FINITE-TIME REGRET BOUNDS FOR REINFORCE IN TABULAR MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish finite-time expected cumulative regret bounds for REINFORCE in finite-horizon tabular MDPs with unknown transitions and bounded deterministic rewards. The central challenge is that policies at earlier stages determine which states are visited at later stages, while inaccurate policies at later stages can make optimal actions at earlier stages appear unfavorable. We tackle this challenge through two contributions. First, we bound the value errors that can reduce the probabilities of optimal actions by the regret incurred at later stages. This bound allows a geometrically weighted potential to control cumulative regret. Second, we show that reducing an action's probability requires cumulative use of that action, yielding quantitative bounds on state visitation along the mean flow. Together with stochastic tracking and probability control, these contributions yield $O_{\mathcal{M}}(\log^2 K)$ expected regret from zero initialization, using one trajectory per update and no explicit exploration or regularization. The step size is $\eta = \bar{\eta}/\log K$, with an instance-dependent scale $\bar{\eta}$, and remains constant within each run. A step size chosen using only the episode budget yields $O_{\mathcal{M}}(\log^3 K)$ regret. Finally, deterministic two-action chains exhibit linear expected regret over an initial window whose length grows as an exponential tower in the horizon, under a sufficiently small uniform step-size cap.

1 INTRODUCTION

REINFORCE and its variants (Williams, 1992; Sutton et al., 1999) have been applied to problems ranging from neural combinatorial optimization to language-model alignment (Bello et al., 2017; Ahmadian et al., 2024). In its basic form, REINFORCE couples exploration and exploitation through a single stochastic policy: updates aimed at increasing expected return also determine how future experience is sampled. Theory shows that these updates can sustain implicit exploration (Mei et al., 2023; Robertson et al., 2025). Understanding when this implicit exploration is sufficient, and how much reward is lost during learning, can clarify the capabilities of the basic algorithm and inform the design of exploration mechanisms for policy-gradient methods.

In tabular Markov decision processes (MDPs), Robertson et al. (2025) established asymptotic convergence for REINFORCE with any positive constant learning rate. Under these dynamics, almost surely, the algorithm visits all reachable state–action pairs infinitely often and its value converges to the optimal value.[1] To quantify the cost of learning, we consider expected regret from initialization. This criterion includes the reward lost before the policy becomes accurate. We therefore ask:

> *What finite-time regret bounds can we establish for REINFORCE in tabular MDPs? How does its expected regret depend on the episode budget and the MDP horizon?*

Challenge. The central challenge is to quantify learning progress and cumulative loss when policies at earlier stages determine visitation at later stages, while the downstream policy determines the learning signals received upstream. Backward analysis must control the cumulative regret incurred while

[1] Robertson et al. (2025, Appendix E, Theorem E.3) state an eventual average suboptimality rate $O(\log T/T)$ after a pathwise entry time. An expected regret bound from initialization additionally requires control of the expected transient loss and the expectations of the random rate coefficients.

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.

the downstream policy improves. Forward analysis must bound the time required to collect enough learning opportunities at downstream states. We address this challenge through two contributions:

Contribution 1: Bellman Debt Bounded by Downstream Regret. We establish that the aggregate downstream value error deducting from the drift of optimal logits, the *Bellman debt*, is bounded by the sum of downstream stage regrets under the same current policy. This bound applies to every downstream policy and precedes the argument establishing the probability floor. By assigning geometric weights to stages, we compensate for these upstream deductions and construct a linear potential. Under a uniform positive floor on the total probability of all optimal actions at every state and stage, the conditional expected increase of this potential controls per-episode regret.

Contribution 2: Cumulative Action Use Constrains Probability Decay. We show that, along the mean flow from zero initialization, reducing an action's probability below a window-dependent baseline requires cumulative use of that same action. We quantify this constraint through a logarithmic lower bound on cumulative action use in terms of cumulative occupancy of the parent state. The resulting bounds propagate along reachable paths, guaranteeing any prescribed cumulative occupancy in finite instance-dependent time without requiring correct Q-value ordering in advance.

We show that REINFORCE initialized with zero logits achieves an expected regret bound of order $\log^2 K$ in finite-horizon tabular MDPs. The environment has H stages with disjoint state sets, unknown transitions, and deterministic rewards in $[0, 1]$ observed on visited state–action pairs. The regret Reg_K measures the cumulative shortfall in policy value relative to an optimal policy across episodes $k = 1, \ldots, K$. For every fixed instance $\mathcal{M}$ with at least one reachable suboptimal action, there exist instance-dependent constants $\bar{\eta} > 0$ and $C_{\mathcal{M}} < \infty$ independent of K. For each announced budget $K \geq 2$, selecting the step size $\eta = \bar{\eta}/\log K$ and holding it constant throughout the run yields

$$\mathbb{E}[\mathrm{Reg}_K] \leq C_{\mathcal{M}} \log^2 K + 2H = O_{\mathcal{M}}(\log^2 K). \tag{1}$$

There exists a choice of step size that depends only on the budget and remains constant during each run, under which REINFORCE achieves expected regret $O_{\mathcal{M}}(\log^3 K)$ (Corollary 3.3).

