the conditional moments in (12) are finite. If, at every reachable state with $p(z) < p_\star$,*

$$V(z)\{\lambda(z)(1-\rho_1(z)) + (1-\lambda(z))(1-\rho_0(z))\} \geq a > 0, \tag{13}$$

then, at any fixed initialization,

$$\mathbb{E}T_\star \leq V(Z_0)/a, \qquad \mathbb{P}(T_\star > n) \leq \min\{1, V(Z_0)/(an)\}. \tag{14}$$

A sufficient set of conditions is $\lambda(z) \geq c_- p(z)^\alpha$, $\rho_1(z) \leq \bar{\rho}_1 < 1$, and $(1-\lambda(z))(\rho_0(z)-1) \leq b\lambda(z)$ with $0 \leq b < 1-\bar{\rho}_1$. An explicit choice of the uniform drift constant is $a = c_-(1-p_\star)^\alpha(1-\bar{\rho}_1-b)$.

The two terms in (13) account separately for feedback and background progress. Their sum yields the budget bound by additive drift, even when individual updates have either sign. This global guarantee requires control throughout the reachable pre-target region. For local audits, Appendix G.10 accumulates progress over short blocks and allows adverse episodes. Independent continuations from restored checkpoints estimate the conditional quantities. Subsequent training logs then test the resulting forecasts against observed progress and waiting costs.

The terms $\rho_0 < 1$ and $\rho_0 > 1$ distinguish favorable and adverse background motion. Near zero, the sufficient condition requires the adverse exponential-moment excess per inactive step to scale as $O(\lambda)$. A fixed adverse drift violates this balance. Appendix F gives localized exit guarantees, a background-transfer example, and robustness bounds for measured conditional moments. Every comparison counts elapsed steps toward the same downstream accuracy target.

For independent groups of K solutions with binary verification, $\lambda_K(p)$ is known exactly and $\alpha = 1$ near zero. The remaining measurements concern actual policy progress. At restored checkpoints, evaluate success before and after complete optimizer steps, retaining momentum and inactive motion. The resulting log-odds increments estimate the conditional moments in (12), while separate rollout cohorts test feedback forecasts. Appendix G.9 collects these measurements in one protocol and distinguishes local estimates from the additional uniform control needed for a global guarantee.

4.6 EQUAL FRESH DATA AND EQUAL MEAN UPDATES LEAVE PROGRESS UNIDENTIFIED

Having separated the sources of cost, now we can ask whether equal data and equal mean updates imply equal progress here. At each step, draw B independent fresh groups of K trajectories and compute all normalized updates $G_1, \ldots, G_B$ from (7). Compare their uniform average with

$$\Delta\theta_{\boldsymbol{w}} = \frac{1}{B}\sum_{j=1}^{B} w_{\sigma(j)} G_j, \qquad w_j > 0, \quad \sum_{j=1}^{B} w_j = B, \tag{15}$$

where σ is a fresh uniform permutation independent of the groups. Every trajectory contributes to a group gradient, and every gradient has positive weight. Averaging over σ gives the uniform pre-optimizer gradient conditional on the entire pool. The specified linear update preserves this identity. Adaptive coordinate scaling, clipping, and Adam require a separate post-optimizer calculation.

Theorem 7 (Equal-data, matched-mean separation)**.** *The two updates have identical conditional means at every policy and use the same BK fresh trajectories. Let J be uniform on $\{1, \ldots, B\}$ and $D_{\boldsymbol{w}} = \delta w_J/B$. Here $T_\star^{\boldsymbol{w}}$ counts attempted batch steps, each using BK trajectories. For fixed B,*

$$K\varepsilon T_\star^{\boldsymbol{w}} \Rightarrow B^{-1} W_{1,\mathcal{L}(D_{\boldsymbol{w}})}, \qquad \lim_{\varepsilon\downarrow 0} \mathbb{E}[K\varepsilon T_\star^{\boldsymbol{w}}] = \frac{1}{B[1 - B^{-1}\sum_{j=1}^{B} e^{-\delta w_j/B}]}. \tag{16}$$

The trajectory count is $BKT_\star^{\boldsymbol{w}}$, whose ε-scaled limit is $W_{1,\mathcal{L}(D_{\boldsymbol{w}})}$. Uniform weights uniquely minimize this limiting mean among positive weights with sum B. Nonuniform weights separate activation cost while preserving the complete conditional mean update at each policy.

For $B = 4$ and $\delta = 1$, uniform weights yield limiting mean cost 1.1302, while a random permutation of $(0.25, 0.25, 0.25, 3.25)$ yields 1.3550. These are analytical predictions with identical fresh-sample and group-gradient counts. The intervention changes the distribution of progress conditional on an informative batch. Convexity of the future waiting cost then converts mean-preserving update variability into an increase in activation cost. Both values quantify the effect of update variability at a fixed downstream accuracy target under on-policy sampling.

Appendix D retains the original independent-versus-duplicated construction as a separate dependence example. That construction changes the number of unique trajectories and therefore answers a different question. The equal-data theorem establishes the main nonidentification statement without relying on repeated trajectories. Dynamic sampling likewise changes the accepted-update clock (Yu et al., 2025). A rollout-budget comparison must still count groups generated before acceptance and retain any background updates during those attempts in the budget.

4.7 Regularization acts during the waiting period

Batching changes feedback-driven progress. A regularization penalty also acts between those events, so its accumulated effect depends on how long feedback takes to arrive. Appendix E studies the exact logit update $\Delta\theta = g_K(N) - \eta\gamma\partial_\theta D_f(p\|\varepsilon)$ for a fixed reference and f-divergences matched at $f''(1) = 1$. With relative log-odds $x = \theta - \operatorname{logit}\varepsilon$, the rare-feedback limit has random reward jumps and deterministic restoring motion. In time measured by accumulated feedback rate, the jumps have rate one and the restoring force is $(\eta\gamma/K)f'(e^x)$. Comparing these forces gives the finite-relative-level order in Theorem 13. For an actual finite initialization and accuracy target, the relevant check instead uses the exact penalty derivative on the visited interval. Appendix E.4 supplies a discrete reachability condition, and the experiments test finite targets directly. Thus equal local curvature fixes a common scale, while the complete restoring force determines motion during waiting periods.

5 Experiments

Design. The theory predicts both a sampling cost and the mechanisms that change it. We first isolate temporal dependence and update variability, then test the effects of a trainable representation, Adam, and regularization. Each comparison uses a common accuracy target and retains all unfinished runs in its distributions and restricted means. Finally, independent language-model outputs test the feedback-arrival law used in the analysis. Appendix H reports all conditions and numerical-resolution checks.

Dependence changes recovery cost. For $q = 0.8$, $\alpha = \delta = 1$, and $\rho \in \{0, 0.1, 0.2, 0.5\}$, the predicted moment thresholds are 1.386, 1.075, 0.847, and 0.405. Every process has favorable-update frequency 0.8 and mean increment 0.6. At $\varepsilon = 10^{-4}$, the fractions still waiting at normalized budget 100 are 0.47%, 1.07%, 2.10%, and 8.995%. At $\rho = 0.1$, even a horizon of 10^4 gives capped mean 9.248, against exact unrestricted mean 20.820. Finite reports can substantially understate recovery cost while preserving all unfinished runs.

Short histories conceal costly recovery. Exact enumeration verifies identical history laws through lengths 3, 5, 9, and 17. Each law has 20,000 runs to target $p_\star = 0.2$. In the last pair, median normalized costs are 47.567 and 48.097. At budget 10^6, every sampled fast run finishes, while 9.63% of slow runs remain. Capped means are 921.9 and 116056.2 (Figure 2). The nearly equal medians coexist with mean-cost regimes established by the theorem, independently of these finite measurements.

Equal fresh data and conditional means. With $K = B = 4$, $\delta = 1$, and $\varepsilon = 10^{-4}$, uniform and skewed weighting give normalized mean costs 1.1325 and 1.3512, versus predictions 1.1302 and 1.3550. Their coefficients of variation are 0.3516 and 0.5263, versus 0.3526 and 0.5266. Both generate 16 fresh trajectories and compute all four group gradients at every attempt. Exact enumeration verifies equal mean updates, isolating conditional variability at fixed data access and mean field.

Separating feedback and optimizer motion. We next apply the equal-data intervention to a fully trainable policy. Skewed weighting raises mean SGD activation cost in all six pairs by factors 1.075–1.530. At $\varepsilon = 0.001$ and step size 0.6, the ratio 1.530 decomposes into 1.128 times as many active events and 1.356 times the pooled cost per event. To test optimizer memory, the

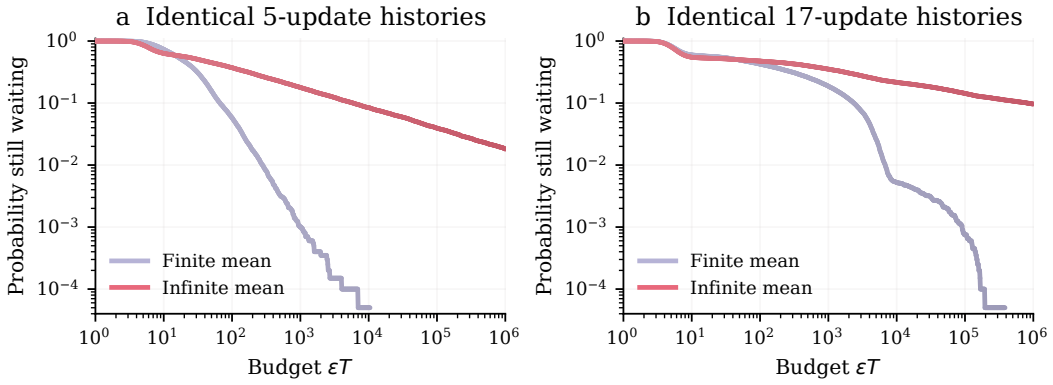


Figure 2: **The same short histories can conceal opposite cost regimes.** Each pair has identical distributions of every sign history through the stated length and the same positive average progress. Curves show 20,000 runs per law, with all unfinished runs retained. The finite/infinite mean labels follow from Theorem 5. The curves measure finite-budget survival. At right, the medians are nearly equal, while 9.63% of the infinite-mean cohort still waits at normalized budget 10^6.

Adam runs execute every step, including zero-gradient steps, and retain their moment buffers. Costs increase by 7.4%–30.5% across all four Adam pairs. Thus the effect persists in this adaptive setting, where the matched quantity is the raw conditional gradient before Adam's nonlinear transformation. Table 14 reports all pairs and inactive parameter motion.

Finite targets and background transfer. We also test the two ways a policy can move while selected feedback is absent. All 16 discrete regularization cohorts reach their targets within budget. At $\varepsilon = 0.003$ and $p_\star = 0.2$, Jensen–Shannon and Hellinger median costs are 1.968 and 1.944, showing why the finite process requires its own comparison. With shared-task learning, following the evolving hazard gives first-feedback prediction error 0.0042, versus 0.936 when freezing it. Appendix H supplies the full matrices and resolution checks.

Prospective feedback forecasts on a language model. We collect 4,608 fresh completions from Qwen2.5-3B-Instruct and a GRPO-family fine-tuned checkpoint on 12 preselected MATH test questions (Qwen et al., 2024; Hendrycks et al., 2021). Independent forecast and observation cohorts test group sizes four and eight. Across all four comparisons, predicted mean mixed-group activity differs from observation by 0.05–2.22 percentage points. The declared decoder includes both naturally terminated and length-capped answers. Thus this audit tests feedback arrival under independent generation, complementing the neural experiments on post-update recovery cost. Appendix H.11 reports all questions, waiting forecasts, and simultaneous 95% confidence intervals for the checkpoint comparisons.

Question-level forecasts use each prompt's success rate before averaging. Each group answers one question, so contrast must be evaluated within prompts. This distinction matters because an easy and a hard question can have intermediate aggregate accuracy while each produces few mixed groups. A forecast based only on that average would misrepresent feedback availability. By holding prompts and decoding rules fixed, the checkpoint comparison tests policy-induced changes in verified sampling, with uncertainty assessed separately for the corresponding predictions.

6 Conclusion

An RLVR policy determines both its progress and the feedback available for its next update. Consequently, favorable average movement can coexist with expensive recovery from regressions. We characterize this tension through exact dependence thresholds and finite-history separations, then retain background learning in finite-budget bounds. Equal-data comparisons and adaptive neural experiments trace how update variability changes cost, while language-model audits test feedback arrival. Together, these results make the generated-data budget an essential part of understanding how reasoning policies learn from the verified outcomes of their own solutions.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of ACL*, pp. 12248–12267, 2024. https://aclanthology.org/2024.acl-long.662/.

Jean Bertoin and Marc Yor. Exponential functionals of Lévy processes. *Probability Surveys*, 2:191–212, 2005. https://arxiv.org/abs/math/0511265.

Nicholas Hugh Bingham, Charles M. Goldie, and Jozef Lodewijk Maria Teugels. *Regular Variation*. Cambridge University Press, 1987. https://doi.org/10.1017/CBO9780511721434.

Mark Herbert Ainsworth Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society, Series B*, 46(3):353–376, 1984. https://doi.org/10.1111/j.2517-6161.1984.tb01308.x.

Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning? *arXiv:2510.13651*, 2025. https://arxiv.org/abs/2510.13651.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*, 2025. https://arxiv.org/abs/2501.12948v1.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks*, 2021. https://arxiv.org/abs/2103.03874.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the emergence of implicit curriculum in RLVR learning dynamics. *Proceedings of ICML*, 2026. https://arxiv.org/abs/2602.14872.

Haoyu Hu, Xuandong Zhao, Xuhai "Orson" Xu, and Nori Jacoby. DUET: Optimize token-budget allocation for reinforcement learning with verifiable rewards. *arXiv:2605.08441*, 2026. https://arxiv.org/abs/2605.08441.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021. https://arxiv.org/abs/2106.09685.

Juhani Karhumäki, Aleksi Saarela, and Luca Q. Zamboni. On a generalization of Abelian equivalence and complexity of infinite words. *arXiv:1301.5104*, 2013. https://arxiv.org/abs/1301.5104.

Soeun Kim and Albert No. Where rollouts begin: Low-load, high-leverage first-token diversification for RLVR. *arXiv:2605.28295*, 2026. https://arxiv.org/abs/2605.28295.

Diederik Pieter Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015. https://arxiv.org/abs/1412.6980.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of SOSP*, 2023. https://arxiv.org/abs/2309.06180.

Long Li, Zhijian Zhou, Tianyi Wang, Weidi Xu, Zuming Huang, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. DyJR: Preserving diversity in reinforcement learning with verifiable rewards via dynamic Jensen–Shannon replay. *arXiv:2603.16157*, 2026. https://arxiv.org/abs/2603.16157.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv:2305.20050*, 2023. https://arxiv.org/abs/2305.20050.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv:2503.20783*, 2025. https://arxiv.org/abs/2503.20783.

Xiaodong Lu, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Zhijun Chen, Yu Luo, Fuzhen Zhuang, Yikun Ban, and Deqing Wang. Contextual rollout bandits for reinforcement learning with verifiable rewards. *arXiv:2602.08499*, 2026. https://arxiv.org/abs/2602.08499.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. https://doi.org/10.1214/aop/1176990746.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 2021. https://arxiv.org/abs/2110.15572.

Arpan Mukherjee, Marcello Bullo, Debabrota Basu, and Deniz Gündüz. Test-time verification via optimal transport: Coverage, ROC, & sub-optimality. *Proceedings of ICLR*, 2026. https://arxiv.org/abs/2510.18982.

Wenhua Nie, Jianan Wu, Junlin Liu, Ziwei Li, Zheng Lin, Zhang Zijian, Yilong Fan, Haoran Zheng, and Jyh-Shing Roger Jang. Gradient starvation in binary-reward GRPO: Why group-mean centering fails and why the simplest fix works. *arXiv:2605.07689*, 2026. https://arxiv.org/abs/2605.07689.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. https://arxiv.org/abs/2203.02155.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv:2412.15115*, 2024. https://arxiv.org/abs/2412.15115.

Alexander Roitershtein. One-dimensional linear recursions with Markov-dependent coefficients. *The Annals of Applied Probability*, 17(2):572–608, 2007. https://doi.org/10.1214/105051606000000844.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*, 2024. https://arxiv.org/abs/2402.03300.

Joe Suk and Yaqi Duan. On the optimization dynamics of RLVR: Gradient gap and step size thresholds. *arXiv:2510.08539*, 2025. Version 4, 2026. https://arxiv.org/abs/2510.08539.

Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. *arXiv:2509.21519*, 2025. https://arxiv.org/abs/2509.21519.

Akifumi Wachi, Hirota Kinoshita, Shokichi Takakura, Rei Higuchi, and Taiji Suzuki. A relative-budget theory for reinforcement learning with verifiable rewards in large language model reasoning. *arXiv:2602.01523*, 2026. https://arxiv.org/abs/2602.01523.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D. Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. *Proceedings of the Conference on Learning Theory*, 291:5632–5711, 2025. https://proceedings.mlr.press/v291/wang25a.html.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv:2503.14476*, 2025. https://arxiv.org/abs/2503.14476.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Proceedings of COLT*, volume 65, pp. 1980–2022, 2017. https://proceedings.mlr.press/v65/zhang17b.html.

Heming Zou, Qi Wang, Yun Qu, Yuhang Jiang, Lizhou Cai, Yixiu Mao, Ru Peng, Xin Xu, Weijie Liu, Kai Yang, Saiyong Yang, and Xiangyang Ji. TRACE: A unified rollout budget allocation framework for efficient agentic reinforcement learning. *arXiv:2606.11119*, 2026. https://arxiv.org/abs/2606.11119.

Appendix

CONTENTS

A Notation and Scope **14**

B Feedback Kernels and Waiting Times **14**

B.1 Autoregressive trajectories and budget accounting 15

B.2 Verified policy updates and feedback costs 17

C The Activation Law and its Error Bound **20**

C.1 Geometric waiting and a finite event cutoff 21

C.2 Coupling up to a progress threshold 21

C.3 Remainders and convergence of moments 22

C.4 Moment calculation 22

C.5 Budget probabilities and sharp moment bounds 22

D Finite Groups and Matched-Mean Batches **23**

D.1 Population normalization and exact conditional means 23

D.2 The duplicated-group dependence intervention 23

D.3 Deriving the normalized update 24

D.4 Equal-data weighting and its exact activation law 25

E Regularized Limits and Escape **26**

E.1 Regularization changes finite-target progress 26

E.2 The divergence gradient and its scaling 27

E.3 Curvature-matched divergence generators 28

E.4 Finite-level comparison and exact discrete reachability 28

E.5 Proof of the escape classification 29

F Scope, Approximation, and Observable Diagnostics **30**

F.1 Proof of the background-transfer example 31

F.2 A quantitative check on freezing the background 32

F.3 Signed activation and finite-initialization error 32

F.4 Exact finite-initialization mean for signed unit increments 34

F.5 Full-state activation and conditional-moment errors 35

F.6 Graded verifiers, batching, and optimizer memory 36

G Temporal Dependence and Fixed-Accuracy Thresholds **36**

G.1 The embedded chain and tilted matrix 36

G.2 Finite moments and almost-sure activation 37

G.3 A lower tail bound including the critical moment 37

G.4 An exact finite-initialization mean 38

G.5 Arbitrary feedback rates and resource costs 39

G.6 Distributional limit and its numerical tail certificate 41

G.7 Finite-window matching and opposite cost regimes 41

G.8 Noise-aware conditional-moment certificates 43

G.9 Practical estimation and forecasts 44

G.10 Finite-budget certificates with signed blocks 46

G.11 Sampling controllers 47

H Experimental Protocol and Complete Results **48**

H.1 Exact finite-group simulation 48

H.2 Regularization and numerical resolution 49

H.3 A shared-network prediction with independent cohorts 50

H.4 Signed and state-dependent feedback 51

H.5 Correlated signs at fixed initialization 51

H.6 Identical finite-window histories 53

H.7 An equal-fresh-data intervention 54

H.8 Fully trainable neural activation 54

H.9 Adaptive-optimizer audit 55

H.10 Fixed-accuracy divergence comparisons 57

H.11 Language-model checkpoint audits 57

I Ethics, Reproducibility, and AI Use **59**

A Notation and Scope

Table 1: Important notation. Times count attempted transitions unless another clock is explicitly specified.

Symbol	Meaning
Z_n, P	Complete training state and its transition kernel
I_n, $\lambda(z)$	Selected feedback event and conditional event probability
K_0, K_1	Conditional transitions without and with selected feedback
L_0	Substochastic kernel retaining paths without selected feedback
T_1, $T_\star$	First-feedback time and success-threshold activation time
ε, $p_\star$	Initial success probability and fixed success target
ν_p, ν, D_j	Active-increment law, its limiting law, and a limiting draw
α, $m(r)$	Feedback-rate exponent and $\mathbb{E}e^{-r\alpha D}$
C_j, R_ρ, π	Feedback-index sign, transition matrix, stationary sign law
ρ, M_r, ξ_ρ	Sign persistence, tilted matrix, critical moment order
χ, ℓ, $c(p)$	Spatial decay $\log(A/B)$, slowly varying factor, mean attempt cost
Ψ, B_0, T_B	Capped reciprocal-odds potential, cap, and block-boundary hitting time
V, ρ_i, a	Reciprocal-odds potential, conditional exponential moments, drift bound
$\boldsymbol{w}$, σ	Positive group weights and their independent permutation
S_j, E_j, W_ν	Cumulative increments, exponential waits, and limiting activation time
K, B, η, δ	Group size, group count, step size, singleton-success jump
D_f, γ, κ	Divergence penalty, its coefficient, and $\eta\gamma/K$
X, u, t	Relative log-odds, operational time, normalized physical time
ψ_f, ψ_∞	Restoring force $f'(e^x)$ and its large-x limit
L, Q_ω, P_L	Window parameter, block transition kernel, and hidden-state count
b, H_m	Background log-odds growth rate and harmonic sum $\sum_{j=1}^{m} 1/j$
d_{BL}, d_{TV}, CV	Bounded-Lipschitz distance, total variation, coefficient of variation

Logical dependencies. Appendix C then couples the waiting times to an exponential functional, and Appendix D applies that result to finite groups. Next, Appendix E retains motion during the waits. Appendix F develops signed and full-state extensions, which make explicit the assumptions needed beyond the positive-increment model. Finally, Appendix G proves the fixed-target dependence and finite-history results and explains their measurement. Appendix H reports the corresponding experimental protocols and complete outcomes.

B Feedback Kernels and Waiting Times

We begin with the exact training process. Conditioning on whether the selected feedback arrives gives the waiting-time identity. Conditioning further on verifier outcomes then identifies which complete solutions drive the update. Both steps preserve the policy and optimizer state used in training.

State and conditional distributions. We take the state space to be standard Borel, so regular conditional distributions exist. Parameters, finite or countable token histories, random-number state if retained, optimizer variables, and discrete iteration counters fit this setting. Conditional kernels on events of conditional probability zero can be defined arbitrarily, since their mixture coefficient is zero. Time-inhomogeneous training becomes homogeneous after adjoining the iteration index. The indicator $\mathbf{1}$ in Proposition 1 denotes the constant-one function.

Proof of Proposition 1. Let $s_n(z) = \mathbb{P}_z(T_1 > n)$. We have $s_0 = \mathbf{1}$. Conditioning on the first transition, survival requires $I_0 = 0$, which has probability $1 - \lambda(z)$. Conditional on that event, the

next state follows K_0. The Markov property therefore gives

$$s_{n+1}(z) = (1 - \lambda(z)) \int s_n(z')K_0(z, dz') = L_0 s_n(z).$$

Induction yields $s_n = L_0^n \mathbf{1}$. Expanding the iterated integrals produces the product expectation in (2). This expectation is over the conditional background kernel K_0, with the survival weights retained explicitly.

When $K_0(z, \cdot) = \delta_z$, the state remains z until a selected event occurs. The probability of a wait of exactly $m \geq 1$ attempted transitions and a subsequent state in a measurable set A is

$$(1 - \lambda(z))^{m-1} \lambda(z) K_1(z, A).$$

This factorization makes the waiting time geometric on $\{1, 2, \ldots\}$ and independent of the subsequent active transition conditional on z. Iterating this factorization gives conditional independence of waits given the embedded state path. If $\lambda(z) = 0$, the wait is infinite and the embedded chain cannot leave that state through the selected event. □

A target-set recursion. The first-feedback calculation extends naturally to reaching a target after several updates. Define $v_n(z) = \mathbb{P}_z(T_A > n)$, with $v_0 = \mathbf{1}_{A^c}$. Conditioning on the first transition now retains both event outcomes and gives

$$v_{n+1}(z) = \mathbf{1}_{A^c}(z) \left[(1 - \lambda(z)) \int v_n(z')K_0(z, dz') + \lambda(z) \int v_n(z')K_1(z, dz') \right].$$

Thus a feedback event may move the policy either toward or away from the target. Meanwhile, the background kernel may move it even before that event arrives. The scalar laws developed below solve particular rare-feedback limits of this recursion, while the full-state bounds retain both kinds of movement.

B.1 Autoregressive trajectories, estimators, and token budgets

To obtain the event probability from an actual language-model rollout, let a stopped autoregressive trajectory have law P_z and terminal verifier $R \in \{0, 1\}$. At a fixed prompt and checkpoint, independent trajectories produce independent Bernoulli outcomes with parameter $p = P_z(R = 1)$. Consequently, summing over the entire prefix tree gives

$$N = \sum_{j=1}^{K} R_j \sim \mathrm{Bin}(K, p), \qquad \mathbb{P}(0 < N < K) = 1 - (1 - p)^K - p^K.$$

Thus token dependence within each trajectory leaves the group-level identity exact. A finite generation cap becomes part of P_z and the verifier convention. Correlation between separate trajectories, adaptive stopping within a group, or policy changes during generation require their actual joint law. For heterogeneous prompts drawn once per group, the mixed-group probability is $\mathbb{E}_x[\lambda_K(p_x)]$, which generally differs from $\lambda_K(\mathbb{E}_x p_x)$.

For the explicitly filtered event $r \leq N < K$, where $1 \leq r < K$, the leading probability is $\binom{K}{r} p^r$. By contrast, in a specified d-stage model with terminal success $p = q^d$, first-stage feedback has probability $q = p^{1/d}$. These examples change either the acceptance event or the feedback source. A longer reasoning chain, or graded rewards by themselves, specify neither change. Dense process supervision supplies additional labels and must be represented by a different event and background kernel (Lightman et al., 2023).

Graded verification and the conditional update law. Fix the complete training state z and a prompt. Let P_z be the law of a complete generated solution, including the stopping rule, and let R take distinct values $r_0, \ldots, r_m$ with probabilities v_j. Draw $K \geq 2$ solutions independently from this same law. The event I that their scores differ has probability

$$\lambda(z) = \mathbb{P}_z(I) = 1 - \sum_{j=0}^{m} v_j^K.$$

This is an exact finite-label identity. The scores can be partial-credit grades from a test suite, as well as a binary correctness check. If a verifier uses independent randomization, include that randomness in the sampled solution record. Correlated verifier errors across solutions require the joint law instead of this product construction.

Suppose $v_0 = 1 - u$ with $0 < u < 1$, and let N count scores different from r_0. Let $\widetilde{K}_1(z, \cdot)$ be the following conditional update law: draw one solution from $P_z(\cdot \mid R \neq r_0)$ and $K - 1$ independently from $P_z(\cdot \mid R = r_0)$, place the exceptional solution at a uniform random group position, then apply the actual update rule to the resulting group and the stored optimizer state. This is an analytical representation of a conditional group, with the same optimizer as training. It requires no generative access to the conditional laws. If the optimizer has additional randomization, use the same conditional randomization in both kernels.

Proposition 8 (One exceptional score in an informative group)**.** *For the true conditional kernel* $K_1(z, \cdot) = \mathcal{L}(Z_{n+1} \mid Z_n = z, I)$*,*

$$d_{\mathrm{TV}}(K_1(z, \cdot), \widetilde{K}_1(z, \cdot)) \leq \min\left\{1, \frac{(K-1)u}{2(1-u)^{K-1}}\right\}. \tag{17}$$

Moreover $\lambda(z) = Ku + O(K^2u^2)$ *when* $Ku \to 0$*. The bound remains valid after mapping the updated state to success probability, log-odds, or any other measurable progress variable.*

Proof. The event $N = 1$ implies I and has probability $Ku(1-u)^{K-1}$. Conditional on $N = 1$, independence gives exactly the group law defining $\widetilde{K}_1$. Thus the conditional law given I is a mixture containing that law with weight $\mathbb{P}(N = 1)/\mathbb{P}(I)$. Its remaining weight is at most

$$\frac{\mathbb{P}(N \geq 2)}{\mathbb{P}(N = 1)} \leq \frac{\binom{K}{2}u^2}{Ku(1-u)^{K-1}} = \frac{(K-1)u}{2(1-u)^{K-1}}.$$

The numerator bound follows from $\mathbf{1}\{N \geq 2\} \leq N(N-1)/2$. Applying a common randomized optimizer cannot increase total variation, proving the kernel statement and its measurable-image version. Finally,

$$Ku(1-u)^{K-1} \leq \lambda(z) \leq \mathbb{P}(N \geq 1) \leq Ku.$$

The difference between the extreme bounds is at most $K(K-1)u^2$, which proves the expansion. □

For a reliable binary verifier with rare success, the exceptional solution is correct. Its conditional trajectory distribution determines the active update, while the total mass p determines how often it can be sampled. Two equally accurate policies can therefore induce different updates because their correct reasoning paths differ. With graded verification, the same argument uses the mass outside the dominant score, which may differ from full-answer accuracy. The reduction itself is statewise and allows unrestricted token dependence and trainable representations. To obtain a limiting increment law as the policy changes, one must additionally control the conditional trajectory laws and optimizer map.

Leave-one-out and critic-based feedback. For terminal rewards, the leave-one-out advantage is

$$A_i = R_i - \frac{1}{K-1}\sum_{j \neq i} R_j = \frac{K}{K-1}(R_i - \bar{R}).$$

It vanishes on homogeneous reward groups, exactly as the centered reward component of a group-relative update. Leave-one-out estimation can change variance and update size without changing this event's arrival probability (Ahmadian et al., 2024). With a learned critic, token advantages, entropy terms, or a divergence penalty, homogeneous terminal scores can still yield a nonzero update (Ouyang et al., 2022). Such updates belong in K_0 for this chosen contrast event. Consequently the full-state decomposition applies, while an inactive-holding scalar law requires a separate check. The comparison here is algebraic, with no empirical performance ranking among training algorithms.

Generated tokens are a separate clock. Let C_n be the tokens generated during attempted group n, and let $\mathcal{C}_\star = \sum_{n<T_\star} C_n$. If $c_{\min} \leq C_n \leq c_{\max}$, then $c_{\min}T_\star \leq \mathcal{C}_\star \leq c_{\max}T_\star$ pathwise. More generally, conditional expected costs uniformly between positive constants give the analogous expectation bounds by conditioning on the pre-attempt state and summing nonnegative terms. A fixed generation cap supplies an upper bound. Without such bounds, state-dependent completion lengths contribute another waiting-cost mechanism.

B.2 VERIFIED POLICY UPDATES AND FEEDBACK COSTS

We now use those conditional trajectory laws to connect a realized update to future feedback. The first identity holds for the complete optimizer step. The explicit softmax example then shows why an accurate update can nevertheless make recovery feedback more expensive.

An exact identity for the complete optimizer step. Fix a prompt, a parameter-independent correctness event A, and the current trajectory law P_z, with $p = P_z(A) \in (0,1)$. For any realized updated state z', suppose $P_{z'}$ and P_z are mutually absolutely continuous and let $L = dP_{z'}/dP_z$. The expectations below are over fresh reference trajectories after fixing z' and use no reuse of the group that produced that state. Direct integration gives

$$p' = P_{z'}(A) = p\mathbb{E}_{P_z(\cdot|A)}L, \qquad 1-p' = (1-p)\mathbb{E}_{P_z(\cdot|A^c)}L.$$

Taking their ratio proves (3). An autoregressive model with strictly positive token probabilities on a common capped support satisfies the support condition. A change in hard decoding support requires a separate change-of-measure treatment. The identity is otherwise independent of how the optimizer obtains z'. In particular, it covers tokenwise and sequencewise objectives, adaptive scaling, replay, and penalties as part of the realized update, without asserting that their gradient estimators agree.

Proof of Proposition 2. Consider three complete solutions c, f_+, f_- with verifier $R(c) = 1$ and $R(f_+) = R(f_-) = 0$. For trainable parameters $(u, w) \in \mathbb{R}^2$, assign softmax logits

$$\ell_c = \log\varepsilon + u/10 + w, \qquad \ell_{f_+} = \log((1-\varepsilon)/2) + 2w, \qquad \ell_{f_-} = \log((1-\varepsilon)/2) - 2w.$$

All probabilities are strictly positive. The success log-odds equal

$$\theta(u,w) = \operatorname{logit}\varepsilon + u/10 + w - \log\cosh(2w). \tag{18}$$

At $w = 0$, varying u attains every success probability in $(0,1)$. Thus the policy has no representational ceiling on the accuracy target. Both parameters are updated in the calculation below.

Initialize $(u,w) = (0,0)$ and use the ordinary centered-reward gradient

$$G = \frac{1}{2}\sum_{i=1}^{2}(R_i - \bar{R})\nabla_{(u,w)}\log P_{u,w}(\tau_i), \qquad (u',w') = (u,w) + \frac{2}{5}G.$$

A homogeneous group gives $G = 0$. Conditional on mixed rewards, there is one correct solution and one failure, with f_+ and f_- equally likely. The common softmax normalizer cancels from the difference of the two score vectors. For f_+ the feature difference is $(1/10, -1)$, and for f_- it is $(1/10, 3)$. Consequently

$$G = \tfrac{1}{4}(1/10,-1) \quad \text{or} \quad G = \tfrac{1}{4}(1/10,3), \qquad (u',w') = (1/100,-1/10) \quad \text{or} \quad (1/100,3/10).$$

Substituting in (18) gives the exact two-point law

$$D_- = \frac{1}{1000} - \frac{1}{10} - \log\cosh\frac{1}{5}, \qquad D_+ = \frac{1}{1000} + \frac{3}{10} - \log\cosh\frac{3}{5}.$$

Define $a = e^{D_-}$, $b = e^{D_+}$, $M_+ = (a+b)/2$, and $M_- = (a^{-1}+b^{-1})/2$. These constants satisfy the strict bounds

$$\begin{gathered} 0.00599 < \frac{D_- + D_+}{2} < 0.00601, \\ 1.01386 < M_+ < 1.01388, \qquad 1.00177 < M_- < 1.00179. \end{gathered} \tag{19}$$

These inequalities can be established directly by the exponential power series. For $0 \le x \le 1$, summing through degree 12 bounds e^x from below, and adding $x^{13}/[13!(1-x/14)]$ bounds it from above. Reciprocals give the bounds for negative exponents. Applying these bounds to the displayed formulas for a, b, and ab proves $ab > 1$, hence the positive mean log-odds claim, and the stated numerical intervals. The decimal values are evaluations of exact expressions, independent of any sampled experiment.

For initial odds $o = \varepsilon/(1-\varepsilon)$, direct subtraction gives

$$\frac{1}{2}\left(\frac{oa}{1+oa} + \frac{ob}{1+ob}\right) - \frac{o}{1+o} = \frac{o[(M_+ - 1) - oab(M_- - 1)]}{(1+o)(1+oa)(1+ob)}. \tag{20}$$

The right-hand side is positive whenever $o < (M_+ - 1)/[ab(M_- - 1)]$, whose value lies between 7.6 and 7.8. In particular it is positive for $\varepsilon = 0.01$, where the mean new success probability is $0.0101357\ldots$. This proves expected accuracy improvement as well as expected log-odds improvement.

After the update, keep applying the same two-sample rule. Until the next mixed group, its centered reward gradient is zero and the parameters stay fixed. The conditional waiting time is therefore geometric with mean $1/\lambda_2(p')$. For a realized log-odds change D, elementary algebra gives

$$\frac{\lambda_2(\varepsilon)}{\lambda_2(p')} = (1-\varepsilon)^2 e^{-D} + 2\varepsilon(1-\varepsilon) + \varepsilon^2 e^{D}.$$

Taking expectations yields

$$\lambda_2(\varepsilon)\mathbb{E}\frac{1}{\lambda_2(p')} = 1 + (1-\varepsilon)^2(M_- - 1) + \varepsilon^2(M_+ - 1) > 1. \tag{21}$$

This increase holds for every $\varepsilon \in (0,1)$ and is $1.00174488\ldots$ at $\varepsilon = 0.01$. Every attempt uses two fresh solutions, so the same ratio holds for generated-solution cost. This is an exact one-update separation. It makes no statement that this particular policy has infinite total activation time. Repeated-update statements require the transition assumptions of the activation theorems.

How successive updates induce dependence. The updated state also determines the law of the next update. This gives an exact connection between temporal dependence and the optimizer without identifying a neural training run with a two-state sign chain. To isolate it, first assume that inactive attempts preserve the state, and let Y_j be the state after feedback event j. Its transition kernel is K_1. Write $D_j = \theta(Y_j) - \theta(Y_{j-1})$, and define the positive operator

$$(\mathcal{M}_r f)(z) = \int e^{-r\alpha[\theta(z')-\theta(z)]} f(z') K_1(z, dz'), \qquad m_r(z) = \mathcal{M}_r \mathbf{1}(z).$$

Successive conditioning gives

$$\mathbb{E}_z e^{-r\alpha\sum_{j=1}^n D_j} = \mathcal{M}_r^n \mathbf{1}(z), \qquad \mathbb{E}_z e^{-r\alpha(D_1+D_2)} = \mathbb{E}_z[e^{-r\alpha D_1} m_r(Y_1)]. \tag{22}$$

Indeed, the claim for one transition is the operator definition. Conditioning on Y_1 proves the two-transition identity, and induction proves the general case. These are identities of nonnegative expectations, so they also hold when a moment is infinite.

When the moments are finite, the second expression equals

$$m_r(z)\mathbb{E}_z m_r(Y_1) + \mathrm{Cov}_z\left(e^{-r\alpha D_1}, m_r(Y_1)\right).$$

The covariance makes the mechanism concrete. A harmful update has a large first factor. If it also moves the optimizer and policy into states with a larger subsequent cost multiplier, the covariance is positive and the two-update cost factor increases. Thus shared representations and optimizer memory enter through the conditional law of the updated state. In the translation-invariant sign model of Theorem 4, restricting this operator to functions of the current sign gives the matrix M_r. Its parameter ρ summarizes those conditional transitions only when that sign model applies. Neither an optimizer coefficient nor within-group rollout correlation identifies it by itself.

For a neural run, one can instead estimate successive conditional moments from restored-state continuations, including the optimizer and sampler state. Appendix G.8 describes finite-state upper comparisons when a justified state partition is available. If inactive steps also move the policy, their transitions must be included before the next feedback event. In that case, the full-state and block bounds retain them directly, avoiding an incorrect geometric-wait approximation. The operator identity here is the standard conditional-expectation calculation that connects those measurements to the model's temporal dependence.

Conditional score distributions for centered update rules. Let ϕ denote the policy parameters, $s(\tau) = \nabla_\phi \log P_\phi(\tau)$, and $\theta(\phi) = \mathrm{logit}\, P_\phi(A)$. Assume differentiability and integrability sufficient to differentiate the success and failure probabilities under the integral. At the current parameters,

$$\mu_+ = \mathbb{E}[s \mid A] = \nabla \log p, \qquad \mu_- = \mathbb{E}[s \mid A^c] = \nabla \log(1-p), \qquad \mu_+ - \mu_- = \nabla\theta.$$

For K independent on-policy solutions, let N be the number passing the verifier, and consider

$$G = \frac{1}{K}\sum_{i=1}^{K} c_N(R_i - N/K)s(\tau_i), \qquad U = \eta M G,$$

where $\eta \geq 0$, $c_n > 0$ depends only on the reward count and M is a symmetric positive semidefinite matrix fixed before the group is sampled. Define $G = 0$ for $N = 0, K$. Group centering uses $c_n = 1$, leave-one-out uses $c_n = K/(K-1)$, and division by the empirical reward standard deviation uses $c_n = K/\sqrt{n(K-n)}$ on mixed groups. A positive numerical stabilizer simply changes c_n. This calculation concerns a fresh, single-step score-gradient update. A preconditioner computed from the current group, as in Adam's current moments, must instead remain inside the exact optimizer law.

Conditional on $N = n \in \{1, \ldots, K-1\}$, the correct and incorrect solutions are independent draws from their respective conditional laws after random permutation. Let $\bar{s}_+$ and $\bar{s}_-$ be their sample means. Then

$$G = c_n \frac{n(K-n)}{K^2}(\bar{s}_+ - \bar{s}_-), \qquad \mathbb{E}[G \mid N = n] = c_n \frac{n(K-n)}{K^2}\nabla\theta. \tag{23}$$

Thus the conditional mean of the linearized log-odds increment is nonnegative for this entire class. For its exponential moment, put

$$v = M\nabla\theta, \quad V_\pm = v^\top s(\tau_\pm), \quad \mathcal{M}_\pm(t) = \mathbb{E}e^{tV_\pm}, \quad a_n = \eta c_n \frac{n(K-n)}{K^2}.$$

Independence within each conditional class gives the exact formula for $D_{\text{lin}} = \nabla\theta^\top U$:

$$\mathbb{E}[e^{-\alpha D_{\text{lin}}} \mid N = n] = \mathcal{M}_+(-\alpha a_n/n)^n \mathcal{M}_-(\alpha a_n/(K-n))^{K-n}. \tag{24}$$

The identity is valid as a nonnegative extended-valued expectation, and is finite when the indicated moments are finite. No bounded-score assumption is needed. Averaging with weights $\binom{K}{n}p^n(1-p)^{K-n}/\lambda_K(p)$ gives the conditional mixed-group moment. Equation (17) explains when the $n = 1$ conditional law dominates. Total variation controls bounded observables only, so approximating an unbounded exponential moment additionally requires a tail bound or uniform integrability.

For the finite-step error, assume $\|\nabla^2\theta\|_{\text{op}} \leq L_\theta$ along every segment between ϕ and $\phi + U$ under consideration. Taylor's integral formula gives

$$|D - D_{\text{lin}}| \leq \tfrac{1}{2}L_\theta\|U\|^2,$$

and hence, conditional on any specified reward count or mixed-group event,

$$\mathbb{E}e^{-\alpha D_{\text{lin}} - \alpha L_\theta\|U\|^2/2} \leq \mathbb{E}e^{-\alpha D} \leq \mathbb{E}e^{-\alpha D_{\text{lin}} + \alpha L_\theta\|U\|^2/2}. \tag{25}$$

The inequalities remain valid with an infinite upper bound. A finite certificate requires the displayed exponential integrability. In the three-solution example the exact likelihood-ratio calculation retains curvature and establishes a cost increase even though the linearized exponential moment is below one. This is why the main criterion uses actual post-optimizer progress.

A verifier error floor separates activity from correctness signal. Let $Y = \mathbf{1}_A$ be true correctness and R the observed binary verifier output. Assume independent, parameter-independent verifier errors across trajectories, with $\mathbb{P}(R = 1 \mid Y = 0) = a$ and $\mathbb{P}(R = 1 \mid Y = 1) = b$, where $0 < a < b < 1$. Conditional on Y, the verifier randomization is independent of the trajectory. Write $p = \mathbb{P}(Y = 1)$, $\theta = \operatorname{logit} p$, and $I = \{0 < \sum_i R_i < K\}$ for mixed observed labels. Then $\widetilde{p} = \mathbb{E}R = a + (b-a)p$. For the unnormalized centered gradient $G = K^{-1}\sum_i(R_i - \bar{R})s(\tau_i)$,

$$\mathbb{P}(I) = \lambda_K(\widetilde{p}), \qquad \mathbb{E}G = \frac{K-1}{K}(b-a)\nabla p, \qquad \mathbb{E}[G \mid I] = \frac{(K-1)(b-a)p(1-p)}{K\lambda_K(\widetilde{p})}\nabla\theta. \tag{26}$$

To prove the gradient formula, use $\mathbb{E}s = 0$ and independence of different trajectories to obtain $\mathbb{E}[K^{-1}\sum_i \bar{R}s_i] = K^{-1}\mathbb{E}[Rs]$. Conditional independence of the verifier gives $\mathbb{E}[Rs] = (b-a)\mathbb{E}[Ys] = (b-a)\nabla p$. Since G vanishes off I, division by $\mathbb{P}(I)$ gives the last identity.

As $p \downarrow 0$, activity approaches $\lambda_K(a) > 0$, while the coefficient of $\nabla\theta$ in the conditional mean is $O(p)$. If $\nabla\theta$ stays bounded, the entire conditional mean vanishes. For the exact verifier $a = 0, b = 1$,

the corresponding coefficient tends to $(K-1)/K^2$. Thus a persistent mixed-score rate can arise from verifier errors without a persistent correctness-directed update. The result concerns the stated class-conditional error model. Systematic trajectory-dependent errors must be retained in the joint verifier and trajectory law. Observed-score calibration alone then describes observed feedback, and an independent correctness audit is needed to identify its relation to true success.

C The Activation Law and its Error Bound

Appendix B identified the arrival rate and conditional update. We now use them to approximate the complete time to a target. We first treat positive increments, for which a progress-based coupling gives an explicit error bound. The signed extension in Appendix F.3 uses the same waiting-time representation but must additionally control returns to less successful states.

A local model for sharp activation laws. Write $p = \mathrm{sigmoid}(\theta)$ and initialize $p_0 = \varepsilon$. Fix $p_\star \in (0,1)$. For Theorem 9, inactive steps preserve θ, and active steps add a random increment with law ν_p. Assume constants $c, c_-, C, d_-, d_+ > 0$ satisfy, for $0 < p \le p_\star$,

$$|\lambda(p) - cp| \le Cp^2, \quad \lambda(p) \ge c_- p, \quad d_{\mathrm{TV}}(\nu_p, \nu) \le Cp, \quad \mathrm{supp}(\nu_p) \cup \mathrm{supp}(\nu) \subset [d_-, d_+]. \tag{27}$$

The linear rare-feedback rate and a limiting conditional increment are local regularity conditions. The positive lower increment is a substantive restriction, verified for the scalar finite-group example below. Background motion and signed updates remain in (1), and the regularized analysis treats a separate case with background motion. This separation avoids imposing favorable scalar dynamics on a general neural optimizer in the full-state description.

Let $D_1, D_2, \ldots$ be independent draws from ν, let $S_0 = 0$, $S_j = \sum_{i=1}^{j} D_i$, and let E_j be independent unit-rate exponential variables, independent of the increments. Define

$$W_\nu = \sum_{j=0}^{\infty} e^{-S_j} E_j, \qquad m_r = \mathbb{E} e^{-rD_1}. \tag{28}$$

Each term is the waiting cost after j feedback updates. A larger increment shortens subsequent waiting by increasing the chance of the next informative group. This infinite sum is a classical exponential functional of an embedded random walk. The theorem identifies when a finite RLVR process approximates that object and controls the error at a nonzero initialization.

Theorem 9 (Activation law with finite-initialization control)**.** *Under* (27)*, let* $T_\star = \inf\{n : p_n \ge p_\star\}$*. As* $\varepsilon \downarrow 0$*,*

$$c\varepsilon T_\star \Rightarrow W_\nu, \qquad d_{\mathrm{BL}}\big(\mathcal{L}(c\varepsilon T_\star), \mathcal{L}(W_\nu)\big) \le C_\star\big(\sqrt{\varepsilon} + \varepsilon \log(1/\varepsilon)\big). \tag{29}$$

The distance tests functions bounded by one with Lipschitz constant at most one. The constant depends on (27) *and* $p_\star$*. Every fixed positive integer moment converges. In particular,*

$$\mathbb{E} W_\nu = \frac{1}{1-m_1}, \qquad \mathrm{CV}^2(W_\nu) = \frac{1-2m_1+m_2}{1-m_2}. \tag{30}$$

Proof Sketch. The proof separates an early region, where the two processes can be closely coupled, from a later region, where little normalized waiting cost remains. We couple increments until cumulative logit progress reaches half of $\log(1/\varepsilon)$. Before this cutoff, mismatch probability is $O(\sqrt{\varepsilon})$. Afterwards, the two cost tails have expectation $O(\sqrt{\varepsilon}+\varepsilon\log(1/\varepsilon))$. Choosing the cutoff by progress is what yields the stated error. The remaining moment identities follow from the classical perpetuity recursion.

A finite-budget prediction. The distributional bound also controls the probability of activation before a specified budget. The limiting variable is the sum of a unit exponential and an independent nonnegative remainder, so its density is bounded by one. Smoothing an indicator at width h converts a bounded-Lipschitz error r_ε into an error $r_\varepsilon/h + h$ for a distribution function. With $h = \sqrt{r_\varepsilon}$, Theorem 9 therefore implies

$$\sup_{b \ge 0} |\mathbb{P}(c\varepsilon T_\star \le b) - \mathbb{P}(W_\nu \le b)| \le 2\sqrt{r_\varepsilon}, \quad r_\varepsilon = C_\star\big(\sqrt{\varepsilon} + \varepsilon\log(1/\varepsilon)\big), \tag{31}$$

whenever $r_\varepsilon \le 1$. For a constant active increment, the Laplace transform is $\mathbb{E} e^{-sW_\delta} = \prod_{j\ge 0}(1 + se^{-j\delta})^{-1}$. These formulas predict a complete budget-success curve from the transition law. They require no fit of a separate activation-time parameter to the observed runs.

Which feedback statistics suffice? Even a known feedback rate and mean active increment leave the expected activation time undetermined. Let an active increment D lie in $[a, b]$ with mean d, where $0 < a < d < b$. Jensen's inequality and the chord bound for e^{-D} give the sharp interval

$$\frac{1}{1-e^{-d}} \le \mathbb{E}W_\nu \le \left[1 - \frac{(b-d)e^{-a} + (d-a)e^{-b}}{b-a}\right]^{-1}. \tag{32}$$

The lower endpoint is attained by a constant increment d, and the upper endpoint by increments at a and b with mean d. Thus the conditional Laplace moment $\mathbb{E}e^{-D}$, beyond $\mathbb{E}D$, controls the activation clock. The bound follows from classical convexity applied to the activation identity, and its operational implication is exact for the specified limiting activation law.

To see the resulting nonidentification directly, fix the same rate function $\lambda(p)$ for two processes and choose their active increments independently of p, with the same mean d. Both have expected update $\lambda(p)d$ at every policy. Their limiting mean times can nevertheless differ by an arbitrarily large factor: take one increment constantly d, and the other at $a = d/b$ and b, with its mixture weights preserving the mean, then send $b \to \infty$ after taking the rare-success limit for each fixed b. The upper endpoint of (32) diverges. This construction concerns the general transition class. The batching results realize an exact matched-mean separation using ordinary finite-group updates and an intervention solely on group dependence.

C.1 Geometric waiting and a finite event cutoff

Let Y_j denote the logit immediately after j active updates, and put $Y_0 = \mathrm{logit}(\varepsilon)$. Let M be the number of active updates required to reach $\theta_\star = \mathrm{logit}(p_\star)$. Since every active increment lies in $[d_-, d_+]$,

$$M \le N_\varepsilon := 1 + \left\lceil \frac{\theta_\star - Y_0}{d_-} \right\rceil, \qquad Y_0 + jd_- \le Y_j \le Y_0 + jd_+ \quad (j \le M). \tag{33}$$

All claims below are for sufficiently small $\varepsilon < p_\star$, so this deterministic bound is positive. Define G_j to be the wait at Y_j for $j < M$, and zero for $j \ge M$. Proposition 1 gives $T_\star = \sum_{j=0}^{N_\varepsilon - 1} G_j$. For a geometric variable of rate $q \in (0,1)$, an exact coupling with $E \sim \mathrm{Exp}(1)$ is

$$G = \lceil E/[-\log(1-q)] \rceil. \tag{34}$$

The case $q = 1$ has $G = 1$ and can be treated separately in the tail estimates. On the early rare-success segment, $q < 1/2$ for small enough ε.

C.2 Coupling up to a progress threshold

Set $h_\varepsilon = \frac{1}{2}\log(1/\varepsilon)$ and $J = \inf\{j : S_j \ge h_\varepsilon\}$ for the independent limiting increments. Since $D_j \in [d_-, d_+]$, we have $J \le 1 + h_\varepsilon/d_-$ and $S_J \in [h_\varepsilon, h_\varepsilon + d_+]$. For small initialization, this entire segment lies below the fixed target. Maximally couple each actual increment to its limiting draw until either disagreement or the common progress threshold. Let $\mathcal{F}_j$ be the joint increment and coupling history through event j, excluding exponential waiting variables. Define $\tau = \inf\{k \ge 1 : \text{actual increment } k \ne D_k\}$ and $A = \{\tau > J\}$. Thus $A \in \mathcal{F}_J$ includes agreement of the increment crossing the cutoff. Before disagreement,

$$p_j = \frac{\varepsilon e^{S_j}}{1 - \varepsilon + \varepsilon e^{S_j}} \le 2\varepsilon e^{S_j}.$$

Since successive S_j increase by at least d_-, summing backwards from the last state below h_ε gives the pathwise bound $\sum_{j<J} e^{S_j} \le e^{h_\varepsilon}/(1 - e^{-d_-})$. Conditional union bounds and the local total-variation estimate therefore imply

$$\mathbb{P}(\text{a disagreement before the progress cutoff}) \le C\varepsilon e^{h_\varepsilon} = C\sqrt{\varepsilon}. \tag{35}$$

The coupled limit increments retain their independent law because each coupling uses a fresh limiting draw conditionally on the past.