A constructed class of deterministic finite-horizon chains exhibits an initial window of linear expected regret for REINFORCE from zero initialization, with steps bounded by a sufficiently small uniform cap. Its length grows as an exponential tower in the horizon H (Section 5).

We present the problem setting in Section 2 and develop the main results and theoretical analysis in Sections 3–4.2. Appendices H and I discuss related work and limitations, respectively.

2 Setting and Algorithm

2.1 MDP and Notation

For $m \in \mathbb{N}_+$, let $[m] := \{1, \ldots, m\}$. We consider a finite-horizon tabular MDP $\mathcal{M} = (H, (\mathcal{S}_h)_{h=1}^H, \mathcal{A}, \rho, (P_h)_{h=1}^{H-1}, (r_h)_{h=1}^H)$ (Puterman, 1994; Sutton & Barto, 2018). State sets $\mathcal{S}_h$ are finite and pairwise disjoint across stages $h \in [H]$, and $\mathcal{A}$ is a finite action set with $A := |\mathcal{A}|$. With initial distribution $\rho \in \Delta(\mathcal{S}_1)$, transitions satisfy $P_h(\cdot \mid s, a) \in \Delta(\mathcal{S}_{h+1})$, and rewards $r_h(s, a) \in [0, 1]$ are deterministic. The learner knows H, the state and action sets, and episode budget K. rewards are observed only at visited state–action pairs.

For stages $h < j$, we call stage h upstream of stage j, and stage j downstream of stage h. Relative to stage h, the downstream policy is $(\pi_j)_{j=h+1}^H$. If $P_h(x \mid s, a) > 0$ for some action a, we call $s \in \mathcal{S}_h$ a parent state of $x \in \mathcal{S}_{h+1}$. We call x a child state of s along that transition.

A Markov policy is $\pi = (\pi_h)_{h=1}^H$ with $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$. For $\mathcal{B} \subseteq \mathcal{A}$, write $p(\mathcal{B}) := \sum_{a \in \mathcal{B}} p(a)$ for any $p \in \Delta(\mathcal{A})$. Let V_h^π, Q_h^π denote the state and action-value functions for the undiscounted return, with advantage $\mathrm{Adv}_h^\pi(s, a) := Q_h^\pi(s, a) - V_h^\pi(s)$ and expected return $J(\pi) := \mathbb{E}_{S_1 \sim \rho}[V_1^\pi(S_1)]$. We write $V_h^\star, Q_h^\star$ for optimal values, and define $\mathcal{A}_h^\star(s) := \arg\max_{a \in \mathcal{A}} Q_h^\star(s, a)$, fixing an optimal Markov policy $\pi^\star$. The state occupancy is $d_h^\pi(s) := \mathbb{P}_{\mathcal{M}, \pi}(S_h = s)$.

Algorithm 1 REINFORCE

Require: Announced episode budget K and step size $\eta > 0$, held constant throughout the run

Initialize $\theta_{1,h}(s,a) \leftarrow 0$ for every $h \in [H]$, $s \in \mathcal{S}_h$, and $a \in \mathcal{A}$
for $k = 1, \dots, K$ **do**
  Form the policy $\pi_k = \pi_{\theta_k}$
  Sample under π_k one trajectory $(S_{k,1}, A_{k,1}, \dots, S_{k,H}, A_{k,H})$
  Set $G_{k,H+1} \leftarrow 0$
  for $h = H, H-1, \dots, 1$ **do**
    $G_{k,h} \leftarrow r_h(S_{k,h}, A_{k,h}) + G_{k,h+1}$
  $\theta_{k+1} \leftarrow \theta_k$
  for $(h,a) \in [H] \times \mathcal{A}$ **do**
    $\theta_{k+1,h}(S_{k,h}, a) \leftarrow \theta_{k,h}(S_{k,h}, a) + \eta \widehat{g}_{k,h}(S_{k,h}, a)$