Use independent unit exponentials for the geometric waits. Put $a = c\varepsilon$ and $q_j = \lambda(p_j)$. Uniformly before the cutoff,

$$\frac{a}{-\log(1-q_j)} = \frac{\varepsilon}{p_j}(1 + O(p_j)) = e^{-S_j} + O(\varepsilon).$$

Here $\varepsilon/p_j = (1-\varepsilon)e^{-S_j} + \varepsilon$ and positive progress gives $e^{-S_j} \le 1$. The rounding error in the inverse-transform geometric coupling consequently satisfies

$$\mathbf{1}_{\{j<J,\tau>j\}}\,\mathbb{E}\left[\left|aG_j - e^{-S_j}E_j\right| \mid \mathcal{F}_j\right] \le C'\varepsilon\,\mathbf{1}_{\{j<J,\tau>j\}}. \tag{36}$$

There are at most $1 + h_\varepsilon/d_-$ such terms. Multiplying their sum by $\mathbf{1}_A$ and taking expectations gives $O(\varepsilon\log(1/\varepsilon))$. The cutoff depends only on increments, independently of the exponential draws. Each future limiting increment retains law ν conditionally on $\mathcal{F}_J$.

C.3 Remainders and convergence of moments

On the event of successful coupling up to J, both chains reach cumulative progress at least h_ε. For the actual chain, every later increment is at least d_- and at most $N_\varepsilon = O(\log(1/\varepsilon))$ events occur before the target. On A, conditioning on $\mathcal{F}_J$, the reciprocal-rate bound gives

$$\mathbb{E}\left[a\sum_{J\le j<N_\varepsilon} G_j \,\middle|\, \mathcal{F}_J\right] \le C\left[\varepsilon N_\varepsilon + \frac{e^{-h_\varepsilon}}{1-e^{-d_-}}\right]. \tag{37}$$

This follows termwise from $a\,\mathbb{E}G_{J+k} \le C(\varepsilon + e^{-h_\varepsilon - kd_-})$ while the state is live. The limiting remainder has conditional mean $e^{-S_J}/(1-m_1) \le e^{-h_\varepsilon}/(1-m_1)$. A bounded one-Lipschitz test function incurs at most twice the disagreement probability on the failed-coupling event. On its complement, use the early waiting error and these two remainder bounds. Since $e^{-h_\varepsilon} = \sqrt{\varepsilon}$, this proves (29). The improvement uses positive increments and leaves the signed-state-dependent argument separate.

To obtain moments, for each integer $r \ge 1$ a geometric variable of rate q satisfies $\mathbb{E}G^r \le C_r q^{-r}$. One proof uses $\mathbb{P}(G > m) \le e^{-qm}$ and sums the tail representation of the moment. Conditional on Y_j and on $j < M$, this bound and (33) imply

$$\|aG_j\|_{L^r} \le C_r'(\varepsilon + e^{-jd_-}).$$

Minkowski's inequality, with $G_j = 0$ after stopping, gives

$$\|aT_\star\|_{L^r} \le C_r' \sum_{j<N_\varepsilon} (\varepsilon + e^{-jd_-}) \le C_r''.$$

For convergence of the rth moment, apply this bound with exponent $r+1$ to obtain uniform integrability, then use convergence in distribution. The same geometric-series argument gives every moment of W_ν.

C.4 Moment calculation

Splitting off the first term in (28) gives the distributional identity

$$W_\nu \overset{d}{=} E_0 + e^{-D_1}W_\nu',$$

where all three quantities on the right are independent and W_ν' has the law of W_ν. With $m_r = \mathbb{E}e^{-rD_1} < 1$, taking the first and second moments gives

$$\mathbb{E}W_\nu = 1 + m_1\mathbb{E}W_\nu, \qquad \mathbb{E}W_\nu^2 = 2 + 2m_1\mathbb{E}W_\nu + m_2\mathbb{E}W_\nu^2.$$

Thus $\mathbb{E}W_\nu = (1-m_1)^{-1}$ and $\mathbb{E}W_\nu^2 = 2/[(1-m_1)(1-m_2)]$. Subtracting the squared mean and dividing by it yields (30). This calculation completes the proof of Theorem 9.

C.5 Budget probabilities and sharp moment bounds

Since $W_\nu = E_0 + R$ with R independent and nonnegative, its density is $f_W(t) = \mathbb{E}[e^{-(t-R)}\mathbf{1}_{R\le t}] \le 1$. For any b, upper and lower continuous approximations to $\mathbf{1}_{(-\infty,b]}$ can be chosen with transition width h, values in $[0,1]$, and Lipschitz constant $1/h$. If two laws have bounded-Lipschitz distance r and $0 < h \le 1$, their expectations of either approximation differ by at most r/h. The density bound makes the difference between either approximation and the limiting distribution function at most h. Choosing $h = \sqrt{r}$ proves (31). Independence of the exponentials gives the finite-product Laplace transform, and dominated convergence gives the infinite product stated in the text.

For (32), convexity gives $e^{-d} \le \mathbb{E}e^{-D}$. The chord inequality on $[a, b]$ gives $e^{-D} \le [(b - D)e^{-a} + (D - a)e^{-b}]/(b - a)$. Take expectations and use the increasing map $m \mapsto (1 - m)^{-1}$. A constant $D = d$ and the two-endpoint distribution attain the respective bounds. For $a = d/b$, the endpoint probability at b is $(d - a)/(b - a) \sim d/b$, so $\mathbb{E}e^{-D} \to 1$ and the upper bound diverges. Each fixed endpoint pair obeys the activation theorem's assumptions. This argument takes the initialization limit before varying the endpoints, with no uniform claim about its approximation constant.

D Finite Groups and Matched-Mean Batches

The coupling theorem becomes directly testable once the arrival and increment laws are obtained from a training rule. We derive them here for finite response groups. This calculation also distinguishes two approximations that can otherwise obscure the effect: replacing empirical normalization by population normalization, and replacing the resulting random update by its mean.

Corollary 10 (Finite-group activation moments)**.** *For (7), the limiting law and mean-recursion limit are (8). Its squared coefficient of variation is* $\tanh(\delta/2)$*, and* $\delta W_\delta \to 1$ *in mean square as* $\delta \downarrow 0$*.*

D.1 Population normalization and exact conditional means

There are two common ways to average a finite-group training process. One replaces the empirical reward mean and standard deviation inside each group by their population values. The other retains the exact finite-group expectation and replaces its stochastic updates by a deterministic recursion. These operations address different randomness. The first changes the expected vector field because normalization is nonlinear. The second preserves that field and removes the waiting-time and conditional-update fluctuations. Corollary 10 isolates the second operation by computing the finite-group mean exactly at every intermediate success probability.

For the scalar normalized update, its exact mean is

$$F_K(p) = \eta \sum_{n=1}^{K-1} \binom{K}{n} p^n (1-p)^{K-n} \sqrt{\frac{n}{K}\left(1 - \frac{n}{K}\right)}. \tag{38}$$

At fixed K and small p, $F_K(p) = \eta\sqrt{K-1}\,p + O(p^2)$. At fixed $p \in (0, 1)$ and large K, $F_K(p) \to \eta\sqrt{p(1-p)}$. The former records singleton successes in mostly inactive groups. The latter corresponds to concentration of the empirical success fraction. Consequently, using a population-normalized objective in the rare-success region can already change the speed of its deterministic prediction before any stochastic activation effect is considered during learning.

The crossover is governed by the expected number of successes per group. Put $p = \rho/K$ and let $K \to \infty$, holding $\rho > 0$ fixed. The binomial count converges to $N_\rho \sim \mathrm{Poisson}(\rho)$, giving

$$\frac{\sqrt{K}}{\eta} F_K(\rho/K) \longrightarrow \mathbb{E}\sqrt{N_\rho}, \qquad \lambda_K(\rho/K) \longrightarrow 1 - e^{-\rho}. \tag{39}$$

Moreover, $\mathbb{E}\sqrt{N_\rho} = \rho + O(\rho^2)$ as $\rho \downarrow 0$, and $\mathbb{E}\sqrt{N_\rho}/\sqrt{\rho} \to 1$ as $\rho \to \infty$. The same expression therefore connects the rare singleton regime to the population-normalized regime. Appendix D proves the Poisson-limit calculation and states its precise scope in the analysis of finite-group training.

These regimes suggest two separate questions when interpreting a training trace. Does each group contain enough successes for its empirical normalization to concentrate? Are the active parameter increments small enough for the activation clock to concentrate? Increasing the number of independently averaged groups can address the second question while leaving the first unchanged, because each group still uses its own normalization. Repeating the same group addresses neither source of randomness, even though the nominal batch count increases. The following construction holds the exact finite-group mean field fixed and makes the resulting activation difference explicit.

D.2 The duplicated-group dependence intervention

At a fixed policy, form a batch of B groups of size K and average their updates. In the independent construction, all groups are independent. In the correlated construction, one sampled

group is repeated B times. Every group has the same marginal law in both constructions, and their expected averaged updates agree exactly for every p. This is a controlled dependence intervention. Repeated groups contain fewer unique samples, so the comparison isolates dependence and makes no claim of equal information per generated trajectory.

Theorem 11 (Matched-mean separation)**.** *For each fixed B, the two constructions have the same expected update at every policy. Their activation times satisfy*

$$K\varepsilon T_\star^{\mathrm{corr}} \Rightarrow W_\delta, \qquad K\varepsilon T_\star^{\mathrm{ind}} \Rightarrow B^{-1} W_{\delta/B}. \tag{40}$$

The independent limit has mean $[B(1-e^{-\delta/B})]^{-1}$ and squared coefficient of variation $\tanh(\delta/(2B))$. Taking $\varepsilon \downarrow 0$ first and then $B \to \infty$, the independent limit converges in mean square to $1/\delta$. The correlated limit remains W_δ for every B and preserves its nonzero relative dispersion.

The independent construction receives approximately B times as many active batches in the rare region. Each typically contains one informative group and changes the logit by approximately δ/B. These changes cancel in the mean update and remain separately in the activation law. The full mean vector field therefore fails to identify the distribution of training progress. The theorem gives an exact counterexample to identification, together with a deterministic-recovery limit that serves as a negative control for the mechanism at small increment sizes.

D.3 Deriving the normalized update

For $a_i \in \{0,1\}$ and $\widehat{p} = N/K$, define $s = \sqrt{\widehat{p}(1-\widehat{p})}$ and $A_i = (a_i - \widehat{p})/s$ when $0 < N < K$. The gradient of the Bernoulli log-likelihood is $a_i - p$. Centering gives $\sum_{i=1}^K A_i = 0$, eliminating the baseline term p. Holding advantages fixed during the on-policy gradient evaluation,

$$\frac{\eta}{K}\sum_{i=1}^K A_i(a_i - p) = \frac{\eta}{Ks}\sum_{i=1}^K (a_i - \widehat{p})a_i = \frac{\eta}{s}\widehat{p}(1-\widehat{p}) = \eta s.$$

For $N = 0$ or $N = K$, the specified update is zero. This proves (7). The population standard deviation convention is explicit. An unbiased sample standard deviation multiplies every nonzero jump by $\sqrt{(K-1)/K}$. Adding a positive numerical stabilizer changes the jump sizes as well, and the appropriate singleton jump must then be used.

Proof of Corollary 10. For fixed K, $\lambda(p) = Kp + O(p^2)$ as $p \downarrow 0$. Its ratio to p is continuous and positive on $(0, p_\star]$ and has limit K at zero, so it has a positive lower bound. Every mixed-group jump lies in $[\delta, \eta/2]$. Conditional on a mixed group, the probability of $N = 1$ is $1 - O(p)$, and its jump is δ. For $K = 2$, every mixed group already has $N = 1$. Hence (27) holds with $c = K$ and $\nu = \mathrm{Dirac}(\delta)$. Theorem 9 gives the limiting series and moments.

For the exact-mean recursion, let $F(p) = \mathbb{E}[g_K(N) \mid p]$. The singleton-success term gives $F(p) = K\delta p + O(p^2)$. Write $q = e^{-\theta} = (1-p)/p$. A deterministic step satisfies

$$q - q_{\mathrm{next}} = q(1 - e^{-F(p)}) = K\delta + O(p).$$

Fix a large M. While $q \geq M$, the decrement lies between $K\delta - C/M$ and $K\delta + C/M$. The time to reach M is therefore bracketed by the initial reciprocal odds, up to fixed overshoot constants, divided by these two bounds. Since $\varepsilon q_0 \to 1$, multiplying time by $K\varepsilon$, sending $\varepsilon \downarrow 0$, and then sending $M \to \infty$ gives $1/\delta$. From the fixed level M to the target, $F(p)$ has a positive minimum on the resulting compact success interval, so the remaining deterministic time is bounded independently of ε.

Finally, $\mathbb{E}[\delta W_\delta] = \delta/(1-e^{-\delta}) \to 1$ and $\mathrm{Var}(\delta W_\delta) = \delta^2/(1-e^{-2\delta}) \to 0$. These establish the stated small-jump mean-square limit. □

The crossover calculation. Equation (38) follows by averaging the exact binomial update. Its small-p expansion retains the $N = 1$ term, and the fixed-p, large-K limit follows by bounded convergence from $N/K \to p$ in probability. For (39), $N_K \sim \mathrm{Bin}(K, \rho/K)$ converges in distribution to N_ρ. The normalized update is $\sqrt{N_K(1 - N_K/K)}$. It converges in distribution to $\sqrt{N_\rho}$, and

its square has expectation at most $\mathbb{E}N_K = \rho$, giving uniform integrability and convergence of expectations. The event-rate limit follows directly from $1-(\rho/K)^K-(1-\rho/K)^K$.

For small ρ, the $N_\rho = 1$ contribution is $\rho e^{-\rho}$. The contribution from $N_\rho \geq 2$ is bounded above by $\mathbb{E}[N_\rho \mathbf{1}_{N_\rho \geq 2}] = \rho(1-e^{-\rho}) = O(\rho^2)$. This proves the first asymptotic. For large ρ, $N_\rho/\rho \to 1$ in probability and $\mathbb{E}[N_\rho/\rho] = 1$. Uniform integrability of its square root gives $\mathbb{E}\sqrt{N_\rho/\rho} \to 1$. No joint activation limit with K diverging is inferred from this one-step calculation.

Proof of Theorem 11. Let $U_b = g_K(N_b)$. Linearity gives $\mathbb{E}[B^{-1}\sum_b U_b] = \mathbb{E}U_1$ for either coupling, with no approximation. Under perfect correlation, $B^{-1}\sum_b U_b = U_1$ pathwise, so Corollary 10 applies directly.

Under independence, a batch is active with probability

$$\lambda_B(p) = 1-(1-\lambda(p))^B = BKp + O(p^2).$$

Conditional on activity, exactly one group has exactly one success with probability $1-O(p)$, for fixed B, K. All other groups have zero successes with that same limiting conditional probability. The averaged jump then equals δ/B. Every nonzero batch update lies in $[\delta/B, \eta/2]$, and $\lambda_B(p)/p$ has a positive lower bound up to $p_\star$. Theorem 9 gives $BK\varepsilon T_\star^{\mathrm{ind}} \Rightarrow W_{\delta/B}$. Dividing by B proves (40).

The first two moments follow from Corollary 10. In particular,

$$\mathbb{E}[B^{-1}W_{\delta/B}] = \frac{1}{B(1-e^{-\delta/B})} \to \delta^{-1}, \qquad \mathrm{Var}(B^{-1}W_{\delta/B}) = \frac{1}{B^2(1-e^{-2\delta/B})} \to 0.$$

This proves mean-square convergence. The proof takes the rare-success limit for each fixed batch size before increasing B. A joint limit with B depending on ε requires a separate uniform argument. □

D.4 Equal-data weighting and its exact activation law

For any realized fresh pool, regard the computed group updates as fixed numbers $G_1, \ldots, G_B$. A uniform random permutation of a weight vector with sum B satisfies $\mathbb{E}_\sigma w_{\sigma(j)} = 1$ for each j. Hence

$$\mathbb{E}_\sigma[\Delta\theta_{\boldsymbol{w}} \mid G_1, \ldots, G_B] = B^{-1}\sum_{j=1}^{B} G_j.$$

Taking expectation over the pool proves equality of the conditional mean at every policy. The argument also holds for vector-valued gradients before a nonlinear optimizer transformation. In the theorem the optimizer is the specified linear scalar update, so no additional transformation intervenes.

For fixed B, K, a batch is informative precisely when at least one group is mixed, since all weights are positive and all mixed-group scalar updates are positive. Its rate is

$$\Lambda_B(p) = 1-(1-\lambda_K(p))^B = BKp + O(p^2).$$

Conditional on an informative batch, with probability $1-O(p)$ there is exactly one successful trajectory in the entire batch. Its group index is uniform by exchangeability, its group update is δ, and its permuted weight is uniform among $w_1, \ldots, w_B$. Thus the limiting active increment is $D_{\boldsymbol{w}} = \delta w_J/B$.

The finite-p active-increment law differs from this law in total variation by $O(p)$, because the probability of any exceptional multi-success configuration conditional on an informative batch is $O(p)$. Every active increment is at least $\delta \min_j w_j/B > 0$ and at most $\eta/2$, because the weights sum to B and each group update is at most $\eta/2$. The hypotheses of Theorem 9 hold with rate constant BK. Its distributional and mean convergence give (16). The limiting squared coefficient of variation is

$$\frac{1-2m_1+m_2}{1-m_2}, \qquad m_r = B^{-1}\sum_{j=1}^{B} e^{-r\delta w_j/B}.$$

This calculation uses a fixed number of fresh trajectories per attempted update in both conditions. Multiplying time by BK converts to generated-trajectory cost without changing their relative comparison.

Strict convexity gives $B^{-1}\sum_{j=1}^{B} e^{-\delta w_j/B} \geq e^{-\delta/B}$, with equality exactly when all weights equal one. Since $x \mapsto (1-x)^{-1}$ is strictly increasing on $(0,1)$, uniform weights uniquely minimize

Table 2: Curvature-matched penalties. Factors 4 and 2 normalize Jensen–Shannon and squared Hellinger to $f''(1) = 1$, matching both Kullback–Leibler penalties. All forces have derivative one at the reference.

Penalty	Restoring force $\psi_f(x)$
Reverse Kullback–Leibler	$1 - e^{-x}$
$4\times$ Jensen–Shannon	$2\log(2/(1 + e^{-x}))$
$2\times$ squared Hellinger	$2(1 - e^{-x/2})$
Forward Kullback–Leibler	x
$\frac{1}{2} \times \chi^2$	$e^x - 1$

the limiting mean. For $B = 4$ and $\delta = 1$, the two weight vectors in the text give analytical means 1.1302029160 and 1.3549887610. Their ratio is approximately 1.199. These are consequences of the theorem, distinct from empirical simulation estimates.

The comparison equalizes trajectory generation, access to the sampled trajectories, and the number of group gradients computed. The intervention deliberately changes how the same information is weighted in the update. Information retained in the resulting parameter step can vary, as weighting is the intervention being studied. All trajectories are fresh, and the exact matched-mean identity holds conditional on each realized pool, a stronger statement than equality after averaging over pools.

E Regularized Limits and Escape

The finite-group model holds its policy fixed on homogeneous groups. A penalty changes this part of the process because it continues to act while reward contrast is absent. We therefore first derive the resulting jump-and-drift limit, then compare finite targets, and only afterwards consider escape through unbounded relative-odds levels.

E.1 Regularization changes finite-target progress

A divergence penalty acts throughout the waiting period. For a fixed reference ε, let f be convex and three-times continuously differentiable with $f(1) = f'(1) = 0$ and $f''(1) = 1$. Consider the exact discrete update

$$\begin{aligned} D_f(p\|\varepsilon) &= \varepsilon f(p/\varepsilon) + (1-\varepsilon) f((1-p)/(1-\varepsilon)), \\ \Delta\theta &= g_K(N) - \eta\gamma\, \partial_\theta D_f(p\|\varepsilon). \end{aligned} \tag{41}$$

The reference and Euclidean logit gradient are part of this specified model. A moving reference or adaptive optimizer changes the background kernel and requires its own scaling calculation.

Theorem 12 (The regularized rare-feedback limit)**.** *On bounded state intervals, localized at continuity levels of the exit functional, the process $X^\varepsilon(t) = \theta_{\lfloor t/(K\varepsilon)\rfloor} - \mathrm{logit}(\varepsilon)$ converges over finite times to a piecewise-deterministic Markov process with generator*

$$\mathcal{A}_f\phi(x) = e^x[\phi(x+\delta) - \phi(x) - \kappa\psi_f(x)\phi'(x)], \quad \kappa = \eta\gamma/K, \quad \psi_f(x) = f'(e^x). \tag{42}$$

In operational time u, it has rate-one Poisson jumps of size δ, drift $-\kappa\psi_f(X)$, and physical time $t(u) = \int_0^u e^{-X(v)}dv$, preserving the attempted-group budget along the evolving policy state.

Theorem 13 (Comparison at every finite relative-odds target)**.** *Couple the limiting processes using the same Poisson jumps and common δ, κ. Let T_H^f be physical time to relative log-odds $H > 0$. Then*

$$T_H^{\text{reverse KL}} \le T_H^{\text{JS}} \le T_H^{\text{Hellinger}} \le T_H^{\text{forward KL}} \le T_H^{\chi^2} \quad \textit{almost surely}. \tag{43}$$

Every such finite target has finite expected hitting time. More generally, for a nonnegative increasing, locally Lipschitz force on $[0,\infty)$ with $\psi(0) = 0$, choose $h > 0$ and an integer k such that $k\delta - \kappa\psi(H)h > H$. With $b_H = \mathbb{P}(\mathrm{Poisson}(h) \ge k) > 0$,

$$\mathbb{P}(T_H > jh) \le (1 - b_H)^j, \qquad \mathbb{E}T_H \le h/b_H.$$

The ordering follows from comparison of the restoring forces over the entire visited interval, followed by comparison of the physical clocks. It addresses each fixed finite relative log-odds threshold H in the continuous limiting process. Fixed accuracy $p_\star$ requires $H = \mathrm{logit}(p_\star) - \mathrm{logit}(\varepsilon)$, which grows as $\varepsilon \downarrow 0$. Discrete overshoots and higher-order gradient terms remain outside this pathwise ordering. The tail bound also shows that returning dynamics can reach every finite level. Appendix E keeps infinite-level escape as a separate ordered-limit result and supplies a finite-ε reachability condition for the exact discrete update. Thus a statement about escape through arbitrarily high relative odds cannot be substituted for a claim about finite-model accuracy.

E.2 The divergence gradient and its scaling

Differentiating (41) with respect to p, followed by $\partial_\theta p = p(1-p)$, gives the exact identity

$$\partial_\theta D_f(p\|\varepsilon) = p(1-p)\left[f'(p/\varepsilon) - f'((1-p)/(1-\varepsilon))\right]. \tag{44}$$

Set $\theta = \mathrm{logit}(\varepsilon) + x$. Uniformly for x in a fixed compact interval,

$$p = \varepsilon e^x + O(\varepsilon^2), \qquad p/\varepsilon = e^x + O(\varepsilon), \qquad (1-p)/(1-\varepsilon) = 1 + O(\varepsilon).$$

Smoothness and $f'(1) = 0$ therefore imply

$$\partial_\theta D_f(p\|\varepsilon) = \varepsilon e^x f'(e^x) + O(\varepsilon^2). \tag{45}$$

Constants in this expansion depend on the chosen interval. No uniform-in-x expansion is used. This distinction matters when subsequently discussing arbitrarily large relative odds.

Proof of Theorem 12. Write $a = K\varepsilon$ for one step of normalized physical time. The regularization increment $b_\varepsilon(x)$ obeys

$$b_\varepsilon(x) = -a\kappa e^x \psi_f(x) + O(a^2).$$

On a fixed compact interval, $\mathbb{P}(N = 1) = ae^x + O(a^2)$ and $\mathbb{P}(2 \le N \le K-1) = O(a^2)$. Homogeneous groups have no reward jump. Thus, for a twice continuously differentiable test function, Taylor expansion on a compact interval enlarged by the maximum jump gives

$$\frac{1}{a}\mathbb{E}[\phi(X_{n+1}) - \phi(X_n) \mid X_n = x] = -\kappa e^x \psi_f(x)\phi'(x) + e^x[\phi(x+\delta) - \phi(x)] + O(a).$$

The term from evaluating the regularization shift on an active step is $O(a^2)$ before division by a. Multi-success events contribute the same order. This proves local convergence of generators.

Here is a direct process argument for this setting. Stop at exit from a bounded interval, retaining the exit state, and enlarge the interval by the maximum possible jump. On that enlarged interval the limiting drift is Lipschitz and the limiting jump rate is bounded and Lipschitz. Construct candidate jump times from a Poisson process with a constant rate larger than all these rates. Accept a candidate at state x with probability e^x divided by the dominating rate. Between candidates solve the locally Lipschitz ordinary differential equation. This constructs a unique stopped limit with finitely many candidate jumps on each finite physical-time interval.

Partition physical time into mesh-a intervals. A dominating Poisson process has at least two candidates in a given interval with probability $O(a^2)$, so the probability of any such interval before a fixed time is $O(a)$. The discrete singleton-event probability differs from the corresponding accepted Poisson probability by $O(a^2)$ plus a times the difference in the coupled states. Bernoulli maximal coupling, summed over the mesh, bounds the chance of a jump mismatch by a constant times a plus the time integral of the current state error. The $O(a^2)$ drift remainder and the Euler approximation similarly bound the accumulated drift error by $O(a)$ plus that integral. Conditional on a bounded number of candidate jumps, Gronwall's inequality makes both errors tend to zero. The candidate-count tail is Poisson and can then be removed. Aligning discrete jump times with their continuous counterparts requires a time change of maximum size $O(a)$. This proves convergence in the Skorokhod topology up to the selected exit, with the usual restriction to continuity levels for an exit-time functional.