Let $\mathcal{S}_h^+ \subseteq \mathcal{S}_h$ denote the states reachable with positive probability along some root-to-state path. The reachability scale is

$$p_{\min} := \min\Big(\{\rho(s) : \rho(s) > 0\} \cup \{P_h(s' \mid s,a) : s \in \mathcal{S}_h^+,\, P_h(s' \mid s,a) > 0\}\Big) > 0,$$

omitting transitions if $H = 1$. Define $\mathcal{I}_+ := \{(h,s,a) : s \in \mathcal{S}_h^+,\, a \notin \mathcal{A}_h^\star(s)\}$. When $\mathcal{I}_+ \neq \varnothing$, the minimum suboptimality gap is $\Delta_{\min} := \min_{(h,s,a)\in\mathcal{I}_+}(V_h^\star(s) - Q_h^\star(s,a)) > 0$ (Simchowitz & Jamieson, 2019; Xu et al., 2021); otherwise every reachable action is optimal and regret vanishes identically. Once the step size η is fixed, the REINFORCE update does not use the instance parameters $\Delta_{\min}$ and $p_{\min}$. Certified tuning assumes a supplied conservative instance-dependent scale, whose construction depends on those parameters.

For any policy π, the policy pseudo-regret is $\mathcal{R}(\pi) := J(\pi^\star) - J(\pi)$. Executing policy π_k in episode k incurs pseudo-regret $\mathcal{R}_k := \mathcal{R}(\pi_k) \in [0, H]$, with cumulative pseudo-regret $\mathrm{Reg}_K := \sum_{k=1}^K \mathcal{R}_k$ (Shani et al., 2020; Cai et al., 2020; Dann et al., 2023). We measure performance by $\mathbb{E}[\mathrm{Reg}_K]$ and use regret to mean pseudo-regret throughout.

2.2 REINFORCE

Algorithm 1 presents REINFORCE (Williams, 1992). The learner initializes $\theta_1 = 0$, inducing uniform policies, and in each episode $k \in [K]$ forms the softmax policy $\pi_k = \pi_{\theta_k}$:

$$\pi_{k,h}(a \mid s) := \frac{\exp(\theta_{k,h}(s,a))}{\sum_{b\in\mathcal{A}} \exp(\theta_{k,h}(s,b))}.$$

The learner samples one trajectory $(S_{k,1}, A_{k,1}, \dots, S_{k,H}, A_{k,H})$ under π_k from $S_{k,1} \sim \rho$ (Section G). A backward pass computes return-to-go $G_{k,h} := \sum_{j=h}^H r_j(S_{k,j}, A_{k,j}) \in [0, H-h+1]$. Using the pre-update policy π_k and step size $\eta > 0$ held constant throughout the run, the learner forms the estimator and updates logits simultaneously across all stages:

$$\begin{aligned} \widehat{g}_{k,h}(s,a) &:= \mathbf{1}\{S_{k,h} = s\}\big(\mathbf{1}\{A_{k,h} = a\} - \pi_{k,h}(a \mid s)\big) G_{k,h}, \\ \theta_{k+1,h}(s,a) &= \theta_{k,h}(s,a) + \eta \widehat{g}_{k,h}(s,a). \end{aligned} \tag{2}$$

Mean Update and Mean Flow We define the mean update vector field $g(\theta)$ by

$$g_h(\theta; s, a) := d_h^{\pi_\theta}(s)\pi_{\theta,h}(a \mid s)\,\mathrm{Adv}_h^{\pi_\theta}(s,a).$$

Letting $\mathcal{F}_{k-1}$ denote the history of complete episodes before episode k, the estimator satisfies $\mathbb{E}[\widehat{g}_k \mid \mathcal{F}_{k-1}] = g(\theta_k)$ (Sutton et al., 1999). Its continuous-time counterpart is the mean flow ODE (Borkar, 2008; Kushner & Yin, 2003)

$$\dot{\bar{\theta}}(u) = g(\bar{\theta}(u)), \qquad \bar{\theta}(0) = 0, \qquad \bar{\pi}_u := \pi_{\bar{\theta}(u)}, \qquad u \geq 0. \tag{3}$$

For the mean flow, define the inverse coordinates and the minimum inverse coordinate among optimal actions by

$$\bar{X}_{h,s,a}(u) := e^{-\bar{\theta}_h(u;s,a)}, \qquad \bar{X}_{h,s}^{\star,\min}(u) := \min_{a^+ \in \mathcal{A}_h^\star(s)} \bar{X}_{h,s,a^+}(u).$$

Definition 2.1 (Readiness)**.** A reachable state–stage block (h, s) is *ready* at time u along the mean flow if every suboptimal action $b \notin \mathcal{A}_h^\star(s)$ satisfies

$$\bar{X}_{h,s,b}(u) - \bar{X}_{h,s}^{\star,\min}(u) \geq L_h^{\mathrm{sep}} + 2,$$

where L_h^{sep} is the separation threshold for stage h defined in Appendix D. Stage h is ready when all its reachable state–stage blocks are ready.