Finally, factor the limiting generator as $\mathcal{A}_f = e^x \mathcal{B}_f$, where

$$\mathcal{B}_f\phi(x) = \phi(x+\delta) - \phi(x) - \kappa\psi_f(x)\phi'(x).$$

The latter is the generator of drift $-\kappa\psi_f$ and constant-rate-one jumps. The time change $dt = e^{-X(u)}du$ gives generator $\mathcal{A}_f$ until any possible explosion in physical time. Starting at zero, the flow stays nonnegative and every jump is positive. Operational time has a well-defined process for all finite u, since there are finitely many Poisson jumps and the drift points toward zero on the positive half-line. □

Hitting-time continuity and the first jump. The localization in Theorem 12 retains its continuity-level qualification. Continuous Poisson jump times alone leave a height atom: the first jump lands at δ with probability one, because the process stays at zero before that jump. For the listed forces with $\kappa > 0$, later jump heights are atomless. Conditional on a positive post-jump state, the restoring flow is strictly decreasing with elapsed time and the next interarrival time has a density. Thus any fixed $H > \delta$ is crossed strictly by a jump almost surely, permitting the hitting-time mapping argument. For $0 < H \le \delta$, the specified discrete process reaches that relative level at its first mixed group. Directly, $T_H^\varepsilon \sim \mathrm{Geom}(\lambda_K(\varepsilon))$ and $K\varepsilon T_H^\varepsilon \Rightarrow \mathrm{Exp}(1)$. This handles the height atom by a separate exact calculation. With $\kappa = 0$, the process is a positive-jump renewal process and lattice levels require their direct event-count argument.

E.3 Curvature-matched divergence generators

The generators used in Table 2 are

$$f_{\mathrm{KL}}(r) = r\log r - r + 1, \qquad f_{\mathrm{RKL}}(r) = -\log r + r - 1,$$
$$f_{\mathrm{JS}}(r) = 2\left[r\log\frac{2r}{r+1} + \log\frac{2}{r+1}\right], \qquad f_{\mathrm{H}}(r) = 2(\sqrt{r}-1)^2,$$
$$f_{\chi^2}(r) = \tfrac{1}{2}(r-1)^2.$$

All have $f(1) = f'(1) = 0$ and $f''(1) = 1$. Substitution into $f'(e^x)$ gives the table. Because $d[f'(e^x)]/dx = f''(e^x)e^x$, their forces all have derivative one at zero. For Jensen–Shannon, $\psi_\infty - \psi(x) = 2\log(1 + e^{-x})$. For reverse Kullback–Leibler it equals e^{-x}, and for Hellinger it equals $2e^{-x/2}$. Each satisfies the critical-case decay condition in Theorem 14.

E.4 Finite-level comparison and exact discrete reachability

Ordering the curvature-matched forces. For $x \ge 0$, put $y = e^{-x/2} \in (0, 1]$. The forces satisfy

$$1 - e^{-x} \le 2\log\frac{2}{1+e^{-x}} \le 2(1 - e^{-x/2}) \le x \le e^x - 1.$$

For the first inequality, the difference as a function of $v = e^{-x}$ is $2\log(2/(1+v)) - (1-v)$. It vanishes at $v = 1$ and has derivative $(v-1)/(v+1) \le 0$, so it is nonnegative on $0 < v \le 1$. For the second inequality, half the difference of its right and left sides is $1 - y - \log(2/(1+y^2))$. This also vanishes at one, and its derivative is $-(1-y)^2/(1+y^2) \le 0$. The last two inequalities are the elementary exponential tangent bounds.

Pathwise comparison and the physical clock. All the forces vanish at zero, are nonnegative on $[0, \infty)$, and are locally Lipschitz. Starting at zero, their operational-time processes remain nonnegative. Couple every process to the same unit-rate Poisson process and the same positive jump size δ. Between jumps, the scalar differential-equation comparison theorem orders states inversely to their restoring forces. A common additive jump preserves that order. Induction over the locally finite jump times therefore gives

$$X_{\text{reverse KL}}(u) \ge X_{\mathrm{JS}}(u) \ge X_{\text{Hellinger}}(u) \ge X_{\text{forward KL}}(u) \ge X_{\chi^2}(u)$$

for every finite u. Suppose $X_a \ge X_b$, and let τ_a, τ_b be their operational first-hitting times of H. Then $\tau_a \le \tau_b$ and

$$T_H^a = \int_0^{\tau_a} e^{-X_a(u)}\,du \le \int_0^{\tau_a} e^{-X_b(u)}\,du \le \int_0^{\tau_b} e^{-X_b(u)}\,du = T_H^b.$$

This proves (43) on the event of finite operational hitting. The following argument proves that this event has probability one for every finite H and gives a finite expected physical time.

An explicit finite-level tail bound. Before hitting H, the downward speed in operational time is at most $\kappa\psi(H)$. From any state in $[0, H)$, k jumps in an interval of length h force a crossing whenever $k\delta - \kappa\psi(H)h > H$. Otherwise, at the kth jump the state would be at least this quantity, contradicting failure to cross. Such k, h exist for every finite H, since the force is finite there. Independent Poisson increments imply, by conditioning at successive block boundaries,

$$\mathbb{P}(\tau_H > jh) \le (1 - b_H)^j, \qquad b_H = \mathbb{P}(\mathrm{Poisson}(h) \ge k) > 0.$$

Nonnegativity of X gives $T_H \le \tau_H$, proving the tail bound in Theorem 13. Integration over intervals $[jh, (j+1)h)$ yields $\mathbb{E}T_H \le h \sum_{j\ge 0}(1 - b_H)^j = h/b_H$. This bound is conservative but completely finite-level. It includes forward Kullback–Leibler and the unbounded chi-square force.

A sufficient condition at fixed initialization and accuracy. The limiting comparison alone establishes no pathwise ordering for the exact discrete updates. Their binomial success counts and nonlinear group normalization must also be retained. The following elementary bound instead treats the finite process directly. Write $\theta_0 = \mathrm{logit}(\varepsilon)$, $\theta_\star = \mathrm{logit}(p_\star)$, and $r_f(\theta) = \eta\gamma\, \partial_\theta D_f(\mathrm{sigmoid}(\theta)\|\varepsilon)$. Suppose, on the fixed interval $[\theta_0, \theta_\star]$,

$$0 \le r_f(\theta) \le \theta - \theta_0, \qquad a = \delta - \sup_{\theta\in[\theta_0,\theta_\star]} r_f(\theta) > 0. \tag{46}$$

Let $M = \lceil(\theta_\star - \theta_0)/a\rceil$ and $\ell = \min_{\theta\in[\theta_0,\theta_\star]} \lambda_K(\mathrm{sigmoid}(\theta)) > 0$. The exact discrete process then satisfies

$$\mathbb{P}(T_\star > jM) \le (1 - \ell^M)^j, \qquad \mathbb{E}T_\star \le M/\ell^M. \tag{47}$$

Indeed, the first condition prevents any update from falling below the reference before activation. Each mixed group increases the logit by at least a, because $g_K(N) \ge \delta$ for $1 \le N \le K - 1$. Conditional on any pre-target state, M consecutive mixed groups thus force activation with probability at least ℓ^M. Conditioning at block boundaries proves the stated bound. Its parameters involve the actual finite initialization and target accuracy. This is a qualitative finite-mean certificate. The factor ℓ^{-M} can grow super-polynomially as initialization vanishes, and supplies no order-sharp rare-feedback bound.

For every fixed $\varepsilon < p_\star < 1$ and each smooth divergence considered here, conditions (46) hold when the positive penalty coefficient is sufficiently small. To see this, $r_f(\theta_0) = 0$, and the unscaled derivative is continuously differentiable on the compact interval. Its ratio to $\theta - \theta_0$ is therefore bounded, with a finite continuous extension at θ_0. Its maximum is also finite, so decreasing the coefficient enforces both inequalities. At a prescribed larger coefficient these conditions must be checked again. The result gives a sufficient reachability guarantee, with no assertion of uniform sharpness or reachability under every penalty strength.

Theorem 14 (Escape in the limiting process)**.** *Assume $\psi(0) = 0$, ψ is continuous and strictly increasing on $[0, \infty)$, and $\psi_\infty = \lim_{x\to\infty}\psi(x)$. Define $t_\infty = \int_0^\infty e^{-X(u)}du$ for the operational-time process in Theorem 12. If $\psi_\infty < \infty$ and $\delta > \kappa\psi_\infty$, then $X(u)/u \to \delta - \kappa\psi_\infty$ and $t_\infty < \infty$ almost surely. If $\delta < \kappa\psi_\infty$, including $\psi_\infty = \infty$, then $t_\infty = \infty$ almost surely. At $\delta = \kappa\psi_\infty$, the latter conclusion holds when $\psi_\infty - \psi(x) = o(1/x)$, as holds for all three bounded forces in Table 2.*

The theorem concerns escape through arbitrarily large *relative* log-odds levels in the limiting process. The order of limits is first $\varepsilon \downarrow 0$ on bounded levels, then the level tending to infinity. Reaching fixed accuracy at a finite ε is a separate hitting-time question. A recurrent process can eventually visit every finite level. Within the stated limit, forward Kullback–Leibler has an unbounded restoring force, whereas Jensen–Shannon has a finite threshold $\delta > 2\kappa\log 2$. This identifies a dynamical role for divergence choice beyond local curvature.

E.5 Proof of the escape classification

Proof of Theorem 14. In operational time, the process has the pathwise representation

$$X(u) = \delta N(u) - \kappa\int_0^u \psi(X(v))\,dv, \tag{48}$$

where N is a unit-rate Poisson process. Nonnegativity follows from the restoring flow and positive jumps. We treat the three cases separately.

Strictly positive limiting drift. Suppose $\psi_\infty < \infty$ and $a_\infty = \delta - \kappa\psi_\infty > 0$. Since $\psi(x) \le \psi_\infty$, (48) implies $X(u) \ge \delta N(u) - \kappa\psi_\infty u$. The strong law $N(u)/u \to 1$ therefore gives $\liminf X(u)/u \ge a_\infty$. Consequently $X(u) \to \infty$, so $\psi(X(u)) \to \psi_\infty$ and its time average tends to ψ_∞. Substituting back into (48) yields $X(u)/u \to a_\infty$. Eventually $X(u) \ge a_\infty u/2$, which makes $\int_0^\infty e^{-X(u)}du$ finite almost surely.

Strictly negative limiting drift. Suppose $\delta < \kappa\psi_\infty$. Choose $R > 0$ and $b > 0$ with $\delta - \kappa\psi(R) \le -b$. The operational generator applied to $V(x) = x$ satisfies $\mathcal{B}V(x) \le -b$ for $x \ge R$. Let τ_R be the first entrance into $[0, R]$, starting above R. Dynkin's formula, stopped also at time u and a finite upper level, gives $b\mathbb{E}(u \wedge \tau_R) \le x$ after removing the upper localization. Removal is justified by $0 \le X(v) \le x + \delta N(v)$ on finite intervals. Sending $u \to \infty$ gives $\mathbb{E}\tau_R \le x/b$, hence almost-sure return from every finite starting state.

For later use, define σ_0 as the first entrance into $[0, R]$, and recursively let σ_{j+1} be the first entrance after time $\sigma_j + 1$. These stopping times are almost surely finite. At each σ_j, the conditional probability of no Poisson jump in the next unit interval is e^{-1}. On that event the state remains below R, so the contribution to physical time is at least e^{-R}. The intervals are disjoint. Iterated conditioning bounds the probability of avoiding this event for m subsequent trials by $(1 - e^{-1})^m$. Infinitely many trials succeed almost surely, proving $t_\infty = \infty$.

The boundary case. Assume $\delta = \kappa\psi_\infty$ and $\psi_\infty - \psi(x) = o(1/x)$. For $V(x) = \log(1 + x)$, Taylor expansion yields

$$\begin{aligned}\mathcal{B}V(x) &= \log\left(1 + \frac{\delta}{1+x}\right) - \frac{\kappa\psi(x)}{1+x} \\ &= \frac{\kappa(\psi_\infty - \psi(x))}{1+x} - \frac{\delta^2}{2(1+x)^2} + O(x^{-3}) < 0\end{aligned}$$

for all sufficiently large x. Choose such an $R > 0$. Starting at $x > R$, stop on entering $[0, R]$ or reaching $H > x$. Lower exit occurs continuously, so $X = R$ there. Upper exit has $X \ge H$, and monotonicity gives $V(X) \ge \log(1 + H)$. The stopped process is bounded on $[R, H + \delta]$, so optional stopping of the nonnegative supermartingale $V(X)$ gives

$$\mathbb{P}_x(\text{reach } H \text{ before } [0, R]) \le \frac{\log(1+x)}{\log(1+H)}.$$

Exit from $[R, H]$ occurs almost surely. Indeed, the drift magnitude there is at least $\kappa\psi(R) > 0$, so a sufficiently long jump-free interval forces a lower exit. Such intervals have a fixed positive probability on disjoint trials. A path avoiding $[0, R]$ must therefore reach every finite upper level. Sending $H \to \infty$ in the displayed bound shows that the probability of avoidance is zero. The return-and-dwell argument from the preceding case again yields $t_\infty = \infty$. □

Interpreting the classification. Let τ_H be the first operational time at which X reaches H, and let $t_H = \int_0^{\tau_H} e^{-X(v)}dv$. Under positive limiting drift, $\tau_H \to \infty$ and $t_H \uparrow t_\infty < \infty$. In the returning cases, either some levels are never reached or their first-passage times tend to infinity with total physical time. For this positive-jump model all finite levels are in fact reachable by sufficiently many Poisson jumps in a short interval, and repeated compact returns provide repeated opportunities. The finite-level process can therefore reach a high level even in a returning regime. The theorem distinguishes accumulation of arbitrarily high levels on the limiting rare-feedback clock, with no interchange of the initialization and level limits.

F Scope, Approximation, and Observable Diagnostics

The previous appendices treated positive reward updates and a specified penalty. We now address the two features needed for more general training: movement without the selected feedback, and updates that temporarily reduce accuracy. The background example first shows why freezing the policy can fail. Signed coupling and full-state drift then give the corresponding distributional and budget guarantees.

F.1 PROOF OF THE BACKGROUND-TRANSFER EXAMPLE

The static model associates activation with a sequence of target-task updates. Shared-parameter learning permits another route: the target success probability can increase before its own informative groups arrive. A solvable example makes the distinction quantitative. Suppose that, conditional on receiving no target feedback, background learning increases target log-odds at speed $b > 0$. Target groups of size K are attempted at the times of an independent unit-rate Poisson process. This continuous-time example makes attempt times explicit and preserves the exact mixed-group probability $\lambda_K(p) = 1 - p^K - (1-p)^K$ at the current state.

Proposition 15 (A finite feedback window under background transfer)**.** *In this example, let* $p(t) = \mathrm{sigmoid}(\mathrm{logit}(\varepsilon) + bt)$ *along the no-target-feedback path. The probability that an informative target group never arrives is*

$$\mathbb{P}(T_1 = \infty) = \exp\left\{-\frac{1}{b}\int_{\varepsilon}^{1} \frac{1 - p^K - (1-p)^K}{p(1-p)}\, dp\right\} \xrightarrow[\varepsilon \downarrow 0]{} \exp\left\{-\frac{2H_{K-1}}{b}\right\}, \tag{49}$$

where $H_{K-1} = \sum_{j=1}^{K-1} j^{-1}$*. On the event of no target feedback, a fixed success level is still reached at time* $[\mathrm{logit}(p_\star) - \mathrm{logit}(\varepsilon)]/b$*, measured in the same physical time as the group-attempt process.*

The event of reaching the target before receiving any target feedback instead has probability

$$\mathbb{P}(T_1 > t_\star) = \exp\left\{-\frac{1}{b}\int_{\varepsilon}^{p_\star} \frac{\lambda_K(p)}{p(1-p)}\, dp\right\} > \mathbb{P}(T_1 = \infty), \qquad t_\star = \frac{\mathrm{logit}(p_\star) - \mathrm{logit}(\varepsilon)}{b}.$$

The infinite-horizon avoidance probability is therefore a lower bound on pre-activation avoidance, with the two events evaluated over different time horizons.

Faster background transfer carries the policy more rapidly through the intermediate-success region where mixed groups are common. Consequently, the target can become accurate while supplying little of its own learning signal. At fixed transfer speed, enlarging groups increases the total integrated feedback intensity as $2H_{K-1}/b$, which grows logarithmically with K. This differs from the linear $K\varepsilon$ arrival rate at a fixed rare-success state. The two expressions describe different trajectories through training, and both follow from keeping the background state evolution in the feedback calculation.

Proposition 15 is an exact worked example with a specified background path and randomized attempt times. A general shared network has a state-dependent, stochastic background path, for which Proposition 1 retains the appropriate expectation over paths. The example therefore identifies an interpretable mechanism to inspect in a trace: distinguish improvement caused by the selected task's feedback from improvement inherited before that feedback. This distinction also clarifies why an inactive target group and an inactive entire optimizer step have different meanings.

Until the first informative target group, the background path in Proposition 15 is deterministic. Thinning the unit-rate attempt process by the mixed-group probability produces first-event survival $\exp[-\int_0^t \lambda_K(p(v))dv]$. Since $dp/dt = bp(1-p)$, changing variables gives the first expression in (49). The identity

$$\frac{1 - p^K - (1-p)^K}{p(1-p)} = \frac{1 - p^{K-1}}{1-p} + \frac{1 - (1-p)^{K-1}}{p} = \sum_{j=0}^{K-2} \left[p^j + (1-p)^j\right]$$

shows that the integral over $[0, 1]$ equals $2H_{K-1}$. The integrand is a finite polynomial, so the integral from ε to one converges to that value. The fixed-success crossing time follows by solving the background logit equation. Even on paths that eventually receive feedback, the no-feedback path suffices to determine first-event survival because the future after the first event plays no role. This completes the proof of Proposition 15.

For fixed-grid attempts, the corresponding exact survival is the product $\prod_n [1 - \lambda_K(p(n))]$, as in Proposition 1. We distinguish this product from the Poisson-attempt expression. The harmonic formula belongs to the stated continuous-time example, while the shared-network experiment uses the discrete product and actual attempted groups.

F.2 A QUANTITATIVE CHECK ON FREEZING THE BACKGROUND

The example shows that background motion can matter. To quantify when freezing it is accurate, we compare the two survival products directly. For $a_j, b_j \in [0,1]$, telescoping gives

$$\left|\prod_{j<n}(1-a_j) - \prod_{j<n}(1-b_j)\right| \le \sum_{j<n}|a_j - b_j|.$$

Applying this pathwise to Proposition 1 yields the exact diagnostic bound

$$|\mathbb{P}_z(T_1 > n) - (1-\lambda(z))^n| \le \sum_{j<n}\mathbb{E}_z^{K_0}|\lambda(Z_j) - \lambda(z)|. \tag{50}$$

If λ is L-Lipschitz under a chosen metric, the right side is bounded by $L\sum_{j<n}\mathbb{E}d(Z_j, z)$. The bound depends on accumulated motion over the waiting interval, so a small one-step background update can still matter over many inactive steps. No global bound on log-likelihood or score gradients is required by this calculation.

For a hazard estimate $\widehat{\lambda}$ evaluated along the same K_0 paths, the same argument gives

$$\left|\mathbb{E}^{K_0}\prod_{j<n}(1-\widehat{\lambda}(Z_j)) - \mathbb{P}(T_1 > n)\right| \le \sum_{j<n}\mathbb{E}^{K_0}|\widehat{\lambda}(Z_j) - \lambda(Z_j)|.$$

If the estimated background kernel also differs, couple its paths with the true paths and add their disagreement probability. This elementary bound identifies which errors must be controlled for a prospective prediction. In the shared-network experiment the hazard is available exactly, so the test isolates the effect of evolving background states.

F.3 SIGNED ACTIVATION AND FINITE-INITIALIZATION ERROR

The positive-increment proof controlled the remaining wait by monotone progress. With signed increments, a path may return to a less successful state, so we instead use positive long-run drift and exponential-moment bounds. We begin with independent increments and then allow their law to depend on the current success probability.

Independent increments with finite first moment. Put $\theta_0 = \mathrm{logit}(\varepsilon)$, $b_\varepsilon = \mathrm{logit}(p_\star) - \theta_0$, and $N_\varepsilon = \inf\{j : S_j \ge b_\varepsilon\}$. The strong law gives $S_j/j \to \mu = \mathbb{E}D > 0$. Consequently $N_\varepsilon < \infty$ almost surely for each fixed ε, and $N_\varepsilon \to \infty$ almost surely as $\varepsilon \downarrow 0$. The latter follows because each finite initial segment of the random walk has a finite maximum. These statements permit arbitrary negative increments with an integrable absolute value.

Before event $j+1$, the success probability is

$$p_j^\varepsilon = \mathrm{sigmoid}(\theta_0 + S_j) = \frac{\varepsilon e^{S_j}}{1 - \varepsilon + \varepsilon e^{S_j}}.$$

Conditional on the embedded chain, use the independent E_j to generate the waiting times

$$G_j^\varepsilon = \left\lceil \frac{E_j}{-\log(1-\lambda(p_j^\varepsilon))} \right\rceil, \qquad T_\star = \sum_{j<N_\varepsilon} G_j^\varepsilon.$$

When a rate equals one, define its geometric waiting time to be one. For each fixed j, eventually $j < N_\varepsilon$, $p_j^\varepsilon/\varepsilon \to e^{S_j}$, and

$$c\varepsilon^\alpha G_j^\varepsilon \longrightarrow e^{-\alpha S_j}E_j \quad \text{almost surely.}$$

It remains to justify summing these eventwise limits, including after downward excursions. The lower rate bound and the pre-target inequality $p_j^\varepsilon < p_\star$ provide the needed domination. For $\varepsilon < 1/2$,

$$\varepsilon^\alpha/\lambda(p_j^\varepsilon) \le c_-^{-1}\varepsilon^\alpha(p_j^\varepsilon)^{-\alpha} \le Ce^{-\alpha S_j}, \qquad \varepsilon^\alpha \le Ce^{-\alpha S_j}.$$

The second inequality follows from $\varepsilon e^{S_j} < (1-\varepsilon)p_\star/(1-p_\star)$. Since $-\log(1-\lambda) \geq \lambda$ and $\lceil x \rceil \leq 1 + x$,

$$c\varepsilon^\alpha G_j^\varepsilon \mathbf{1}\{j < N_\varepsilon\} \leq Ce^{-\alpha S_j}(1 + E_j). \tag{51}$$

The series on the right is finite almost surely. Indeed, eventually $S_j \geq \mu j/2$, while $E_j \leq e^{\alpha\mu j/4}$ eventually almost surely by its exponential tail and the Borel–Cantelli lemma. Dominated convergence for the counting measure on j proves the almost-sure coupled convergence in (5). It also proves finiteness of the limiting series, independently of its moments.

Moment convergence and the exact first-moment boundary. Almost-sure convergence alone leaves expected cost open. To control its qth moment, assume $q \geq 1$ and $m(q) < 1$. The same dominating series is then summable in L^q, since Minkowski's inequality gives

$$\left\| \sum_{j \geq J} e^{-\alpha S_j}(1 + E_j) \right\|_q \leq \|1 + E_0\|_q \sum_{j \geq J} m(q)^{j/q} = \frac{\|1 + E_0\|_q \, m(q)^{J/q}}{1 - m(q)^{1/q}}.$$

Thus the coupled convergence is in L^q. Tonelli's theorem, which needs no integrability assumption for nonnegative summands, gives

$$\mathbb{E}W_{\alpha,\nu} = \sum_{j \geq 0} \mathbb{E}[e^{-\alpha S_j} E_j] = \sum_{j \geq 0} m(1)^j.$$

This proves (6), including the infinite-mean case. When $m(1) \geq 1$, Fatou's lemma additionally implies $\liminf_{\varepsilon \downarrow 0} \mathbb{E}[c\varepsilon^\alpha T_\star] = \infty$. A finite sample average in that regime supplies no estimate of a finite population mean.

State-dependent signed increments. The conditional update distribution generally changes with the policy. The next theorem controls this change by coupling its early updates to a limiting law and bounding the later waiting cost through a uniform fractional moment. The fractional exponent is sufficient for distributional forecasts even when the mean cost is infinite.

Theorem 16 (State-dependent signed activation)**.** *Assume inactive steps preserve the scalar state and $\lambda(p) \geq c_- p^\alpha$ before the fixed target. For state-dependent active laws ν_p, suppose $|D| \leq M$ under ν_p and ν, and, for some $\beta > 0$ and $0 < r \leq 1$,*

$$|\lambda(p)/(cp^\alpha) - 1| \leq Cp^\beta, \quad d_{\mathrm{TV}}(\nu_p, \nu) \leq Cp^\beta, \quad \sup_{p < p_\star} \mathbb{E}_{\nu_p} e^{-r\alpha D} \leq \bar{m} < 1.$$

Writing $\zeta = \min\{1, \alpha, \beta\}$ and $\ell_\varepsilon = \log(1/\varepsilon)$, the distance $d_{\mathrm{BL}}(\mathcal{L}(c\varepsilon^\alpha T_\star), \mathcal{L}(W_{\alpha,\nu}))$ is at most

$$C_\star \left[\varepsilon^{\zeta r/2} + \varepsilon^{\alpha r} \ell_\varepsilon^r + \varepsilon^{|\log \bar{m}|/(2M)} + \varepsilon^{\beta/2} \ell_\varepsilon \right]. \tag{52}$$

Constants depend on the displayed hypotheses and the fixed target, and remain independent of ε.

We prove (52) in two stages: first control the tail of the actual process, then couple a finite initial segment. Write Y_j for cumulative actual increments before feedback event $j + 1$, and stop at $N_\varepsilon = \inf\{j : Y_j \geq b_\varepsilon\}$. All probabilities below refer to this actual chain unless the independent limit chain S_j is specified. The conditional exponential-moment assumption implies, by induction,

$$\mathbb{E}[e^{-r\alpha Y_j} \mathbf{1}\{j < N_\varepsilon\}] \leq \bar{m}^j. \tag{53}$$

To check the induction, condition on a live state at event j, multiply by its conditional moment, and discard paths killed at the next event. The discarded terms are nonnegative. In particular, $\mathbb{P}(N_\varepsilon > j) \leq e^{r\alpha b_\varepsilon} \bar{m}^j$, proving that the actual embedded chain reaches the target almost surely. Total-variation convergence on a common bounded support also gives $m(r) \leq \bar{m}$.

Set $J = \lfloor \ell_\varepsilon/(2M) \rfloor$. For sufficiently small ε, neither chain can reach the fixed target during its first J events, since every increment is at most M. Every state in this initial segment satisfies $p_j^\varepsilon \leq 2\varepsilon e^{jM} \leq 2\sqrt{\varepsilon}$. Maximally couple each actual mark with an independent mark from ν. At each step before a mismatch, the conditional discrepancy probability is at most $C(2\sqrt{\varepsilon})^\beta$. Hence

$$\mathbb{P}(\text{a mismatch before } J) \leq CJ\varepsilon^{\beta/2}. \tag{54}$$

Use the same exponential variables to generate the waiting times on matching paths. The rate approximation, the logistic denominator, and $-\log(1-\lambda) = \lambda(1+O(\lambda))$ give the uniform estimate

$$\left| \frac{c\varepsilon^{\alpha}}{-\log(1-\lambda(p_j^{\varepsilon}))} - e^{-\alpha S_j} \right| \le C\varepsilon^{\zeta/2} e^{-\alpha S_j}, \qquad \zeta = \min\{1, \alpha, \beta\}.$$

The logistic denominator contributes $O(\varepsilon + p_j^{\varepsilon})$, which is also bounded by $C\sqrt{\varepsilon}$ even after downward excursions. Summing before J, the coupled waiting-time error is bounded by

$$C\varepsilon^{\zeta/2} \sum_{j<J} e^{-\alpha S_j} E_j + c\varepsilon^{\alpha} J.$$

For $0 < r \le 1$, subadditivity $(\sum a_j)^r \le \sum a_j^r$ gives

$$\mathbb{E}\left(\sum_{j<J} e^{-\alpha S_j} E_j \right)^r \le \Gamma(1+r) \sum_{j<J} m(r)^j \le \frac{\Gamma(1+r)}{1-\bar{m}}.$$

For a bounded one-Lipschitz test function, its difference is at most $\min\{2, |x-y|\} \le 2|x-y|^r$. The contribution of the matched initial segment is therefore at most $C[\varepsilon^{\zeta r/2} + \varepsilon^{\alpha r} J^r]$, while mismatches contribute at most twice (54).

We have controlled the coupled initial segment. For the remaining actual waits, condition on the embedded state. By Jensen's inequality, a geometric variable of rate λ has rth moment at most λ^{-r} when $r \le 1$. Combining the pre-target estimate in (51) with (53) therefore yields

$$\mathbb{E}\left(c\varepsilon^{\alpha} \sum_{j=J}^{N_\varepsilon - 1} G_j^{\varepsilon} \right)^r \le C \sum_{j \ge J} \bar{m}^j = \frac{C\bar{m}^J}{1-\bar{m}}.$$

The independent limiting remainder has the same bound with an additional factor $\Gamma(1+r)$. Since $\bar{m}^J \le \bar{m}^{-1}\varepsilon^{|\log \bar{m}|/(2M)}$, adding the two remainders, the initial waiting-time error, and the coupling error proves (52). Every exponent and every dependence on the conditional exponential moment is explicit.

For state-dependent L^q convergence with $q \ge 1$, a corresponding uniform condition $\sup_{p<p_\star} \mathbb{E}_{\nu_p} e^{-q\alpha D} \le \bar{m}_q < 1$ is sufficient. The preceding killed-moment induction and the geometric qth-moment bound give a summable L^q remainder by Minkowski's inequality. Finite-segment coupling then completes the proof. Convergence of the mean uses the stronger first-moment condition, while the fractional-moment bound already controls bounded budget-success predictions.

Checking the fractional-moment condition from bounded signed drift. A uniform positive conditional mean is sufficient for the fractional condition when increments are bounded. If $|D| \le M$ and $\mathbb{E}_{\nu_p} D \ge \mu > 0$ at every pre-target state, Hoeffding's elementary moment-generating-function bound gives

$$\mathbb{E}_{\nu_p} e^{-r\alpha D} \le \exp[-r\alpha\mu + (r\alpha)^2 M^2/2] \le \exp[-r\alpha\mu/2] < 1$$

for $0 < r \le \min\{1, \mu/(\alpha M^2)\}$. Therefore the distributional approximation admits bounded signed increments with a uniform positive mean, including the infinite-mean activation example. Finite expected training cost calls for control at exponent one, which is a genuinely stronger requirement.

Budget probabilities. For every signed law above, $W_{\alpha,\nu} = E_0 + R$ with $R \ge 0$ independent of E_0. Its density is bounded by one. Thus a bounded-Lipschitz error $r_\varepsilon \le 1$ implies a uniform distribution-function error at most $2\sqrt{r_\varepsilon}$ by the smoothing argument in Appendix C. This gives a finite-budget interpretation even when the limiting mean is infinite.

F.4 Exact finite-initialization mean for signed unit increments

A finite-initialization obstruction. The distinction persists before any asymptotic limit. Take $\lambda(p) = p$ and $D = +1$ with probability $q > 1/2$, otherwise $D = -1$. Let $H = \lceil \mathrm{logit}(p_\star) -$

$\mathrm{logit}(\varepsilon)\rceil$. Every run reaches the target almost surely, yet

$$\mathbb{E}T_\star = \begin{cases} \dfrac{H}{2q-1} + \dfrac{1-\varepsilon}{\varepsilon}\dfrac{1-e^{-H}}{1-qe^{-1}-(1-q)e}, & q > \dfrac{e}{1+e}, \\ \infty, & \dfrac{1}{2} < q \le \dfrac{e}{1+e}. \end{cases} \tag{55}$$

Thus favorable average movement can coexist with infinite mean activation time at a fixed initialization and fixed accuracy target. Budget-success probabilities remain well defined in this regime. The obstruction is caused by rare downward excursions that make subsequent feedback exponentially expensive. It gives a concrete necessity example for controlling negative exponential moments, beyond a technical sufficient condition in a proof. Let S_j be a nearest-neighbor random walk with upward probability $q > 1/2$, and let $N_H = \inf\{j : S_j = H\}$ for the integer H in (55). Positive drift implies $N_H < \infty$ almost surely. A Chernoff bound gives an exponential upper tail after a constant multiple of H, and optional stopping of $S_j - (2q-1)j$ consequently yields $\mathbb{E}N_H = H/(2q-1)$. This argument uses bounded increments and finite expected stopping time, so the optional-stopping hypotheses hold.

At state S_j, the waiting rate is exactly $\mathrm{sigmoid}(\mathrm{logit}(\varepsilon) + S_j)$, and its reciprocal equals $1 + (1-\varepsilon)e^{-S_j}/\varepsilon$. Therefore Tonelli's theorem gives the extended-valued identity

$$\mathbb{E}T_\star = \frac{H}{2q-1} + \frac{1-\varepsilon}{\varepsilon}\,\mathbb{E}\sum_{j<N_H} e^{-S_j}. \tag{56}$$

Put $m = qe^{-1} + (1-q)e$. If $m < 1$, the full series of expected costs equals $(1-m)^{-1}$. The strong Markov property at N_H, where $S_{N_H} = H$ exactly, decomposes that series into the pre-hit part and e^{-H} times an independent full series. Hence the expectation in (56) is $(1-e^{-H})/(1-m)$.

For completeness, we prove divergence also at the boundary $m = 1$. Define a tilted random walk with upward probability $\widetilde{q} = qe^{-1}/m$ and downward probability $1-\widetilde{q} = (1-q)e/m$. For each path of j increments its tilted probability is its original probability multiplied by e^{-S_j}/m^j. Thus

$$\mathbb{E}[e^{-S_j}\mathbf{1}\{N_H > j\}] = m^j\widetilde{\mathbb{P}}(N_H > j).$$

On $1/2 < q \le e/(1+e)$, one has $m \ge 1$ and $\widetilde{q} < 1/2$. The standard nearest-neighbor hitting calculation gives $\widetilde{\mathbb{P}}(N_H = \infty) = 1 - (\widetilde{q}/(1-\widetilde{q}))^H > 0$. It follows that the summands are bounded below by a positive constant, and their series diverges. This proves both branches of (55). Each individual activation time is nevertheless finite because it is a finite sum of almost-surely finite geometric waits.

F.5 Full-state activation and conditional-moment errors

The signed scalar law still treats success probability as a closed state. To retain the optimizer and representation, we instead apply a potential directly to the full training state. By the definition of V and conditioning on the selected event,

$$\mathbb{E}[V(Z_{n+1}) \mid Z_n = z] = V(z)[(1-\lambda(z))\rho_0(z) + \lambda(z)\rho_1(z)].$$

Consequently (13) is exactly the assertion $\mathbb{E}[V(Z_{n+1}) - V(Z_n) \mid Z_n = z] \le -a$ before activation. Stop at $T_\star \wedge n$. Successive conditioning, with the assumed finite conditional moments, gives

$$\mathbb{E}V(Z_{T_\star \wedge n}) + a\,\mathbb{E}(T_\star \wedge n) \le V(Z_0).$$

The first term is nonnegative. Monotone convergence in the second term proves the mean bound in (14). Markov's inequality gives the budget bound, and finite expectation implies almost-sure activation. This proof is the standard additive-drift argument and requires no independence of updates, no bounded increment, and no scalar closure of the full training state.

For the sufficient conditions in Theorem 6,

$$\lambda(1-\rho_1) + (1-\lambda)(1-\rho_0) \ge \lambda(1-\bar{\rho}_1 - b).$$

Since $V(z)p(z)^\alpha = (1-p(z))^\alpha \ge (1-p_\star)^\alpha$ before activation, multiplication by the lower rate bound gives the stated constant a. The condition compares the cost of adverse background movement directly with the gain supplied by selected feedback.

Suppose $\widehat{\rho}_i(z)$ estimates the two conditional moments and $|\widehat{\rho}_i(z) - \rho_i(z)| \le e_i(z)$ on the pre-target region. A valid lower bound on the true one-step drift is

$$V(z)\left[\lambda(z)(1-\widehat{\rho}_1(z)) + (1-\lambda(z))(1-\widehat{\rho}_0(z)) - \lambda(z)e_1(z) - (1-\lambda(z))e_0(z)\right].$$

A uniform positive lower bound on this expression supplies the same guarantee. A finite list of checkpoint measurements establishes this inequality only at those checkpoints. Extending it to the full reachable region needs a structural bound or an appropriate uniform confidence argument. This distinction is essential when the quantities are estimated from neural training traces.

For a localized region $\mathcal{D}$, let $\sigma = \inf\{n : Z_n \notin \mathcal{D}\}$. If the drift condition is proved only on $\mathcal{D} \cap \{p < p_\star\}$, the stopped proof bounds $\mathbb{E}(T_\star \wedge \sigma) \le V(Z_0)/a$, and

$$\mathbb{P}(T_\star > n) \le \frac{V(Z_0)}{an} + \mathbb{P}(\sigma \le n,\ \sigma < T_\star).$$

Thus a local audit yields a local hitting-or-exit guarantee. A global activation bound additionally needs control of the displayed exit probability. This makes explicit what must be supplied beyond measurements at a few checkpoints. When such region-wide control is unavailable, Appendix G.10 gives a different finite-budget statement based on signed progress along successive audited blocks.

F.6 Graded verifiers, batching, and optimizer memory

A graded verifier can define an event through a predeclared contrast threshold, a nonzero gradient contribution, or a selected outcome category. Conditional kernels then account for all updates on the complement. This extension is exact at the transition level. Applying the sharp scalar law additionally requires checking its local rate and increment assumptions for that event. Scores with continuous variation may lead to frequent small updates and a diffusion or deterministic approximation instead of the rare finite-jump limit.

With momentum or adaptive optimization, a group with zero current reward gradient can still move the policy through stored optimizer variables. Those variables belong in Z_n, and their induced motion belongs in K_0. Similarly, another prompt can move a selected task through shared parameters. The shared-network experiment retains this effect directly. Treating every homogeneous group as a static holding step is justified only after verifying that the relevant conditional background transition is effectively stationary over the prediction horizon.

Finally, the scalar examples use one on-policy step from freshly generated groups. They contain no stale-trajectory importance correction. The correlated-batch construction deliberately reuses a group within one simultaneous averaged update, with no intervening policy change. It is a dependence intervention at a fixed policy, so its matched-mean identity uses only trajectories sampled at that same policy.

G Temporal Dependence and Fixed-Accuracy Thresholds

The signed laws identify the exponential cost of regressions. We now determine how temporal dependence changes their accumulation at a fixed initialization and target. The proof first computes exponential moments along the event chain, then converts rare downward excursions into long rollout waits. This same decomposition later permits arbitrary sampling rates and resource costs.

G.1 The embedded chain and tilted matrix

We prove Theorem 4 for every fixed $\varepsilon, p_\star, \alpha, \delta$ in its statement. Write $S_0 = 0$, $S_j = \sum_{k=1}^{j} C_k$, $\theta_j = \operatorname{logit}(\varepsilon) + \delta S_j$, and

$$H = \left\lceil \frac{\operatorname{logit}(p_\star) - \operatorname{logit}(\varepsilon)}{\delta} \right\rceil, \qquad N_H = \inf\{j : S_j = H\}.$$

Here j counts feedback events. Conditional on the sign chain, the physical waits G_j for $j < N_H$ are independent geometric variables with rate $\lambda(\operatorname{sigmoid}(\theta_j))$. Thus $T_\star = \sum_{j<N_H} G_j$. The matrix R_ρ is strictly positive for $q \in (1/2, 1)$ and $\rho < 1$. We assume $q < 1$ throughout

this appendix, as implicit in the signed two-outcome setting. The deterministic case $q = 1$ has all positive moments and can be treated separately.

Let $A = q + \rho(1-q)$ and $B = 1 - q + \rho q$. These are the diagonal entries of R_ρ, with $A > B > 0$, $A + B = 1 + \rho$, and $\det R_\rho = \rho$. For $r > 0$ define

$$M_r = R_\rho \operatorname{diag}(e^{-r\alpha\delta}, e^{r\alpha\delta}).$$

Multiplication of transition probabilities along a path gives

$$\mathbb{E}_i[e^{-r\alpha\delta S_j}] = \boldsymbol{e}_i^\top M_r^j \mathbf{1}. \tag{57}$$

The moment classification is therefore reduced to the powers of a two-by-two matrix. Its determinant is ρ and its trace is $Ae^{-r\alpha\delta} + Be^{r\alpha\delta}$. Its eigenvalues are real and nonnegative because the off-diagonal entries are positive and the determinant is nonnegative. Writing $y = e^{r\alpha\delta} > 1$, the equation $\det(I - M_r) = 0$ is

$$By^2 - (A+B)y + A = (y-1)(By - A) = 0.$$

The positive nonzero crossing is therefore $r = \log(A/B)/(\alpha\delta) = \xi_\rho$. Between zero and this crossing the trace is smaller than $1 + \rho$, so the larger eigenvalue is strictly below one. Above the crossing it is strictly greater than one. This also covers $\rho = 0$, whose second eigenvalue is zero. Consequently, for each $0 < r < \xi_\rho$ there exist $C_r < \infty$ and $a_r < 1$ such that every component in (57) is at most $C_r a_r^j$.

G.2 Finite moments and almost-sure activation

We first establish eventual activation. The ergodic theorem gives $S_j/j \to 2q - 1 > 0$ almost surely, so $N_H < \infty$ almost surely. Since every pre-target rate is positive, each of these finitely many waits is also finite. Expected cost requires a separate argument, however, because the wait can become very large on paths with deep regressions. The matrix bound controls precisely those contributions.

For every $r > 0$, a geometric variable with parameter a obeys $\mathbb{E}G^r \le C_r a^{-r}$. For $r \le 1$ this follows from Jensen's inequality. For larger r, it follows by integrating the geometric tail bound $\mathbb{P}(G > t) \le e^{-a(t-1)}$ for $t \ge 1$. Before N_H, the lower rate assumption and the logistic formula imply

$$\lambda(\operatorname{sigmoid}(\theta_j))^{-r} \le c_-^{-r}(1 + e^{\theta_j})^{r\alpha} e^{-r\alpha\theta_j} \le C_{\varepsilon,p_\star,r} e^{-r\alpha\delta S_j}.$$

The target is fixed, so the prefactor is finite and independent of j. Set $Y_j = G_j \mathbf{1}\{j < N_H\}$. For $r < \xi_\rho$, equation (57) gives $\mathbb{E}Y_j^r \le Ca_r^j$. If $r \le 1$, subadditivity yields $\mathbb{E}T_\star^r \le \sum_j \mathbb{E}Y_j^r < \infty$. If $r > 1$, Minkowski's inequality yields

$$(\mathbb{E}T_\star^r)^{1/r} \le \sum_{j \ge 0} (\mathbb{E}Y_j^r)^{1/r} \le C^{1/r} \sum_{j \ge 0} a_r^{j/r} < \infty.$$

This proof works directly at fixed ε and never replaces the accuracy target by a vanishing one.

G.3 A lower tail bound including the critical moment

The preceding argument proves finiteness below the threshold. For the converse, we need a lower bound on the chance of visiting a state where feedback is very rare. Put $s = \alpha\delta\xi_\rho = \log(A/B)$. At the spectral crossing, a strictly positive vector v satisfies $M_{\xi_\rho} v = v$, so equation (57) implies that

$$h_j = e^{-sS_j} v(C_j)$$

is a martingale. Let τ_{-m} be the first event index with $S_j = -m$, and stop this martingale at $\tau_{-m} \wedge N_H$. The stopped process is bounded. The exit time is almost surely finite because, on the finite strip $[-m, H]$, a fixed-length sequence of positive signs has a probability bounded below uniformly over the current sign. Optional stopping is therefore justified by bounded convergence.

Let $v_{\min}, v_{\max}$ be the smallest and largest components of v. The lower boundary contribution gives the upper bound

$$\mathbb{P}_i(\tau_{-m} < N_H) \le \frac{v(i)}{e^{sm} v_{\min}}.$$

For a lower bound, choose a fixed integer L with $e^{sL}v(-1) > e^{-sH}v_{\max}$. A prefix of L negative signs has positive probability independent of m and reaches $(-L,-1)$ without hitting H. Starting from this state and taking $m > L$, optional stopping and the upper bounds on each terminal value give

$$e^{sL}v(-1) \le e^{sm}v_{\max}\mathbb{P}_{-L,-1}(\tau_{-m} < N_H) + e^{-sH}v_{\max}.$$

Multiplying the resulting positive lower bound by the prefix probability proves that constants $c, C > 0$ exist with

$$ce^{-sm} \le \mathbb{P}(\tau_{-m} < N_H) \le Ce^{-sm}, \qquad m > L. \tag{58}$$

Either initial sign, and hence stationary initialization, satisfies this statement with appropriate constants.

Having bounded the probability of a downward excursion, we translate its depth into waiting cost. At level $-m$, the upper rate assumption gives $\lambda(\mathrm{sigmoid}(\mathrm{logit}\,\varepsilon - \delta m)) \le C_0 e^{-\alpha\delta m}$. For a large physical budget t, choose $m = \lceil \log(2C_0(t+1))/(\alpha\delta) \rceil$. The rate at that state is at most $1/[2(t+1)]$. Its geometric holding time exceeds t with probability at least $1/2$, using $(1-a)^{\lfloor t \rfloor} \ge 1 - \lfloor t \rfloor a$. The future holding draw is independent conditional on the reached state. Combining with (58) yields

$$\mathbb{P}(T_\star > t) \ge c_1 t^{-\xi_\rho} \quad \text{for all sufficiently large } t. \tag{59}$$

For each $0 < r < \xi_\rho$, the finite-moment result gives $\mathbb{P}(T_\star > t) \le C_r t^{-r}$. Taking logarithms and then $r \uparrow \xi_\rho$ proves the logarithmic tail exponent in (11). Integrating (59) against $rt^{r-1}dt$ proves divergence for every $r \ge \xi_\rho$, including the boundary. We assert a logarithmic tail exponent, which accommodates the lattice structure, without assuming a constant-prefactor regular-variation limit.

G.4 An exact finite-initialization mean

For $\lambda(p) = p$ and $\xi_\rho > 1$, put

$$u = (I - R_\rho \,\mathrm{diag}(e^{-\delta}, e^{\delta}))^{-1}\mathbf{1}, \qquad \mu = 2q - 1.$$

The inverse exists because the tilted spectral radius is below one. With initial sign $i \in \{+1,-1\}$, the exact expectation is

$$\mathbb{E}_i T_\star = \frac{H + \rho(1-i)/(1-\rho)}{\mu} + \frac{1-\varepsilon}{\varepsilon}(u_i - e^{-\delta H}u_{+1}). \tag{60}$$

For stationary initialization, average this expression with weights $(q, 1-q)$. If $\xi_\rho \le 1$, Theorem 4 gives infinite mean, while the event-count mean in the first term stays finite.

To prove the first term, observe that $S_j + \rho(C_j - \mu)/(1-\rho) - \mu j$ is a martingale. Indeed, $\mathbb{E}[C_{j+1} \mid C_j] = \rho C_j + (1-\rho)\mu$ makes its conditional increment zero. For a sufficiently small positive exponent, (57) yields $\mathbb{P}(N_H > j) \le \mathbb{P}(S_j < H) \le Ce^{s'H}a^j$ with $a < 1$. Thus N_H has an exponential tail and optional stopping is valid. At its endpoint, $S_{N_H} = H$ and $C_{N_H} = +1$, giving the first term in (60).

For clarity, write $A = q + \rho(1-q)$, $B = 1 - q + \rho q$, and $\Delta = 1 + \rho - Ae^{-\delta} - Be^{\delta}$. Direct inversion gives

$$u_{+1} = \frac{1 - \rho e^{\delta}}{\Delta}, \qquad u_{-1} = \frac{1 - \rho e^{-\delta}}{\Delta}.$$

These expressions are positive in the finite-mean regime. The exact ladder identity uses $S_{N_H} = H$ and terminal sign $+1$, a property of nearest-neighbor increments. General increments require the random boundary overshoot and terminal state to be retained.

For the second term, the expected geometric wait at S_j is $1 + (1-\varepsilon)e^{-\delta S_j}/\varepsilon$. The unrestricted exponential sum has expectation u_i by the matrix geometric series. Applying the strong Markov property at N_H decomposes this sum into its pre-target part and $e^{-\delta H}u_{+1}$. Tonelli's theorem and the event-count identity complete the proof. At $\rho = 0$ and $\delta = 1$, the result reduces exactly to (55).

G.5 Arbitrary feedback rates and the resource being counted

A sampling controller can change the wait at a state and the number of trajectories charged for that wait. To separate these effects from the frequency of visiting the state, keep the embedded sign chain of Theorem 4. Put $A = q + \rho(1-q)$, $B = 1 - q + \rho q$, $r_0 = B/A$, and $\chi = \log(A/B)$. Let $p_k = \mathrm{sigmoid}(\mathrm{logit}\,\varepsilon + \delta k)$ and retain the target H and event count N_H. We now allow arbitrary positive feedback probabilities $\lambda(p_k)$ and count the resource spent during each attempt.

Theorem 17 (Fixed-target cost criterion)**.** *Assume realized attempt costs are nonnegative almost surely, with finite positive means $c(p_k)$ at each level k. Conditional on each entry to that level, assume repeated attempts have independent, identically distributed pairs of feedback indicators and costs. The cost of an attempt may depend on its own outcome. The embedded sign transition remains R_ρ. For total resource cost $\mathcal{B}_\star$ through activation,*

$$\mathbb{E}\mathcal{B}_\star < \infty \quad \Longleftrightarrow \quad \sum_{m\geq 1} e^{-\chi m}\,\frac{c(p_{-m})}{\lambda(p_{-m})} < \infty.$$

In particular, if $\lambda(p) \asymp p^\alpha \ell(1/p)$ as $p \downarrow 0$, with ℓ positive and slowly varying, the attempted-step mean is finite for $\chi > \alpha\delta$ and infinite for $\chi < \alpha\delta$. At equality,

$$\mathbb{E}T_\star < \infty \quad \Longleftrightarrow \quad \sum_{m\geq 1} \frac{1}{\ell(e^{\delta m})} < \infty.$$

Thus $\ell(x) = (\log x)^\beta$ gives finite critical mean exactly when $\beta > 1$.

Here $f \asymp g$ means two-sided comparison by fixed positive constants for sufficiently small p. A positive measurable function is slowly varying if $\ell(tx)/\ell(x) \to 1$ for every fixed $t > 0$ as $x \to \infty$. We use the usual local boundedness conditions. All rates and mean costs at the finitely many remaining pre-target lattice levels are positive and finite. The theorem keeps both initialization and target accuracy fixed.

Proof. The embedded count N_H has an exponential tail by the tilted-matrix argument already proved. Define its killed occupation numbers

$$u_k^{\pm} = \mathbb{E}\sum_{j<N_H} \mathbf{1}\{S_j = k, C_j = \pm 1\}, \qquad g(k) = u_k^+ + u_k^-.$$

They are finite because $\mathbb{E}N_H < \infty$. For every edge $(k, k+1)$ with $k < 0$, a path starting at zero and ending at H has zero net crossings. Expected upward crossings equal $Au_k^+ + (1-B)u_k^-$, while downward arrivals at k equal u_k^-. Hence

$$Au_k^+ = Bu_k^-, \qquad u_k^+ = r_0 u_k^-.$$

For $k \leq -2$, the incoming negative-sign occupation equation gives

$$u_k^- = (1-A)u_{k+1}^+ + Bu_{k+1}^- = r_0 u_{k+1}^-.$$

Consequently,

$$g(-m) = C_H r_0^m, \qquad C_H = (1+r_0)u_{-1}^-/r_0 \in (0,\infty), \quad m \geq 1.$$

Positivity follows because a negative initial move has positive probability under either initial sign. Finiteness follows from $\mathbb{E}N_H < \infty$. This relation counts every return to a level, including those following a long regression.

At a fixed level, write $(I_j, C_j^{\mathrm{cost}})$ for successive attempt outcomes and costs. The event that attempt j is reached depends only on the previous $j-1$ outcomes. Independence across attempts therefore yields

$$\mathbb{E}\sum_{j\geq 1} C_j^{\mathrm{cost}}\mathbf{1}\{I_1 = \cdots = I_{j-1} = 0\} = \sum_{j\geq 1} c(p)(1-\lambda(p))^{j-1} = \frac{c(p)}{\lambda(p)}.$$

No independence between an attempt's cost and its own outcome is required. Applying Tonelli's theorem over visits and levels gives the extended-valued identity

$$\mathbb{E}\mathcal{B}_\star = \sum_{k=0}^{H-1} g(k)\frac{c(p_k)}{\lambda(p_k)} + C_H \sum_{m\geq 1} r_0^m \frac{c(p_{-m})}{\lambda(p_{-m})}.$$

The first sum is finite, proving the equivalence.

For the slowly varying specialization, the exact logistic relation is $p_{-m}^{-1} = 1 + e^{-\operatorname{logit}\varepsilon} e^{\delta m}$. The uniform convergence property of slowly varying functions gives $\ell(1/p_{-m})/\ell(e^{\delta m}) \to 1$ (Bingham et al., 1987). Thus the negative-level summands for $c = 1$ are comparable to

$$\frac{\exp[(\alpha\delta - \chi)m]}{\ell(e^{\delta m})}.$$

Since $\log \ell(e^{\delta m}) = o(m)$, a strictly negative exponent gives convergence and a strictly positive exponent gives divergence. At equality the stated series remains. For the logarithmic example it is a constant multiple of $\sum_m m^{-\beta}$, completing the proof. □

Which tail conclusions remain unchanged? Slow variation supplies, for every sufficiently small $\zeta > 0$, constants with $c_\zeta p^{\alpha+\zeta} \le \lambda(p) \le C_\zeta p^{\alpha-\zeta}$ on all reachable pre-target levels. The lower bound and the tilted-matrix proof imply $\mathbb{E}T_\star^r < \infty$ whenever $r(\alpha + \zeta)\delta < \chi$. The upper bound and the negative-excursion argument imply $\mathbb{P}(T_\star > t) \ge c'_\zeta t^{-\chi/((\alpha-\zeta)\delta)}$ for large t. Letting $\zeta \downarrow 0$ preserves the logarithmic tail exponent $-\chi/(\alpha\delta)$ and the classification of every noncritical moment. The critical mean can change even while that exponent stays fixed. Theorem 4's critical divergence remains valid under its original constant-factor power bounds.