The mean flow characterizes state visitation and reaches readiness in finite instance-dependent time. Downstream readiness supplies the ordering of Q-values under the current policy needed to separate upstream inverse coordinates. Pre-readiness tracking and the post-readiness barrier then control the REINFORCE iterates (see Section 4.2).

3 Main Results

This section establishes regret bounds over an episode budget K for REINFORCE initialized from zero logits. We first state a general regret guarantee across certified step sizes, and then specialize to the largest certified step to obtain regret of order $\log^2 K$.

3.1 Regret Bound

Assume the MDP contains at least one reachable suboptimal action, so that the minimal positive Bellman gap $\Delta_{\min} \in (0, H]$ is well-defined. There exist fixed instance constants $\mu_0 \in (0, 1/A)$ and $\bar{\eta} \in (0, 1)$, depending only on the instance parameters $(H, A, (|\mathcal{S}_h|)_{h=1}^H, p_{\min}, \Delta_{\min})$. These constants are independent of both the episode budget K and the step size η. They serve as the floor on the total probability of optimal actions and the certified step-size scale, respectively, throughout the analysis; their exact constructions appear in Appendices B and F.

For an episode budget $K \geq 2$, the algorithm operates in the *certified step-size regime*

$$K \geq 2, \qquad 0 < \eta \leq \frac{\bar{\eta}}{\log K}. \tag{4}$$

To choose a step guaranteed to satisfy this regime, we assume that a sufficiently conservative instance-dependent scale is supplied. Its construction depends on suboptimality gaps and reachability parameters. The step size η is held constant throughout the execution of the algorithm, with the upper endpoint termed the *largest certified step*.

In this certified step-size regime, Section 4.2 establishes a uniform lower bound on the total probability mass assigned to optimal actions. With probability at least $1-1/K$, the total probability mass assigned to all optimal actions is at least μ_0 simultaneously across all episodes $k \leq K$, stages $h \in [H]$, and states $s \in \mathcal{S}_h$.

To control the minimum logit reached during learning, Appendix A constructs a logit localization threshold $L_K(\eta) > 0$. Across all certified step sizes, this threshold satisfies

$$0 < L_K(\eta) \leq 2 \log K + \log\left(2A \sum_{h=1}^{H} |\mathcal{S}_h|\right). \tag{5}$$

Theorem 3.1 (Regret bound)**.** *Fix a finite MDP satisfying the model and feedback conditions of Section 2, with $\mathcal{I}_+ \neq \varnothing$. Initialize $\theta_1 = 0$, and let $\mu_0 \in (0, 1/A)$ and $\bar{\eta} \in (0, 1)$ be the constants defined above. For every announced budget $K \geq 2$ and step size $0 < \eta \leq \bar{\eta}/\log K$, held constant throughout the run, the REINFORCE update (2) satisfies*

$$\mathbb{E}[\mathrm{Reg}_K] \leq \frac{A\big(L_K(\eta) + \eta H\big)}{\eta \mu_0} \sum_{h=1}^{H} |\mathcal{S}_h|(1 + \mu_0^{-1})^{h-1} + 2H. \tag{6}$$

Consequently, for every fixed instance, $\mathbb{E}[\mathrm{Reg}_K] = O_{\mathcal{M}}(\log K/\eta)$.

3.2 Instance Dependence and Certified Step Selection

The parameters μ_0 and $\bar{\eta}$ are determined by the finite readiness time of the mean flow, constructed via backward recursion across stages (Appendix D). This readiness supplies correct current-policy Q-value ordering and separation margins for pre-readiness tracking and the post-readiness barrier. Repeated composition of the occupancy bounds can make the readiness time and resulting instance coefficient very large, potentially tower-sized in H. Exact expansions are given in Appendix F.

Corollary 3.2 (Largest certified step)**.** *Under the assumptions of Theorem 3.1, for each budget $K \geq 2$, choose the largest certified step $\eta_K = \bar{\eta}/\log K$. There exists a positive finite constant $C_{\mathcal{M}}$, depending only on the fixed instance and independent of K. Holding this selected step size constant throughout the run yields the regret bound*

$$\mathbb{E}[\mathrm{Reg}_K] \leq C_{\mathcal{M}} \log^2 K + 2H = O_{\mathcal{M}}(\log^2 K). \tag{7}$$

Appendix A details the algebraic simplification leading to (7). The step size η_K depends on the episode budget K and is held constant within the run, providing a guarantee for that fixed budget.