Corollary 18 (Group attempts and generated trajectories can have different mean regimes)**.** *At each embedded state, generate independent terminal Bernoulli rewards in groups of a finite integer size $K(p) \ge 2$, with feedback event defined by mixed rewards. Keep the embedded update law fixed. Then*

$$\mathbb{E}T_{\text{groups}} < \infty \iff \sum_{m\ge 1} \frac{e^{-\chi m}}{\min\{K(p_{-m})p_{-m}, 1\}} < \infty,$$

$$\mathbb{E}\mathcal{R}_{\text{trajectories}} < \infty \iff \sum_{m\ge 1} e^{-\chi m} \max\{K(p_{-m}), p_{-m}^{-1}\} < \infty.$$

For $K(p) \asymp p^{-\gamma}$ with $0 \le \gamma \le 1$, the trajectory mean is finite exactly when $\chi > \delta$. At $\chi = \delta$, the allocation $K(p) = \max\{2, \lceil(\log(1/p))^\beta\rceil\}$ with $\beta > 1$ gives finite mean group attempts and infinite mean generated trajectories.

Proof. For $0 < p \le 1/2$ and every $K \ge 2$,

$$e^{-2}\min\{Kp, 1\} \le 1 - (1-p)^K - p^K \le \min\{Kp, 1\}.$$

The upper bound follows from the event of at least one success. If $Kp \le 1$, exactly one success has probability $Kp(1-p)^{K-1} \ge e^{-2}Kp$, using $\log(1-p) \ge -2p$. If $Kp \ge 1$, the contrast probability is at least $1 - e^{-1} - 1/4 > e^{-2}$. Apply Theorem 17 first with cost one and then with cost K. Finitely many levels with $p > 1/2$ have finite contributions. For polynomial allocation with $\gamma \le 1$, $\max\{K(p), p^{-1}\} \asymp p^{-1}$, giving the strict condition $\chi > \delta$. In the logarithmic example, $K(p)p \to 0$, so the singleton lower bound and the union upper bound imply $\lambda_K(p) \sim K(p)p$. The group series is comparable to $\sum_m m^{-\beta}$, while the trajectory series has summands bounded below by a positive constant. This proves the separation at the common fixed target. □

The corollary isolates what can be achieved by changing the sampling clock alone. A bounded group-size controller retains exponent one for independent binary outcomes. Unbounded allocation can instead reduce the number of group attempts, as the logarithmic example shows, while generating enough trajectories to leave mean resource cost infinite. In ordinary group-normalized training, changing K also changes the conditional update. That additional effect belongs in the embedded dynamics and can provide further gains. Counting both effects separately prevents a larger group from appearing cheaper simply because it is recorded as one step.

History-dependent allocation and heterogeneous prompts. If the controller changes between failed attempts, geometric holding is replaced by a conditional comparison. Suppose

that at each attempt before feedback, its hazard lies in $[\lambda_k^-, \lambda_k^+] \subset (0,1]$ and its conditional mean nonnegative cost lies in $[c_k^-, c_k^+] \subset (0,\infty)$. Iterated conditioning bounds survival by the corresponding geometric survivals and gives

$$\frac{c_k^-}{\lambda_k^+} \le \mathbb{E}[\text{holding cost} \mid \text{entry history}] \le \frac{c_k^+}{\lambda_k^-}.$$

If the same embedded sign law is preserved, inserting these bounds into the occupation sum gives two-sided mean-cost bounds. This extension concerns expectations, with no claim about its full tail law. For prompt-specific independent groups with selection probabilities $w_x(z)$, the target-prompt hazard is $w_x(z)\lambda_{K_x}(p_x(z))$. The aggregate hazard is their sum. Substituting an average success rate into a binomial-group formula can obscure changes in prompt allocation, so estimation should retain the prompt or task stratum.

G.6 Distributional limit and its numerical tail certificate

If additionally $\lambda(p)/(cp^\alpha) \to 1$ at zero, the same sign path and independent $E_j \sim \mathrm{Exp}(1)$ couple the process to

$$W_\rho = \sum_{j\ge 0} e^{-\alpha\delta S_j} E_j, \qquad c\varepsilon^\alpha T_\star \longrightarrow W_\rho \quad \text{almost surely.}$$

For each fixed event index, the inverse-transform geometric wait converges after scaling to $e^{-\alpha\delta S_j}E_j$. The lower rate bound dominates the scaled pre-target wait by $Ce^{-\alpha\delta S_j}(E_j+1)$. Since $S_j/j \to \mu > 0$, this dominating series is finite almost surely, and its tail vanishes. Also $N_H \to \infty$ as $\varepsilon \downarrow 0$ on every fixed sign path. Dominated convergence for the series proves the claim. The matrix moment bounds above give L^r convergence for $1 \le r < \xi_\rho$.

For the numerical series cut after J terms and any $0 < r < \min\{1, \xi_\rho\}$, subadditivity supplies the explicit certificate

$$\mathbb{P}\left(\sum_{j\ge J} e^{-\alpha\delta S_j} E_j > b\right) \le \frac{\Gamma(1+r)}{b^r}\, \pi^\top M_r^J (I - M_r)^{-1}\mathbf{1}.$$

This controls omitted series error separately from Monte Carlo error. It remains applicable when the full mean is infinite, because the certificate uses a finite fractional moment. The experiment records this quantity for every displayed dependent limiting law.

G.7 Finite-window matching and opposite cost regimes

We next ask whether short update histories determine the cost classification. The construction must preserve every short word while changing the frequency of long harmful excursions. We achieve this by keeping the interiors and short boundary contexts of two blocks fixed, and changing only how the blocks follow one another. Fix $L, \alpha, \delta, \varepsilon, p_\star$ as in Theorem 5, and write $s = \alpha\delta$, $\theta_0 = \operatorname{logit}\varepsilon$, and $H = \lceil(\operatorname{logit} p_\star - \theta_0)/\delta\rceil$. Use the words

$$A = +^{6L+1}, \qquad B = +^L -^{4L} +^L$$

have signed sums $6L+1$ and $-2L$. Consecutive block types follow

$$Q_\omega = \begin{pmatrix} \omega & 1-\omega \\ 1-\omega & \omega \end{pmatrix}, \qquad 0 < \omega < 1.$$

A hidden state records the block type and the next position to emit. It advances deterministically inside a block and uses Q_ω at a block boundary. There are $P_L = 12L+1$ hidden states. The emitted sign at event j is C_j, and $S_n = \sum_{j=1}^n C_j$.

Stationarity, mixing, and positive event-count progress. Each position has invariant probability $1/P_L$. Interior positions have one predecessor, while incoming transition probabilities at either first position sum to one. The chain is irreducible. Return lengths from the first position

of A include $6L+1$ and $12L+1$, which have greatest common divisor one. It is therefore aperiodic. A finite irreducible aperiodic chain is exponentially mixing, as is any bounded observation of it. Its stationary mean sign is $(4L+1)/P_L > 0$, independent of ω.

The ergodic theorem gives $S_n/n \to (4L+1)/P_L$, so $N_H = \inf\{n : S_n \geq H\} < \infty$ almost surely. In fact N_H has an exponential tail. To verify this, let P_ω be the position kernel and $c(x)$ its emitted sign. The tilted matrix $F_\gamma(x,y) = e^{-\gamma c(x)} P_\omega(x,y)$ has a simple Perron eigenvalue $\kappa(0) = 1$. Differentiation at zero gives $\kappa'(0) = -\sum_x \pi(x)c(x) < 0$. For a sufficiently small $\gamma > 0$, its powers satisfy $\mathbb{E}e^{-\gamma S_n} \leq C\kappa(\gamma)^n$ with $\kappa(\gamma) < 1$. Since $\{N_H > n\} \subseteq \{S_n < H\}$, Markov's inequality yields $\mathbb{P}(N_H > n) \leq Ce^{\gamma H}\kappa(\gamma)^n$. Constants depend on the specified process. No uniform mixing rate across increasing L is assumed.

The exact observation horizon. Every negative run has length $4L$ and every positive run between negative runs has length at least $2L$. A window of length $k \leq 2L+1$ can contain at most one sign change. Its nonzero word probabilities are

$$\begin{aligned} \mathbb{P}(C_{1:k} = -^k) &= (4L-k+1)/P_L, \\ \mathbb{P}(C_{1:k} = +^k) &= (8L+2-k)/P_L, \\ \mathbb{P}(C_{1:k} = +^j -^{k-j}) &= \mathbb{P}(C_{1:k} = -^j +^{k-j}) = 1/P_L, \quad 1 \leq j < k. \end{aligned} \tag{61}$$

Each mixed word identifies one prescribed offset from a negative-run boundary. That boundary has stationary frequency $1/P_L$. The all-negative expression counts positions inside a negative run, and the all-positive expression follows by normalization. None depends on ω. In contrast,

$$\mathbb{P}(C_{1:2L+2} = - +^{2L} -) = \omega/P_L.$$

This word requires two consecutive B blocks, so $2L+2$ is the first window length that identifies the differing transition probability. The equality concerns distributions of individual consecutive windows. Temporal dependence between separated windows, and their associated waiting times, can distinguish the processes.

A finite-cost choice. Put $a_L = e^{-s(6L+1)}$, $b_L = e^{2sL}$, and

$$\omega_{\mathrm{f}} = \frac{1 - a_L b_L}{2(a_L + b_L - 2a_L b_L)}, \qquad \omega_{\mathrm{s}} = \frac{1 + b_L^{-1}}{2}.$$

Since $a_L b_L = e^{-s(4L+1)} < 1$ and $a_L + b_L - 2a_L b_L = (1 - a_L b_L) + (b_L - 1)(1 - a_L)$, we have $0 < \omega_{\mathrm{f}} < 1/2 < \omega_{\mathrm{s}} < 1$. At complete-block boundaries the matrix for negative exponential progress is $K_\omega = \mathrm{diag}(a_L, b_L) Q_\omega$. It satisfies

$$\det(I - K_\omega) = 1 - a_L b_L - \omega(a_L + b_L - 2a_L b_L).$$

At ω_{f} this determinant equals $(1 - a_L b_L)/2 > 0$, while $\det K_{\omega_{\mathrm{f}}} = a_L b_L (2\omega_{\mathrm{f}} - 1) < 0$. The two eigenvalues have opposite signs. Positivity of $\det(I - K)$ therefore places the Perron eigenvalue below one. The matrix series $\sum_{j \geq 0} K_{\omega_{\mathrm{f}}}^j$ converges entrywise.

Before the fixed target, the rate lower bound gives

$$\lambda(\mathrm{sigmoid}(\theta_0 + \delta S_n))^{-1} \leq C_\star e^{-sS_n}, \qquad C_\star = c_-^{-1} e^{-\alpha\theta_0} (1 + e^{\mathrm{logit}\, p_\star})^\alpha.$$

Conditional on the embedded sign path, geometric waits have these reciprocals as their means. The initial partial block has bounded length and downward displacement at most $4L$. Every subsequent full block has length at most $6L+1$ and internal downward displacement at most $3L$ relative to its start. Grouping the nonnegative series by complete blocks bounds its expectation by a constant times $1 + \| \sum_{j \geq 0} K_{\omega_{\mathrm{f}}}^j \mathbf{1} \|_\infty$. Consequently $\mathbb{E}T_\star^{\mathrm{f}} \leq C_\star \mathbb{E} \sum_{n \geq 0} e^{-sS_n} < \infty$. This argument includes stationary initialization inside a block and preserves the fixed target throughout.

An infinite-cost choice without premature target crossing. Let $x_- = (B, L+1)$ be the first negative position. It has stationary probability $1/P_L$. Starting there, selecting B at the next boundary returns to x_- after the word $-^{4L}+^{2L}$, whose sum is $-2L$ and whose every partial

sum is nonpositive. Repeating this cycle j times therefore avoids every positive target. Let E_j be the event of starting at x_- and repeating the cycle j times. Then

$$\mathbb{P}(E_j) = \omega^j / P_L, \qquad S_{6Lj} = -2Lj, \qquad N_H > 6Lj \quad \text{on } E_j.$$

Using $\mathrm{sigmoid}(x) \le e^x$ and the upper rate bound, the expected geometric wait at that event is at least $c_+^{-1} e^{-\alpha\theta_0} b_L^j$. These are distinct waiting terms for different j, even though the events E_j are nested. Tonelli's theorem gives

$$\mathbb{E}T_\star \ge \frac{e^{-\alpha\theta_0}}{c_+ P_L} \sum_{j\ge 0} (\omega b_L)^j.$$

At ω_s, $\omega_s b_L = (1 + b_L)/2 > 1$, proving infinite mean. All paths still activate almost surely because N_H is finite and every individual geometric rate is positive. This completes the proof.

What the construction adds. Finite-word matching has combinatorial precedents (Karhumäki et al., 2013), and tilted matrices are standard Markov-additive tools. The result proved here is their explicit fixed-target consequence: complete finite-window laws, bounded signed updates, positive drift, and exponential mixing together leave activation integrability unidentified. The hidden-state count grows with the requested matching window. Knowledge of a uniform memory bound would add information absent from this theorem.

G.8 Noise-aware certificates for conditional progress

The previous results specify which conditional moments matter. Estimating them requires two sources of randomness to be separated: the training transition and the evaluation of its resulting policy. Fix the complete state z and event stratum i, and restore that state, including optimizer memory, before each of n independent transitions from $K_i(z, \cdot)$. Write $p_0 = p(z)$, $p_j^+ = p(Z_j^+)$, $D_j = \mathrm{logit}\, p_j^+ - \mathrm{logit}\, p_0$, and $Y_j = e^{-\alpha D_j}$. The proposition below combines uncertainty from both stages.

Proposition 19 (A conditional-moment upper certificate)**.** *Assume a known structural regression bound $D_j \ge -M$, and put $R = e^{\alpha M}$. Choose simultaneous valid binomial intervals $p_0 \in [\ell_0, u_0]$ and $p_j^+ \in [\ell_j, u_j]$, with total failure probability at most η_{eval}. Write $o(p) = p/(1-p)$ and define*

$$U_j = \min\{R, [o(u_0)/o(\ell_j)]^\alpha\}, \qquad \overline{\rho}_i(z) = \min\left\{R, \frac{1}{n}\sum_{j=1}^n U_j + R\sqrt{\frac{\log(1/\eta_{\mathrm{samp}})}{2n}}\right\}.$$

An unbounded endpoint ratio is interpreted as $+\infty$. Then $\rho_i(z) \le \overline{\rho}_i(z)$ with probability at least $1 - \eta_{\mathrm{eval}} - \eta_{\mathrm{samp}}$.

Proof. On simultaneous evaluation coverage, monotonicity of odds gives $Y_j \le U_j$ for every j. Independently sampled transitions give independent latent variables $Y_j \in [0, R]$ with common mean $\rho_i(z)$. Hoeffding's inequality bounds this mean by their empirical mean plus the stated radius. A union bound with the evaluation event completes the claim. The two events need no mutual independence. Fixed-sample binomial intervals can be obtained by exact binomial-tail inversion and a union bound. Evaluation sample sizes must be fixed in advance, or replaced by valid sequential confidence bounds. □

A zero-success evaluation can yield $\ell_j = 0$, hence $U_j = R$. This is an honest inconclusive result. Introducing a pseudocount and then exponentiating a point estimate supplies no coverage guarantee. With a known tail bound $\mathbb{E}Y^{1+s} \le M_s$, $s > 0$, the hard bound can be replaced by a predetermined truncation level b. Applying the same argument to $Y \wedge b$ and using $\mathbb{E}(Y - b)_+ \le M_s/b^s$ gives

$$\rho_i(z) \le \frac{1}{n}\sum_j U_j^{(b)} + b\sqrt{\frac{\log(1/\eta_{\mathrm{samp}})}{2n}} + \frac{M_s}{b^s}.$$

The last term accounts for unseen regressions, so the result bounds the full exponential moment.

Table 3: Quantities used in a feedback-cost forecast. Estimates refer to a declared prompt, decoder, verifier, and optimizer state.

Quantity	Measurement	Interpretation
λ, α	Mixed-score frequency and its dependence on success probability	Feedback availability under the actual sampling rule
D, δ	Success log-odds before and after the complete update	D is measured progress. Constant δ requires the specified scalar model
$m(r)$, $\rho_i(z)$	Exponentiated progress from independent restored-state continuations	Conditional recovery-cost multiplier with evaluation and tail uncertainty
ρ	Conditional sign transitions, when a homogeneous two-state law applies	Persistence of actual progress, distinct from optimizer coefficients

A certificate for the full-state theorem. With simultaneous upper bounds $\overline{\rho}_0, \overline{\rho}_1$ and a hazard interval $[\ell_\lambda, u_\lambda]$, define

$$Q(z) = \max_{\lambda \in \{\ell_\lambda, u_\lambda\}} [(1-\lambda)\overline{\rho}_0(z) + \lambda \overline{\rho}_1(z)].$$

The expression is affine in λ, so its maximum is attained at an endpoint. If $Q(z) < 1$ and $V(z) \geq V_-(z) > 0$, then $V_-(z)(1 - Q(z))$ is a certified lower drift at that state. A uniform positive infimum over the reachable pre-target region implies Theorem 6. A finite collection of checkpoints supplies local certificates unless additional regularity controls the remaining states. For a known finite-state dependence model, simultaneous entrywise bounds on a tilted matrix $\overline{M}$ can instead be checked through a positive vector v satisfying $\overline{M}v \leq (1-\gamma)v$, $\gamma > 0$. Marginal histograms discard the transitions needed for that comparison.

Why an unrestricted finite-sample upper certificate is unavailable. Even perfectly observed increments cannot reveal an unconstrained negative tail. Compare a point mass at $d > 0$ with $(1-\zeta)\,\mathrm{Dirac}(d) + \zeta\,\mathrm{Dirac}(-B)$. Their n-sample laws differ in total variation by at most $n\zeta$. Yet the exponential moment of the second law is $(1-\zeta)e^{-\alpha d} + \zeta e^{\alpha B}$. Taking B proportional to $\log(1/\zeta)$ makes this moment arbitrarily large while preserving positive mean increment as $\zeta \downarrow 0$. Thus a known tail restriction carries substantive information. An empirical maximum regression is insufficient as its replacement. Failure of an upper certificate to fall below one also differs from proof of infinite activation cost.

Evaluation noise has two distinct effects. If a noisy increment is $D + \eta$ with conditionally centered η, convexity raises the expected exponential. Binomial plug-in logits also have their own bias and can be undefined at boundary counts, so conditional centering requires justification. Away from those boundaries, independent evaluations of size m give the familiar local expansions

$$b(p) \simeq \frac{2p-1}{2mp(1-p)}, \qquad v(p) \simeq \frac{1}{mp(1-p)}.$$

The relative exponential correction is approximately $1 - \alpha[b(p_+) - b(p_-)] + \alpha^2[v(p_+) + v(p_-)]/2$. This calculation explains sensitivity, while Proposition 19 supplies the finite-sample guarantee without relying on the approximation.

G.9 A PRACTICAL ESTIMATION AND FORECAST PROTOCOL

We now put the measurements together into a forecast. The starting point is the training event and its resource cost, since they determine which hazard and increment should be measured. An ordinary training trace then describes the sequence of visited states, whereas independent continuations from a restored checkpoint estimate a conditional transition law. Keeping these roles separate makes the following quantities interpretable.

1. **Fix the event and cost.** Record every attempted group, its prompt, group size, verified outcomes, and generated tokens. Include rejected groups, replay age, aborted generations, and optimizer steps. A selected-update count answers a different question from a generated-data count.

2. **Determine feedback scarcity.** For independent complete answers at a fixed prompt with fixed K, use the exact function $h_K(p) = 1 - p^K - (1-p)^K$, hence $\alpha = 1$. Its finite-range elasticity is $ph'_K(p)/h_K(p)$ and varies with the operating success range. For a different controller, measure its event directly at comparable frozen states and fit a declared finite-range model with uncertainty. Prompt mixture, K, or acceptance-rule changes require separate strata.
3. **Measure actual progress.** Evaluate success immediately before and after the complete optimizer update. With simultaneous intervals $p_t \in [\ell_t, u_t]$, the logit increment belongs to $[\operatorname{logit} \ell_{t+1} - \operatorname{logit} u_t, \operatorname{logit} u_{t+1} - \operatorname{logit} \ell_t]$. A regression is resolved only when the upper endpoint is negative. The scalar $\delta = \eta\sqrt{K-1}/K$ belongs to the specified Bernoulli SGD model. Neural Adam updates require their measured distribution, including inactive motion.
4. **Audit exponential moments.** Propagate these intervals through $e^{-r\alpha D}$ for a predeclared exponent grid. At a reset state use Proposition 19. Along a changing trace, bounded-variable concentration estimates the average conditional moment over those visited states. A hard truncation estimates a truncated moment unless an explicit tail remainder is added.
5. **Track cumulative regression and recovery.** Report the drawdown $\max_{s \le t} \theta_s - \theta_t$, the following period without informative feedback, and attempts or tokens until the previous level is recovered. Retain unresolved signs and unrecovered episodes. Checkpoint observations measure checkpoint-scale drawdowns and can miss shorter excursions.
6. **Test a forecast on fresh outcomes.** Freeze each prediction before its observation cohort or subsequent training block. Report finite-horizon success and censored cost. A forecast can be rejected using bounded sequential tests. Passing such a test alone gives no certificate of an infinite-horizon mean.

The parameter ρ deserves a separate check. Under the homogeneous two-state sign model, let $A = \mathbb{P}(C_{j+1} = +1 \mid C_j = +1)$ and $B = \mathbb{P}(C_{j+1} = -1 \mid C_j = -1)$. Then

$$\rho = A + B - 1, \qquad q = \frac{1-B}{2-A-B}.$$

These equalities follow by solving the two diagonal entries of (9). Applying its threshold additionally requires constant increment magnitude, the stated feedback-rate bounds, and the absence of inactive motion. If the next-sign law changes after conditioning on longer history or optimizer state, the two-state reduction must be replaced by the corresponding conditional law. In particular, a fitted short-history transition matrix alone cannot certify the tail regime, as Theorem 5 demonstrates.

Evaluation error and finite-range estimation. For each frozen policy and prompt, exact binomial intervals provide fixed-sample coverage. Allocate error across checkpoints and prompts before inspection. Repeated sequential looks require a confidence sequence or a summable allocation such as $\eta_j = 6\eta/[\pi^2(j+1)^2]$. For a proposed model $|\log \lambda_j - c_0 - \alpha \log p_j| \le b_j$, intersect its constraints with the success and event-rate intervals and project the feasible set onto α. The tolerances b_j are model assumptions specified before the test. An empty set rejects the model and a wide set remains inconclusive. This supplies a finite-range analysis, with no identification of an unobserved asymptotic exponent.

A sequential test for delayed feedback. Let $\ell_t < 1$ be a predictable lower forecast for $\lambda_t = \mathbb{P}(I_t = 1 \mid \mathcal{F}_t)$, where $\mathcal{F}_t$ contains the full history before the next outcome. Under $\lambda_t \ge \ell_t$, the process

$$E_n = \prod_{t<n} \frac{1-I_t}{1-\ell_t}$$

is a nonnegative supermartingale starting at one, because each conditional factor has mean $(1-\lambda_t)/(1-\ell_t) \le 1$. Stopping when $E_n \ge 1/\eta$ has false-alarm probability at most η: stop also at any fixed horizon, apply the supermartingale inequality, and let that horizon grow. During a period without informative feedback, the test rejects when $\sum_{t<n} -\log(1-\ell_t) \ge \log(1/\eta)$. It detects waiting inconsistent with the stated forecast. New episodes require allocated error budgets or a predefined mixture. The conclusion is a forecast violation, distinct from identifying an infinite-mean regime.

G.10 FINITE-BUDGET CERTIFICATES WITH SIGNED BLOCK PROGRESS

The global drift theorem requires an inequality on every reachable pre-target state. A finite-budget audit can ask a more local question: has enough conditional progress accumulated across the blocks actually encountered? This allows some blocks to move adversely, and a bounded potential removes the need to bound an unseen exponential tail. Fix $\beta > 0$, a reference θ_{ref}, and a cap B_0 before monitoring, and define

$$\Psi(z) = \min\{B_0, e^{-\beta(\theta(z)-\theta_{\mathrm{ref}})}\}, \qquad v_\star = e^{-\beta(\operatorname{logit} p_\star - \theta_{\mathrm{ref}})} < B_0.$$

Use the continuous endpoint conventions $\Psi(0) = B_0$ and $\Psi(1) = 0$. The bounded potential still separates every pre-target state, where $\Psi > v_\star$, from the target. The cap preserves target separation and allows a finite-budget argument with bounded observations.

Choose block lengths $1 \le h_k \le h_{\max}$ before each block. Let $t_0 = 0$, $t_{k+1} = t_k + h_k$, $\Psi_k = \Psi(Z_{t_k})$, and $T_B = \inf\{k : p(Z_{t_k}) \ge p_\star\}$. The production-step hitting time satisfies $T_\star \le t_{T_B} \le h_{\max} T_B$. Include optimizer buffers and sampling memory in the state, and absorb the block process after T_B.

Proposition 20 (Signed cumulative progress certificate)**.** *Suppose that before each production block a predictable bound ℓ_k satisfies*

$$\ell_k \le \mathbb{E}[\Psi_k - \Psi_{k+1} \mid \mathcal{F}_k], \qquad k < T_B,$$

and $\Psi_0 \le \Psi_0^+$, simultaneously with failure probability at most η_{audit}. The bounds may be negative. Put $C_n = \sum_{k<n} \ell_k$ on surviving paths and

$$b_n = B_0 \sqrt{\frac{n}{2} \log \frac{\pi^2 n^2}{3\eta_{\mathrm{noise}}}}.$$

Then

$$\mathbb{P}\{\exists n \ge 1,\ n < T_B :\ C_n > \Psi_0^+ - v_\star + b_n\} \le \eta_{\mathrm{audit}} + \eta_{\mathrm{noise}}.$$

Proof. For the absorbed process, write $g_k = \mathbb{E}[\Psi_k - \Psi_{k+1} \mid \mathcal{F}_k]$ and $M_n = \sum_{k<n}(\Psi_{k+1} - \mathbb{E}[\Psi_{k+1} \mid \mathcal{F}_k])$. Its martingale increments have conditional interval width at most B_0. The conditional Hoeffding bound gives $\mathbb{P}(|M_n| > b_n) \le 2\exp[-2b_n^2/(nB_0^2)] = 6\eta_{\mathrm{noise}}/(\pi^2 n^2)$. Summing over n bounds every fluctuation simultaneously by η_{noise}. On simultaneous audit coverage and before activation,

$$v_\star < \Psi_n = \Psi_0 - \sum_{k<n} g_k + M_n \le \Psi_0^+ - C_n + M_n.$$

Crossing the stated threshold therefore implies an audit failure or a martingale fluctuation. Their union proves the claim without conditioning the martingale law on the coverage event. □

Paying explicitly for adverse blocks. Choose a deterministic initial bound $U_0 \ge \Psi_0$, for example $U_0 = B_0$. Suppose audited good blocks have $\ell_k \ge a > 0$ and adverse blocks have $\ell_k \ge -d$. If a separate valid bound ensures $\mathbb{P}(T_B > n, N_{\mathrm{bad}}(n) > q_n) \le \eta_{\mathrm{occ}}$, then

$$an - (a+d)q_n > U_0 - v_\star + b_n \quad \Longrightarrow \quad \mathbb{P}(T_\star > h_{\max} n) \le \eta_{\mathrm{audit}} + \eta_{\mathrm{noise}} + \eta_{\mathrm{occ}}.$$

This follows because the surviving cumulative lower drift is at least $an - (a+d)N_{\mathrm{bad}}(n)$. The occupation bound is substantive information, which must come from a justified model, independent replicated continuations, or an executable certified policy. A small observed adverse count in one completed run supplies no such prospective guarantee. The result permits negative episodes and favorable multistep motion even when a uniform positive one-step drift fails.

Obtaining the conditional bounds. At each audited checkpoint, restore the policy, optimizer moments and step counter, scheduler, and sampler state. Run q independent copies of the declared next block, and evaluate their endpoint policies using fresh terminal verifications. Let $[\ell_{\mathrm{cur}}, u_{\mathrm{cur}}]$ and $[\ell_j, u_j]$ be simultaneous success intervals for the current and branch endpoint policies. Since Ψ decreases with success probability, a valid one-sided fixed-q lower bound is

$$\ell_k = \Psi(u_{\mathrm{cur}}) - \min\left\{B_0, \frac{1}{q}\sum_{j=1}^{q} \Psi(\ell_j) + B_0\sqrt{\frac{\log(1/\eta_k)}{2q}}\right\}.$$

Table 4: Verified controller operations and their interface with the feedback model. These are structural mappings, with no empirical ranking of the methods.

Method	Sampling intervention	Quantities requiring a new measurement
CBS	Select fresh rollouts and reuse a historical buffer	Generated versus selected activity, sample age, and the conditional update law
DUET	Allocate rollout counts and marker-gated stopping with reweighting	Joint stopping/event law, update moments, and generated-token cost
TRACE	Allocate root rollouts and prefix continuations in a tree	Prefix-conditional contrast, shared-prefix dependence, and full tree cost
REFT	Stratify continuations across distinct first tokens	Joint group contrast and conditional updates at fixed rollout count

Hoeffding's inequality bounds the conditional mean endpoint potential, and the evaluation intervals bound the latent observations. Production then uses independent future randomness under the same block policy. Allocate η_k and evaluation errors simultaneously across audits. If block length or optimizer settings are selected using these branches, use simultaneous bounds over the candidate set or a fresh audit after selection. Branch and evaluation costs belong in the total resource account. The certificate can be inconclusive, but requires no bound on an unseen exponential tail and makes no scalar closure assumption for Adam.

G.11 How sampling controllers enter the framework

The same quantities clarify what sampling interventions change. Let a controller choose action a from the full state z, and let $\mu_{z,a}$ be the joint binary-outcome law of its selected group. For a nonempty group, its mixed-outcome probability is

$$\lambda(z) = \mathbb{E}_{a|z}[1 - \mu_{z,a}(\text{all zero}) - \mu_{z,a}(\text{all one})].$$

The conditional kernels include the ensuing policy, optimizer, replay-buffer, and controller changes. The cost counts every generated token, including discarded or aborted responses. This description separates three effects: how often useful feedback arrives, what an update does after it arrives, and what obtaining it costs.

The distinction is especially important when selection occurs after generation, as in CBS: selected-group activity can increase while fresh-group contrast is unchanged (Lu et al., 2026). DUET additionally changes rollout counts and stopping, so its cost and event law must be measured jointly. Its unbiasedness argument concerns the stated action-independent-baseline surrogate (Hu et al., 2026). TRACE obtains contrast through root outcomes and prefix reward flips (Zou et al., 2026), whereas REFT changes the group law by balancing distinct first tokens (Kim & No, 2026). These operations can affect future progress, but their effects on the sign-transition parameter require a separate measurement of actual updates. Within-group dependence and cross-step persistence describe different random objects.

For REFT-type stratification, conditional on selected openers $\mathcal{S}$ and m independent continuations per opener with success probabilities p_f, contrast has probability

$$1 - \prod_{f \in \mathcal{S}} p_f^m - \prod_{f \in \mathcal{S}} (1 - p_f)^m.$$

Substitution of the average p_f into a binomial formula discards the group design. For a TRACE-type prefix with factual terminal reward $r \in \{0, 1\}$, k fresh continuations with success probability p_h flip that outcome with probability $1 - [rp_h + (1 - r)(1 - p_h)]^k$. Obtaining the factual path has already incurred cost. Restarting an accounting clock at that prefix leaves this cost outside the displayed conditional probability.

Even preserving the mean update leaves this distinction relevant. To see it directly, retain a scalar increment d with probability s and apply d/s when retained. The mean remains d, while its negative exponential moment becomes $1 - s + se^{-\alpha d/s} \geq e^{-\alpha d}$. For $d > 0$, the smaller active-only moment is offset by the rarer event, since $s(1 - e^{-\alpha d/s}) \leq 1 - e^{-\alpha d}$. This calculation illustrates convexity in a scalar update model. The measured full-state drift and cost remain the appropriate comparison for adaptive neural training.

Changes in the cost regime. Theorem 17 gives a precise way to distinguish a speedup from a change of regime. Suppose a controller preserves the embedded update law and, throughout the reachable pre-target region, its mean cost $\widetilde{c}$ and hazard $\widetilde{\lambda}$ satisfy

$$C^{-1}\frac{c(p)}{\lambda(p)} \le \frac{\widetilde{c}(p)}{\widetilde{\lambda}(p)} \le C\frac{c(p)}{\lambda(p)} \quad \text{for some } 1 \le C < \infty.$$

The two occupation-cost series are then comparable term by term, so their means are finite or infinite together. A bounded improvement in activity alone can therefore shorten waits without changing this classification. Crossing the threshold requires an unbounded relative change in cost per feedback event somewhere in the reachable region, or a change in the embedded update process. In our rare-feedback models, the relevant excursions visit states of vanishing success probability. Reducing persistent regressions is therefore one possible route, provided it is established from actual post-optimizer progress.

This conclusion also explains why an inferred change in α needs a resource convention. Corollary 18 exhibits allocation that improves the group-attempt regime while leaving generated-trajectory cost infinite. In contrast, a controller that changes verified trajectory quality can change both the conditional increment law and the excursion frequency. Its evaluation should consequently compare the full cost-and-progress calculation at the same target, including replay corrections and any inactive optimizer motion.

Finite-initialization regularization guidance. Curvature matching sets a common local scale, while the full force on the visited probability range determines later motion. At fixed initialization, compare the exact discrete penalty derivative and measure progress on both active and inactive steps. Select coefficients using a common declared finite-budget target and retain overshoot and truncation conventions. The fixed-accuracy experiment below reports every divergence and target setting together. The limiting force ordering is a hypothesis for controlled finite-range tests, with no universal optimizer-independent ranking. Appendix E.4 supplies an exact discrete sufficient condition, which can be checked on the stipulated interval without taking an infinite-level limit.

H Experimental Protocol and Complete Results

The experiments separate three questions raised by the analysis. First, do the finite processes follow the predicted waiting-cost laws? Second, do equal-data interventions change cost after a trainable representation and adaptive optimizer are introduced? Third, can verifier outcomes predict feedback availability on language-model outputs?

H.1 Exact finite-group simulation

The main activation grid consists of $K = 4$, $p_\star = 0.2$, three initial success probabilities, and three singleton jumps. We set $\eta = \delta K/\sqrt{K-1}$ and run 40,000 independent replicates per condition. For each state, we draw the exact geometric waiting time at rate $1 - p^K - (1-p)^K$, then draw N from its exact conditional binomial distribution on $\{1, \dots, K-1\}$. This skips inactive computations while preserving their full distribution of elapsed steps.

For independent batches, the number of active groups is binomial with parameters B and $\lambda(p)$, conditioned on being positive. We draw each active group's conditional success count and sum its normalized contribution, dividing by B. The waiting rate is $1 - (1 - \lambda(p))^B$. This procedure samples the exact discrete process, including multi-active batches and multi-success groups. The mean-recursion comparator sums the complete binomial update law at every step. It never substitutes an infinite-group objective.

Limiting series draws truncate after $\lceil 25/\delta \rceil$ terms for W_δ, or after $\lceil 25B/\delta \rceil$ terms for $B^{-1}W_{\delta/B}$. Their omitted expected tails are at most $e^{-25}/(1-e^{-\delta})$ and $e^{-25}/[B(1-e^{-\delta/B})]$, respectively. All nine activation conditions and all three batching conditions are included. At 40,000 observations, the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) gives a distribution-free 95% uniform band of half-width $\sqrt{\log(40)/(2 \cdot 40000)} \approx 0.0068$ for one empirical distribution function. The simulation and limiting-law samples are independent, so their comparison also contains Monte Carlo variation from the latter.

Table 5: All activation conditions. Times are normalized by $K\varepsilon$. The discrepancy is the two-sample Kolmogorov distance to 80,000 independent draws from the limiting series.

δ	ε	Mean	Coefficient of variation	Mean-recursion time	Discrepancy
0.1	10^{-2}	10.752	0.218	10.320	0.0497
0.1	10^{-3}	10.573	0.221	10.088	0.0160
0.1	10^{-4}	10.501	0.223	10.015	0.0036
0.5	10^{-2}	2.609	0.480	2.120	0.0381
0.5	10^{-3}	2.548	0.490	2.028	0.0097
0.5	10^{-4}	2.539	0.493	2.004	0.0029
1.0	10^{-2}	1.635	0.656	1.080	0.0418
1.0	10^{-3}	1.594	0.670	1.020	0.0072
1.0	10^{-4}	1.581	0.679	1.003	0.0066

H.2 Regularization and numerical resolution

We use forward Kullback–Leibler, reverse Kullback–Leibler, four times Jensen–Shannon, and twice squared Hellinger. Exact discrete runs use $K = 4$, $\delta = 0.7$, $\kappa = 0.3$, $\eta = 0.7K/\sqrt{K-1}$, and $\gamma = K\kappa/\eta$. For each family, 3,000 runs at each of $\varepsilon = 10^{-3}$ and 3×10^{-4} are followed to normalized time 30 or relative log-odds four. Every step samples a fresh binomial group and applies the exact gradient (44). Runs that have yet to cross the threshold at time 30 are recorded as censored, retaining their contribution to the survival distribution.

The limiting process is simulated in operational time using fourth-order Runge–Kutta drift integration, Poisson counts on each operational interval, and trapezoidal integration of physical time. This is a discretized approximation, distinct from the exact discrete-group simulation. We use operational mesh widths 0.02 and 0.01, each with 3,000 independent replicates, and an operational cutoff of 300. Every replicate either reaches the target or the physical-time horizon before that cutoff. Comparison of both meshes is retained in the numerical results, including all divergence families.

Table 6: Regularized distribution checks. Columns two and three compare the discrete hitting-time distribution to the limiting simulation with operational mesh 0.01. The mesh column compares the two limiting simulations. Medians use $\varepsilon = 3 \times 10^{-4}$ for the discrete process. Distances are $\sup_{0 \le t \le 30} |\widehat{F}_1(t) - \widehat{F}_2(t)|$, with all uncrossed runs retained as censored.

Penalty	Distance $\varepsilon = 10^{-3}$	Distance $\varepsilon = 3 \times 10^{-4}$	Mesh distance	Median	Limit median
Forward KL	0.0553	0.0303	0.0230	5.176	5.024
Jensen–Shannon	0.0220	0.0183	0.0210	2.519	2.525
Reverse KL	0.0263	0.0207	0.0267	2.253	2.193
Hellinger	0.0193	0.0167	0.0230	2.657	2.665

The resolution comparison includes Monte Carlo variation because its two cohorts are independent. The displayed distances therefore serve as numerical diagnostics, without being asserted as deterministic discretization-error bounds. No finite simulation is used to establish the infinite-level classification of Theorem 14.

Paired event-aligned resolution check. To separate numerical resolution from cohort variation, we draw 3,000 new operational-time Poisson paths per penalty and share each exact exponential interarrival time between meshes 0.02 and 0.01. We jointly integrate the state and physical clock by fourth-order Runge–Kutta, shortening the final step to each actual jump time. A physical-time censor crossing is located before any subsequent jump. The largest paired difference in capped hitting time is 1.09×10^{-9}, with zero disagreements in hit/censor status or hitting-event index across all 12,000 pairs. These are pathwise numerical measurements for the sampled cohort, separate from a uniform error theorem. The new fine-mesh medians and discrete-to-reference distances are given in Table 7.

Table 7: Event-aligned limiting simulation, with 3,000 paired paths per penalty and horizon 30. The maximum difference compares capped hitting times on identical Poisson paths. The last two columns compare independent discrete and event-aligned cohorts using the preceding supremum metric.

Penalty	Median	Censored	Maximum paired mesh difference	$d_{10^{-3}}$	$d_{3\times10^{-4}}$
Forward KL	5.034	7	1.09×10^{-9}	0.0513	0.0307
Jensen–Shannon	2.459	0	1.26×10^{-11}	0.0210	0.0190
Reverse KL	2.214	0	4.62×10^{-12}	0.0190	0.0147
Hellinger	2.616	0	3.38×10^{-11}	0.0177	0.0187

We also retain a paired check of the earlier bin-end jump convention. Sharing Poisson counts removes cohort noise but still moves the jump location when the mesh changes. Its largest paired hitting-time differences are 12.112, 0.0869, 0.1965, and 1.1016 for forward Kullback–Leibler, Jensen–Shannon, reverse Kullback–Leibler, and Hellinger. The event-aligned experiment removes this particular source of threshold sensitivity. Across independent old and new fine-mesh cohorts, distributional distances range from 0.0153 to 0.0247, and all conservative 95% median-difference intervals contain zero. These comparisons retain the earlier diagnostics while distinguishing jump placement, integration error, and sampling variation.

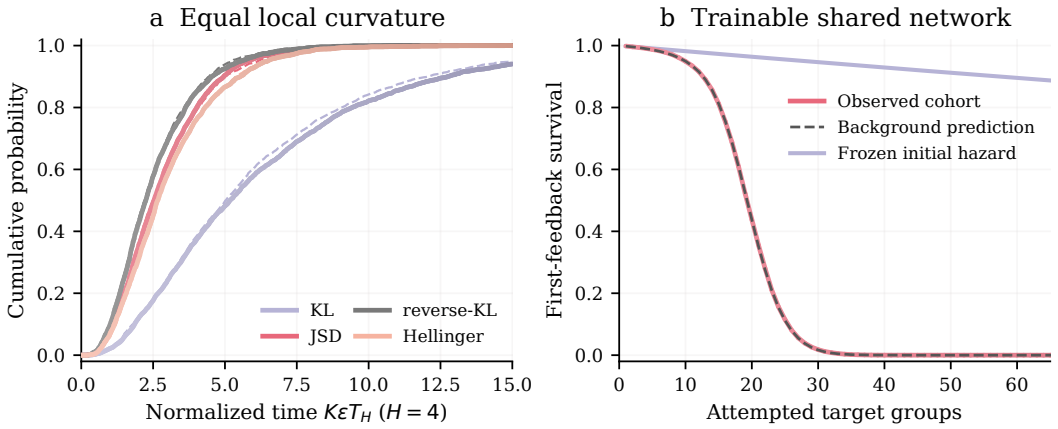


Figure 3: **Learning between selected feedback events.** Left: four curvature-matched regularizers, compared with their limiting predictions at relative log-odds level four. Right: first-feedback survival in a fully trainable shared network. An independent background-training cohort predicts the observation cohort. The initial-hazard curve fixes the initial state throughout. Fixed-accuracy regularization tests at $p_\star = 0.05$ and 0.2 are reported separately in Appendix H for all four penalties.

H.3 A SHARED-NETWORK PREDICTION WITH INDEPENDENT COHORTS

To test the background kernel, we use a shared policy for an easy task and a harder target. The two-layer network $\ell(x) = v^\top \tanh(W^\top x + b) + b_o$ has six inputs, 12 hidden units, and a Bernoulli action at each token, with every parameter trainable. The easy task has one token with feature $(1, 1, 0, 0, 0, 0)$. The target has four token features $(1, 0, 1, 0, 0, 0)$, $(1, 0, 0, 1, 0, 0)$, $(1, 0, 0, 0, 1, 0)$, and $(1, 0, 0, 0, 0, 1)$. Its terminal verifier accepts exactly when all four actions equal one. Since logits depend on the features and shared parameters, the product of the four Bernoulli probabilities gives an exact target hazard throughout learning.

We initialize one fixed network with input weights of scale 0.18, output weights of scale 0.12, zero hidden bias, and output bias $\mathrm{logit}(0.15)$. All replicates share this initialization and differ in subsequent sampled groups. Every step samples an easy-task group of four and applies the exact normalized gradient with learning rate 0.15 to all network parameters. A selected target feedback event is a mixed-reward group of four independent target trajectories. Before this event, the target's centered reward contribution is zero, while easy-task training continues to move the shared network.

We form the prediction from 20,000 independent background-only training paths, computing the target hazard at each step and accumulating (2). A separate cohort of 20,000 runs then records first-feedback times from actual target rollouts over the same 200-step horizon. Because its outcomes enter only the evaluation, the comparison tests a genuine forecast. The maximum survival-curve difference is 0.00418, and both cohorts have median first-feedback time 20. Freezing the initial hazard instead gives maximum difference 0.93614. At step 50, frozen-hazard survival is 0.91216, compared with approximately 1.19×10^{-8} dynamically and zero observed survivors. The agreement concerns the full-state first-feedback identity under shared learning.

H.4 Signed and state-dependent feedback

The signed transition model uses feedback probability $\lambda(p) = p^\alpha$ and increments $+1$ and -1. The positive sign has conditional probability $q(p) = q + \beta_q p$, with $\beta_q = 0$ or 0.4. We vary $q \in \{0.6, 0.8, 0.95\}$ at $\alpha = 1$ and separately use $q = 0.9$ at $\alpha = 0.5$ and 1.5. Each of the seven transition laws is run at $\varepsilon = 10^{-2}, 10^{-3}, 10^{-4}$, with 40,000 observations, target accuracy 0.2, and normalized budget $\varepsilon^\alpha T \le 20$. Inactive transitions leave the state fixed, so inverse-transform geometric skipping is exact. An event that falls beyond the horizon remains censored before its update is applied.

Independent limiting cohorts contain 80,000 series draws, each truncated after 2,000 terms. A fractional-moment Markov bound certifies the probability that the omitted tail exceeds 10^{-4}, as recorded in the result file. The same independent limiting cohort is used for state-dependent laws with a common local q, α. The discrepancy below is measured over the fixed grid from zero to 20, with spacing 0.01. Thus it is a grid-based cumulative-distribution discrepancy, distinct from a supremum over all thresholds. No finite empirical average is used as an estimate of an infinite theoretical mean.

Table 8: Complete signed-update matrix. Here β_q is the slope of $q(p)$, F_5 and F_{20} are observed budget-success fractions, and d is the grid discrepancy from the limiting distribution. Every row has target accuracy 0.2.

q	α	β_q	ε	F_5	F_{20}	d
0.60	1.0	0.0	0.010	0.3784	0.6380	0.0811
0.60	1.0	0.0	10^{-3}	0.3272	0.5974	0.0329
0.60	1.0	0.0	10^{-4}	0.3154	0.5825	0.0182
0.80	1.0	0.0	0.010	0.7829	0.9621	0.0197
0.80	1.0	0.0	10^{-3}	0.7732	0.9607	0.0074
0.80	1.0	0.0	10^{-4}	0.7724	0.9606	0.0040
0.95	1.0	0.0	0.010	0.9544	0.9990	0.0054
0.95	1.0	0.0	10^{-3}	0.9529	0.9990	0.0039
0.95	1.0	0.0	10^{-4}	0.9530	0.9990	0.0043
0.90	0.5	0.0	0.010	0.8768	0.9988	0.1314
0.90	0.5	0.0	10^{-3}	0.8509	0.9987	0.0492
0.90	0.5	0.0	10^{-4}	0.8384	0.9983	0.0181
0.90	1.5	0.0	0.010	0.9066	0.9866	0.0059
0.90	1.5	0.0	10^{-3}	0.9088	0.9871	0.0041
0.90	1.5	0.0	10^{-4}	0.9083	0.9874	0.0064
0.60	1.0	0.4	0.010	0.4179	0.6683	0.1202
0.60	1.0	0.4	10^{-3}	0.3438	0.6095	0.0463
0.60	1.0	0.4	10^{-4}	0.3184	0.5851	0.0207
0.80	1.0	0.4	0.010	0.7991	0.9640	0.0456
0.80	1.0	0.4	10^{-3}	0.7780	0.9591	0.0100
0.80	1.0	0.4	10^{-4}	0.7759	0.9591	0.0048

H.5 Correlated signs at fixed initialization

We hold $q = 0.8$, $\delta = 1$, and target accuracy 0.2 fixed. The primary comparison varies $\rho \in \{0, 0.1, 0.2, 0.5\}$ at $\alpha = 1$. Two additional conditions keep $\rho = 0.2$ and change the feedback exponent to 0.5 or 1.5. All six laws are evaluated at $\varepsilon = 10^{-2}, 10^{-3}, 10^{-4}$, with 40,000 observations and normalized budget 10^4. The latent sign is initially sampled from its stationary distribution and changes only at feedback events. Each event uses a fresh indepen-

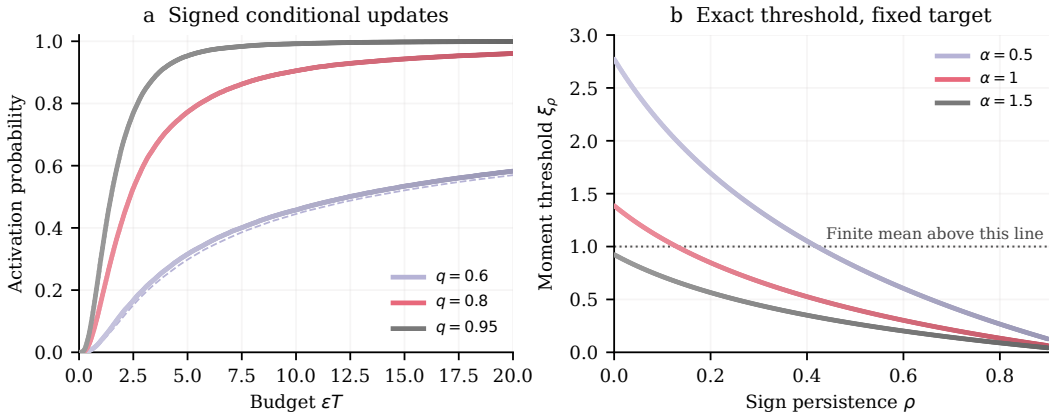


Figure 4: Signed feedback and the exact dependence threshold. Left: observed signed-update distributions at $\varepsilon = 10^{-4}$, with independent limiting predictions dashed. Right: the analytical threshold from Theorem 4, with $q = 0.8$ and $\delta = 1$. The right panel displays a proved formula and contains no fitted data.

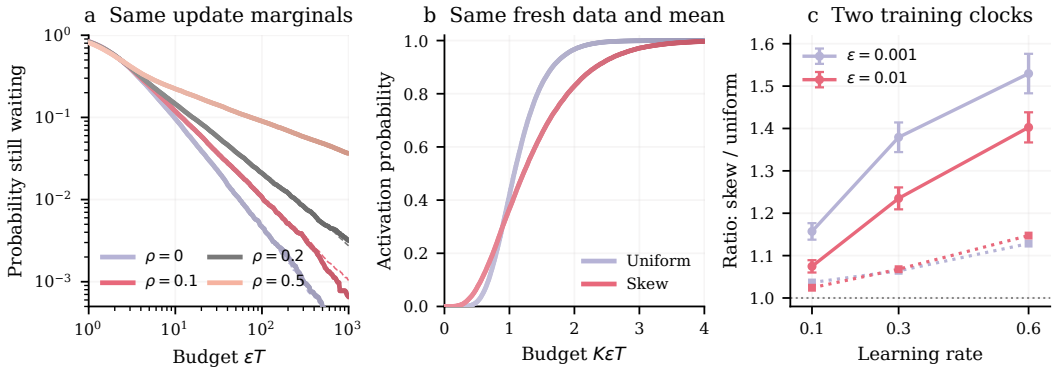


Figure 5: **Three ways that feedback cost escapes a mean-update description.** Left: increasing sign persistence changes activation survival while preserving the stationary one-update distribution ($q = 0.8$, $\varepsilon = 10^{-4}$, $p_\star = 0.2$). Dashed curves use a separate simulation cohort of the corresponding Markov-dependent limiting series. Middle: uniform and skewed positive weighting use the same 16 fresh trajectories per attempt and have identical conditional mean updates. Right: cost ratios (solid, approximate 95% Monte Carlo intervals) and active-event count ratios (dotted) in a fully trainable neural policy. Both clocks change, separating active-event counts from waiting costs.

dent draw from the transition row in (9). The same stationary sign marginal across conditions is an exact property of this construction at deterministic event indices.

The limiting cohorts contain 80,000 draws, with 1,600 terms and the matrix tail certificate in Appendix G. The certificate at omitted cost 10^{-4} is smaller than 10^{-36} in every condition. Distributional discrepancies use zero together with 2,500 logarithmically spaced thresholds from 0.01 to 10^4. The recorded restricted mean is the average of $\min\{\varepsilon^\alpha T_\star, 10^4\}$, retaining the horizon cost for censored runs. It is a finite-budget statistic for both moment regimes. We independently check the critical eigenvalue, the sign-chain Poisson equation, and the Bellman recursion for (60) to relative numerical tolerance 10^{-12}.

What a large finite cohort can establish. At $\varepsilon = 10^{-4}$ and $\rho = 0.1$, the exact normalized mean in (60) is 20.820, while the observed mean truncated at 10^4 is 9.248. The moment threshold is 1.075, close to the boundary of mean integrability. This disparity is consistent with a slowly decaying contribution from rare long waits. The paper therefore uses the exact formula for the full mean and the observed curve for finite-budget prediction. At $\rho = 0.2$ and 0.5, the mean is infinite by

Table 9: Complete dependence matrix. S_{100} is the fraction still waiting at normalized budget 100, C is the count censored at 10^4, d is the grid discrepancy, and RM_{10^4} averages $\min\{\varepsilon^\alpha T_\star, 10^4\}$ over every run.

ρ	α	ε	ξ_ρ	S_{100}	C	d	RM_{10^4}
0.0	1.0	0.010	1.386	0.0042	0	0.0154	5.655
0.0	1.0	10^{-3}	1.386	0.0042	0	0.0043	5.927
0.0	1.0	10^{-4}	1.386	0.0047	0	0.0041	5.961
0.1	1.0	0.010	1.075	0.0105	1	0.0207	9.642
0.1	1.0	10^{-3}	1.075	0.0118	4	0.0036	10.642
0.1	1.0	10^{-4}	1.075	0.0107	0	0.0039	9.248
0.2	1.0	0.010	0.847	0.0210	18	0.0241	21.361
0.2	1.0	10^{-3}	0.847	0.0211	12	0.0073	21.082
0.2	1.0	10^{-4}	0.847	0.0210	18	0.0045	21.995
0.5	1.0	0.010	0.405	0.0790	507	0.0501	211.114
0.5	1.0	10^{-3}	0.405	0.0837	516	0.0245	216.368
0.5	1.0	10^{-4}	0.405	0.0899	574	0.0108	238.198
0.2	0.5	0.010	1.695	0.0040	0	0.1465	6.446
0.2	0.5	10^{-3}	1.695	0.0039	0	0.0533	6.953
0.2	0.5	10^{-4}	1.695	0.0041	0	0.0185	7.220
0.2	1.5	0.010	0.565	0.0470	141	0.0117	77.521
0.2	1.5	10^{-3}	0.565	0.0474	145	0.0051	77.936
0.2	1.5	10^{-4}	0.565	0.0467	147	0.0039	82.933

proof, while the observed restricted means are 21.995 and 238.198. Censoring and the distinction between a restricted and an unrestricted mean are retained in all comparisons.

H.6 Identical finite-window histories

We implement the exact block construction in Theorem 5 for $L \in \{1, 2, 4, 8\}$, $\alpha = 1$, $\delta = 0.2$, $\varepsilon = 0.001$, and $p_\star = 0.2$. The feedback probability is $\lambda(p) = p$. Each law has 20,000 independent trajectories with uniformly sampled stationary hidden phases. During an inactive trial the logit and hidden phase remain fixed. Each geometric holding time is drawn by inverse transformation, followed by the signed update and the hidden-state transition. Runs are observed to normalized budget $\varepsilon T = 10^6$, with the cost at 10^5 read from the same paths. The smaller and larger budgets therefore form a nested observation scheme.

A sparse forward enumeration checks every reachable sign word through length $2L + 2$, including its support and exact stationary probability. All matching-window identities hold to error below 4.45×10^{-16}. We separately verify the first distinguishing word, both block and unrolled tilted matrices, and the side of one on which their spectral radii lie. Seven implementation tests check the construction, update order, geometric inversion, censoring, and unresolved quantiles. Finite-state exponential mixing holds for both laws. Its constants depend on L and ω, and are measured in event index. The construction makes no claim of equal mixing rates or exponential mixing in attempted-step time.

Table 10: Complete finite-window experiment. Mean regime is the analytical classification. C_b counts runs censored at budget b. Each RM_b is the empirical average of $\min\{\varepsilon T_\star, b\}$, including all 20,000 runs. Every displayed average has a finite cap.

L	Window	Mean regime	Median	C_{10^5}	C_{10^6}	RM_{10^5}	RM_{10^6}
1	3	Finite	15.382	0	0	21.6	21.6
1	3	Infinite	25.337	657	315	4975.5	23238.0
2	5	Finite	18.469	0	0	36.1	36.1
2	5	Infinite	32.948	788	360	5900.0	27390.7
4	9	Finite	24.465	0	0	95.2	95.2
4	9	Infinite	46.798	1375	745	9308.4	50432.2
8	17	Finite	47.567	15	0	878.9	921.9
8	17	Infinite	48.097	2856	1926	17758.9	116056.2

The last pair makes the distinction between typical completion and expected cost visible without using a rare-event estimate as proof. Its median costs differ by about 1.1%, while the fractions still waiting at 10^6 are zero and 0.0963. A zero observed survivor count is a finite-cohort observation. The finite-mean certificate for the fast law uses a convergent positive matrix series. The infinite-mean certificate for the slow law uses a lower bound over arbitrarily many adverse block repetitions. Both certificates are exact statements about the specified process and remain separate from the capped empirical costs.

H.7 An equal-fresh-data intervention

At each attempted step, both interventions use $B = 4$ independent groups with $K = 4$ independent fresh trajectories each. All four gradients are computed. Uniform weights are compared with a fresh independent random permutation of $(0.25, 0.25, 0.25, 3.25)$. The singleton jump is $\delta = 1$, the target is 0.2, and the three initializations match the activation grid. Each condition uses 40,000 observations and normalized horizon 20, with no observed censoring. The exact transition samples the number of active groups conditional on at least one being active, then each group's conditional binomial success count. This includes batches with several active groups. Exchangeability allows active positions to be enumerated first, with the heavy-weight position drawn uniformly and independently.

Each corresponding limiting cohort has 80,000 draws truncated after 500 terms. Every limiting increment is at least $0.25/4$, so the omitted expected cost is below 2×10^{-13}. Direct enumeration over all 5^4 success-count vectors and the four heavy-weight positions checks the mean equality at five policies. The maximum absolute numerical difference is 5.6×10^{-17}. This verifies the implementation of the conditional-mean identity separately from the activation-time experiment.

Table 11: All equal-fresh-data conditions. Means and coefficients of variation concern $K\varepsilon T_\star$. The last column is the grid discrepancy over the common budget interval.

Weighting	ε	Mean	Predicted mean	CV	Discrepancy
uniform	0.010	1.2038	1.1302	0.3280	0.0982
uniform	10^{-3}	1.1442	1.1302	0.3482	0.0187
uniform	10^{-4}	1.1325	1.1302	0.3516	0.0044
skew	0.010	1.4163	1.3550	0.5048	0.0505
skew	10^{-3}	1.3667	1.3550	0.5225	0.0100
skew	10^{-4}	1.3512	1.3550	0.5263	0.0049

H.8 Fully trainable neural activation

We next test whether the equal-data cost separation persists when the representation is trainable. The network has the same six inputs and 12 hidden units, and learns the four-token target until its true success probability reaches 0.2. Input weights have scale 0.18, output weights have scale 0.12, and hidden biases start at zero. We choose the shared output bias to give complete-solution probability exactly ε, while keeping the nonzero feature weights. Both interventions then start from this same network and update every parameter simultaneously by stochastic gradient descent.

Each attempted update uses four groups of four trajectories. Conditional on a group being informative, its success count is drawn from the exact conditional binomial law. Successful trajectories have four correct tokens, and failed trajectories are sampled from the product Bernoulli law conditioned on at least one error. Rejection sampling implements this conditioning exactly. The normalized score-gradient sum is calculated before any parameter is changed. A zero-contrast batch has zero gradient and no optimizer memory, so geometric skipping preserves the complete distribution of attempted-step costs. The experiment uses no clipping, momentum, replay, or adaptive optimizer. Those operations would require a separate conditional-mean calculation.

We vary $\varepsilon \in \{0.001, 0.01\}$ and learning rate in $\{0.1, 0.3, 0.6\}$, with 4,000 observations for each of the 12 conditions and normalized horizon 50. Every observed run reaches the target. Figure 5 shows ratios of independent cohort means. Its approximate intervals use the delta-method standard error $r\sqrt{s_s^2/(n\bar{T}_s^2) + s_u^2/(n\bar{T}_u^2)}$ with multiplier 1.96, where $r = \bar{T}_s/\bar{T}_u$. They describe Monte Carlo uncertainty for this controlled network experiment. The exact scalar activation law is compared only with scalar processes meeting its transition assumptions.

Table 12: Complete neural activation matrix. Each attempted step uses $B = 4$ groups of $K = 4$ fresh trajectories. Times are $K\varepsilon T_\star$, with target 0.2. CV is the empirical coefficient of variation, using the population denominator. Every network parameter is trainable.

ε	Step size	Weighting	Mean	Median	CV	Active events
10^{-3}	0.1	uniform	1.4855	1.4280	0.304	28.39
10^{-3}	0.1	skew	1.7189	1.6000	0.450	29.43
10^{-3}	0.3	uniform	0.6026	0.5480	0.496	10.24
10^{-3}	0.3	skew	0.8312	0.7240	0.651	10.90
10^{-3}	0.6	uniform	0.3838	0.3160	0.703	5.66
10^{-3}	0.6	skew	0.5871	0.5160	0.686	6.38
0.010	0.1	uniform	2.3179	2.2600	0.234	21.74
0.010	0.1	skew	2.4909	2.3600	0.369	22.26
0.010	0.3	uniform	0.8986	0.8400	0.384	7.98
0.010	0.3	skew	1.1098	1.0000	0.556	8.52
0.010	0.6	uniform	0.5404	0.4800	0.517	4.45
0.010	0.6	skew	0.7580	0.6800	0.633	5.11

Accounting for progress and waiting separately. For an embedded neural path $z_0, \dots, z_{N-1}$, the batch event probability is $\Lambda(z) = 1 - [1 - \lambda_K(p(z))]^B$. Holding states fixed and drawing only the waiting times gives

$$\mathbb{E}[T_\star \mid z_0, \dots, z_{N-1}] = \sum_{j<N} \Lambda(z_j)^{-1}, \qquad \mathbb{E}T_\star = \mathbb{E}N + \mathbb{E}\sum_{j<N} \frac{1 - \Lambda(z_j)}{\Lambda(z_j)}.$$

This conditional identity holds because inactive updates are exactly zero in the specified network. We replayed every original path without additional random draws and recovered all 48,000 original hitting times exactly. At each active state we also accumulated $\Lambda(z_j)^{-1}$, and compared it with the frozen-rate quantity $N/\Lambda(z_0)$. The latter preserves the same path and active count, while changing only its waiting schedule.

For cohort means, the exact descriptive identity is

$$\frac{\bar{C}_\mathrm{s}}{\bar{C}_\mathrm{u}} = \frac{\bar{N}_\mathrm{s}}{\bar{N}_\mathrm{u}} \frac{(\sum C_\mathrm{s})/(\sum N_\mathrm{s})}{(\sum C_\mathrm{u})/(\sum N_\mathrm{u})}.$$

The final factor is a ratio of pooled totals, distinct from an average of pathwise C/N. Table 13 reports all six comparisons. The intervention changes visited states as well as waiting, so this decomposition measures accounting terms, without assigning unique causal shares to optimization and feedback scarcity.

Equal expected parameter updates leave expected logit progress unidentified. For $h(w) = \operatorname{logit} p(w)$, a Taylor expansion includes $\frac{1}{2}\operatorname{tr}(\nabla^2 h(w)\mathbb{E}[\Delta w \Delta w^\top])$ in addition to $\nabla h(w)^\top \mathbb{E}\Delta w$. At a fixed state before conditioning on batch activity, independent group gradients yield a covariance multiplier $\sum_j w_j^2/B = 2.6875$ for the skewed update. Even scalar positive-increment walks with equal means can have different finite-target event counts through overshoot. Thus a count difference alone identifies neither nonconvex optimization nor pure holding-time convexity.

These measurements separate the number of active events from the cost of obtaining them. The neural ratios also increase with learning rate, consistent with stronger finite-update effects, although their state-dependent increments differ from the scalar model. The next experiment adds optimizer memory. Because momentum can move the policy after a zero-contrast batch, that comparison executes every step and follows the full evolving state instead of skipping inactive computations.

H.9 Adaptive-optimizer audit

To include that memory, we repeat the fresh-data intervention under SGD and Adam (Kingma & Ba, 2015) in the same fully trainable policy. Each attempted step generates four independent groups of four complete, four-token trajectories. All gradients receive either uniform weights or a fresh random permutation of $(0.25, 0.25, 0.25, 3.25)$. Crucially, homogeneous-reward batches also execute an optimizer step. Adam uses moment coefficients $(0.9, 0.999)$,

Table 13: Neural budget decomposition. Each ratio compares skewed with uniform weighting. Their product equals the observed cost ratio. The final pair gives mean path-conditional expected waiting in normalized budget units, without resampling policy updates.

ε	Step size	Cost ratio	Event ratio	Pooled-wait ratio	Expected wait U/S
10^{-3}	0.1	1.157	1.037	1.116	1.4870/1.7274
10^{-3}	0.3	1.379	1.064	1.296	0.5977/0.8302
10^{-3}	0.6	1.530	1.128	1.356	0.3833/0.5914
0.010	0.1	1.075	1.024	1.049	2.3109/2.5005
0.010	0.3	1.235	1.068	1.157	0.8961/1.1013
0.010	0.6	1.403	1.147	1.223	0.5452/0.7552

Table 14: Mean attempted steps to a common accuracy target. Ratios compare skewed to uniform weighting within the same optimizer, step size, and initialization. Standard errors concern independent training paths. The last column is the fraction of inactive attempts that change success log-odds under skewed weighting.

Optimizer	(ε, η)	Uniform	Skewed	Ratio	Inactive motion
SGD	$(.001, .3)$	152.85 ± 2.39	210.92 ± 4.03	1.380	0
SGD	$(.001, .6)$	98.31 ± 2.01	151.77 ± 3.10	1.544	0
Adam	$(.001, .003)$	226.19 ± 1.55	279.91 ± 2.23	1.237	73.8%
Adam	$(.001, .01)$	115.59 ± 1.61	124.12 ± 1.59	1.074	41.6%
SGD	$(.01, .3)$	22.24 ± 0.28	27.33 ± 0.50	1.229	0
SGD	$(.01, .6)$	13.44 ± 0.23	18.76 ± 0.38	1.396	0
Adam	$(.01, .003)$	85.34 ± 0.22	111.34 ± 0.45	1.305	91.2%
Adam	$(.01, .01)$	34.50 ± 0.21	40.89 ± 0.25	1.185	75.1%

stabilizer 10^{-8}, and bias correction at the actual attempted-step index, with zero weight decay. Both optimizers target $p_\star = 0.2$ within 10,000 attempts.

Each condition comprises 1,024 independent training paths. Every path reaches the target within the horizon. The gradients are checked against automatic differentiation, and the manual Adam ascent update is checked against the library implementation on a sequence containing a zero-gradient step. The latter check is essential because Adam's memory keeps moving parameters in such steps. Table 14 contains the entire two-initialization, two-step-size matrix for each optimizer. Multiplying attempted steps by 16 gives the complete fresh-trajectory cost.

Mean matching through the optimizer. Averaging over the random permutation preserves the raw gradient conditional on the fresh pool. SGD carries that equality through its linear update. Adam, however, transforms both moment estimates and their coordinatewise ratio, so the conditional means of its parameter steps can differ. Accordingly, the Adam comparison tests equal fresh data and equal mean raw gradients, with measured costs after the complete optimizer transformation. The scalar theorem supplies the controlled mechanism, while this experiment checks its behavior in the specified adaptive network.

The distinction is already visible on the first nonzero gradient. Suppose all earlier gradients vanish, the numerical stabilizer is zero, and that gradient is a positive scalar multiple ag. At attempted step n, the Adam displacement in a nonzero coordinate is

$$\eta \, \frac{1-\beta_1}{1-\beta_1^n} \sqrt{\frac{1-\beta_2^n}{1-\beta_2}} \, \operatorname{sign}(g).$$

The multiplier a cancels. Subsequent steps retain the accumulated moments, and a positive stabilizer further changes the expression. This elementary calculation explains why transferring a matched-mean scalar-jump theorem through Adam requires measuring the actual post-optimizer transition.

H.10 FIXED-ACCURACY DIVERGENCE COMPARISONS

The regularized limit compares fixed relative-odds levels. To examine the finite task directly, we instead fix accuracy at 0.05 or 0.2 and initialization at $\varepsilon = 0.003$ or 0.01, retaining the curvature normalization in Table 2. The exact discrete update uses $\delta = 0.7$, $\kappa = 0.1$, and $K = 4$. Each condition has 4,000 observations at normalized budget 20. We censor steps beyond that budget, including a final step created by rounding the horizon, so the comparison retains a common resource limit.

All 16 sampled cohorts reach their targets within the horizon. The empirical Jensen–Shannon and Hellinger medians are close and reverse the limiting order in one condition. Their restoring forces agree through second order at zero, while sampling variability and the finite discrete process require separate treatment. We infer no particular cause of the reversal from this table. The table retains every penalty and both budget-success and median statistics. The exact pathwise ordering remains a statement about the limiting process, separate from this fixed-accuracy experiment. At $\varepsilon = 0.01$, normalized time has lattice spacing $K\varepsilon = 0.04$, so equal medians such as 1.60 mean equal median step counts. They imply no equality of the full budget distributions.

Table 15: Complete fixed-accuracy matrix. F_5 is the activation fraction by normalized budget five. All rows have observed activation fraction one by budget 20. FKL and RKL denote forward and reverse Kullback–Leibler, JS denotes Jensen–Shannon, and H denotes Hellinger.

$\varepsilon = 0.003$				$\varepsilon = 0.01$			
$p_\star$	Penalty	Median	F_5	$p_\star$	Penalty	Median	F_5
0.05	FKL	1.932	0.9620	0.05	FKL	1.600	0.9768
0.05	JS	1.824	0.9705	0.05	JS	1.600	0.9820
0.05	RKL	1.764	0.9768	0.05	RKL	1.600	0.9735
0.05	H	1.860	0.9675	0.05	H	1.560	0.9755
0.20	FKL	2.148	0.9525	0.20	FKL	2.040	0.9625
0.20	JS	1.968	0.9647	0.20	JS	1.960	0.9620
0.20	RKL	1.872	0.9663	0.20	RKL	1.920	0.9653
0.20	H	1.944	0.9670	0.20	H	1.960	0.9635

H.11 PROSPECTIVE FEEDBACK AUDITS ON LANGUAGE-MODEL CHECKPOINTS

Finally, we test feedback availability using language-model solutions. We audit Qwen2.5-3B-Instruct (Qwen et al., 2024) and a rank-eight low-rank adapter (Hu et al., 2021) from an existing GRPO-family MATH run (Hendrycks et al., 2021). That run contains 96 attempted groups, including 38 with mixed verified rewards, from 256 training questions. It selected eight trajectories from 20 candidates plus four pilot generations, using one optimization pass, learning rate 5×10^{-6}, generation cap 768, and AdamW with zero weight decay. Saved moment buffers document this optimizer state. We freeze both endpoints and collect fresh generations so that the audit tests the feedback forecast separately from the historical training procedure.

Before sampling, we fix 12 questions from the MATH test split, four each from difficulty levels three, four, and five. We exclude normalized exact matches with the declared fine-tuning questions. This establishes separation from that fine-tuning set, while base-model pretraining exposure remains outside the audit. Question selection uses no measured model outcome. Each checkpoint-question pair receives 64 forecast completions followed by 128 independent observation completions. Every forecast record is saved before its observation requests are submitted. The total design contains 4,608 complete generation requests, with all successes, failures, capped outputs, and verifier exceptions retained.

Both checkpoints use the same chat prompt, temperature one, top-p one, unrestricted top-k, and a 1,024-token completion cap. The prompt requests concise reasoning and a final boxed answer, with no reference solution in the input. Verification compares the extracted boxed expression with the MATH gold answer using symbolic equivalence. A length-capped output retains a correct completed answer if one was already produced. The estimand is success under this declared capped decoder. Four vLLM generation workers (Kwon et al., 2023) partition checkpoints and question IDs without changing the per-request sampling law.

Table 16: Complete checkpoint calibration audit. Activity is a percentage. Waiting is counted in groups and capped at eight opportunities. Raw and smooth columns use the separately saved forecast cohort. Observations use fresh completions. Censoring is retained in every mean.

		Activity		Restricted waiting			
Checkpoint	K	Forecast	Observed	Raw	Smooth	Observed	Censored
Base	8	58.83	59.90	2.627	2.342	2.958	4/24
Base	4	44.39	46.61	3.320	3.103	3.563	11/48
Tuned	8	55.61	56.77	3.111	2.709	2.625	4/24
Tuned	4	42.50	42.45	3.661	3.398	3.354	13/48

EVALUATION PROMPT

System: You are a careful competition mathematics solver.

User: {`MATH question`}

Give concise reasoning and put your final answer in \`boxed`{}.

Predictions and independent checks. For primary group size $K = 8$, compute $h_K(p) = 1 - p^K - (1-p)^K$ from the forecast success rate. We retain the raw plug-in forecast and a separately labeled smoothed forecast with $p = (s + 1/2)/65$. The latter is a point prediction, with no confidence claim. Observation draws form 16 disjoint groups and two disjoint eight-opportunity waiting streams per question and checkpoint. We report mixed-group activity, first-feedback waits censored at eight, and the restricted mean $\mathbb{E}\min(T_1, 8)$. The predeclared $K = 4$ sensitivity shares the same completions and forms four waiting streams. These two group sizes share data and are evaluated as related diagnostics.

Endpoint-decrease candidates are identified using the forecast cohort. Independent observation cohorts then test all 12 base/final pairs using one-sided Fisher exact tests and Holm adjustment over the complete family. Simultaneous difference intervals use exact Clopper–Pearson success intervals and a Bonferroni allocation across the 24 proportions. Every question pair remains in the results, including ties and increases. Since $h_K(p)$ is largest at $p = 1/2$, an accuracy decrease from a high-success state can increase contrast. The audit therefore uses the exact hazard function across the whole range.

The experiment measures prospective feedback under fixed weights. Intervening adaptive recovery, optimizer-specific causality, and a global language-model activation guarantee are distinct targets requiring training-path evidence. Its role is to test the checkpoint-level quantities needed by the practical measurement protocol, using real generated mathematical reasoning and independent prediction and observation samples.

Complete results. All 4,608 requests complete, with 3,087,316 generated tokens, 606 capped outputs, and zero verifier exceptions. Across both checkpoints and both predeclared group sizes, the raw forecast of mean mixed-group activity differs from its independent observation by 0.05–2.22 percentage points. Table 16 reports all four comparisons, both predeclared waiting forecasts, observed restricted waits, and censoring. The point forecasts of waiting are sensitive to boundary success counts. The smoothed estimates are reported alongside the raw estimates under their original definitions.

Five forecast pairs have a lower endpoint success estimate after fine-tuning. The independent observation cohorts resolve none of these as a decrease after the predeclared Holm correction, and every simultaneous difference interval contains zero (Table 17). The interpretation is therefore checkpoint-level feedback calibration with uncertainty-aware screening. These measurements supply observable quantities for the proposed diagnostics, while serial clustering and adaptive recovery require additional ordered training-path measurements. In particular, two endpoint estimates cannot identify the persistence parameter in Theorem 4.

Table 17: All 12 preselected MATH question pairs. Forecast columns count successes out of 64, and observation columns out of 128. Differences and simultaneous 95% intervals are in percentage points. IDs index the fixed MATH test split. Every Holm-adjusted one-sided decrease-test p-value is one.

	Forecast		Observation			
ID	Base	Tuned	Base	Tuned	Difference	Simultaneous interval
00323	62	63	122	127	3.91	$[-6.1, 13.4]$
00385	36	26	49	62	10.16	$[-17.4, 36.7]$
00600	41	40	89	93	3.13	$[-22.0, 27.9]$
00702	63	64	128	126	-1.56	$[-8.4, 5.2]$
01365	62	63	120	127	5.47	$[-5.4, 15.7]$
01495	10	8	24	23	-0.78	$[-22.3, 20.8]$
01818	32	35	62	66	3.13	$[-24.4, 30.3]$
02261	6	5	10	9	-0.78	$[-15.8, 14.3]$
02455	34	31	68	78	7.81	$[-19.7, 34.5]$
02482	8	9	11	10	-0.78	$[-16.4, 15.0]$
02836	0	0	0	0	0.00	$[-5.2, 5.2]$
03072	7	8	24	20	-3.13	$[-23.9, 18.0]$

I Ethics, Reproducibility, and AI Use

Ethics statement This work studies feedback and recovery dynamics in reinforcement learning with verifiable rewards, with the goal of improving our understanding of training efficiency and reliability. Our experiments use controlled synthetic processes, neural policies, and mathematical reasoning tasks, and do not involve human subjects or private or sensitive personal data. The resulting analysis may help make RLVR training more predictable and resource-efficient, but improvements in training dynamics do not by themselves establish the safety or reliability of the resulting models. In particular, automated verifiers can introduce systematic errors that affect the feedback used for learning. Applications should therefore evaluate verifier reliability and downstream model behavior separately, while respecting the usage conditions of the underlying models and datasets.

Reproducibility statement The main text defines the training state, feedback event, activation target, and resource conventions used throughout the analysis. Appendices B–G provide the assumptions, derivations, and complete proofs for the feedback-arrival laws, activation-time results, finite-group and matched-mean analyses, regularization results, full-state extensions, and dependence thresholds. Appendix H specifies the complete experimental protocols and reports the full results for the controlled simulations, trainable neural policies, adaptive-optimizer experiments, fixed-accuracy comparisons, and language-model checkpoint audits, including initialization, optimization, normalization, censoring, sampling, and numerical-resolution conventions. The supplementary source package contains the manuscript source, simulation and plotting programs, and numerical outputs used to produce the reported figures and tables, allowing each reported experimental result to be traced to its corresponding program and result record.

AI use statement In accordance with the ICLR 2027 policy on the use of large language models, we disclose the following. For the manuscript, we used LLM assistants together with automated compiler diagnostics to help debug LaTeX formatting, to suggest grammatical corrections at the word and sentence level, and to check cross-references and the internal consistency of terminology, notation, and reported numbers; these tools produced suggestions only, and every change was reviewed and adopted or rejected by the authors. In the code released as supplementary material, we used Codex to assist with code generation and verification, including but not limited to dataset formalization, data cleaning and reformatting code, generating a subset of the plotting figures from experimental data we had already collected, and producing scripts for batched execution of experiment runs; no experimental result reported in this paper was produced by an AI assistant without author review of the underlying code and outputs. We also used an AI assistant to help condense and tighten the presentation of the analysis and proofs in Appendix, which we had derived beforehand, and the condensed statements and derivations were subsequently checked independently by three of the authors. We have reviewed all AI-assisted work, and the authors take full responsibility for the final content of this paper, including its text, claims, proofs, figures, and experimental results.