Corollary 3.3 (Budget-based step selection)**.** *Fix an MDP satisfying the model and feedback conditions of Section 2 with $\mathcal{I}_+$ nonempty. Initialize $\theta_1 = 0$. For each announced episode budget $K \geq 2$, choose the step size $\eta_K = (1+\log K)^{-2}$, held constant within the run, and execute the REINFORCE update (2). For every fixed instance $\mathcal{M}$, there exists a deterministic finite constant $\tilde{C}_{\mathcal{M}}$, independent of K, such that for all integers $K \geq 2$,*

$$\mathbb{E}[\mathrm{Reg}_K] \leq \tilde{C}_{\mathcal{M}}(1+\log K)^3 = O_{\mathcal{M}}(\log^3 K). \tag{8}$$

Proof. Because $\eta_K \log K = \log K/(1+\log K)^2 \to 0$ as $K \to \infty$, there exists a deterministic integer $K_0 \geq 2$, depending on $\mathcal{M}$ through $\bar{\eta}$, such that $\eta_K \leq \bar{\eta}/\log K$ for all $K \geq K_0$. For any $K \geq K_0$, the certified regime (4) is satisfied. Applying Theorem 3.1 along with (5) gives $\mathbb{E}[\mathrm{Reg}_K] \leq C_{1,\mathcal{M}}(1+\log K)^3$. Here $C_{1,\mathcal{M}}$ is a finite instance-dependent constant independent of K. For $2 \leq K < K_0$, the deterministic bound $\mathrm{Reg}_K \leq HK \leq HK_0$ applies. Setting $\tilde{C}_{\mathcal{M}} := \max\{C_{1,\mathcal{M}}, HK_0\}$ yields (8) for all $K \geq 2$. □

This step selection uses the budget K alone. The certification threshold K_0 and the constant $\tilde{C}_{\mathcal{M}}$ remain instance-dependent.

4 Main Techniques

This section develops two techniques for analyzing regret in REINFORCE despite downstream value errors and upstream constraints on occupancy. First, to control stagewise regret in the presence of downstream value errors, we bound their adverse contribution to the drift of optimal logits by the aggregate downstream regret. We then construct a weighted linear potential to offset these deductions. Its conditional expected drift controls regret in each episode whenever the total probability of optimal actions has a uniform positive floor across all states and stages. Second, to establish this floor, we show how upstream actions supply sufficient cumulative occupancy along the mean flow: lowering an action probability below a window-dependent baseline requires cumulative use of that action. Combining this property with occupancy propagation and backward induction establishes the readiness condition in Definition 2.1 in finite time. Finally, tracking and a barrier argument transfer and preserve the probability floor for the REINFORCE iterates under the certified step-size regime; logit localization and stopped telescoping yield the regret bound.

4.1 A Linear Potential for Regret Control

For any action subset $B \subseteq \mathcal{A}$, let $\theta_h(s,B) := \sum_{a\in B} \theta_h(s,a)$ and $g_h(\theta;s,B) := \sum_{a\in B} g_h(\theta;s,a)$. As a motivating special case, consider a multi-armed bandit ($H=1$), with stage and state indices suppressed. Let $\mathcal{A}^\star$ denote the set of optimal actions. Define $\theta_k(\mathcal{A}^\star) := \sum_{o\in\mathcal{A}^\star} \theta_k(o)$ as their aggregate logit. The corresponding policy mass is $\pi_k(\mathcal{A}^\star) := \sum_{o\in\mathcal{A}^\star} \pi_k(o)$. In this notation, Lattimore (2026b, p. 3) gives

$$\mathbb{E}\left[\theta_{k+1}(\mathcal{A}^\star) - \theta_k(\mathcal{A}^\star) \mid \mathcal{F}_{k-1}\right] = \eta \pi_k(\mathcal{A}^\star)\mathcal{R}(\pi_k).$$

The left-hand side is the conditional expected policy gradient of the aggregate optimal logits, scaled by the step size, while the right-hand side contains the episode-k regret $\mathcal{R}(\pi_k)$ itself. Once $\pi_k(\mathcal{A}^\star)$ is controlled, the bandit regret is therefore recovered directly from the drift of optimal logits.

In MDPs ($H > 1$), however, an inaccurate downstream policy can give optimal actions negative advantages and decrease their mean logit updates at visited states. For $\pi = \pi_\theta$ and $o \in \mathcal{A}_h^\star(s)$, the mean update decomposes as

$$\begin{aligned} g_h(\theta; s, o) &= d_h^\pi(s)\pi_h(o \mid s)\,\mathrm{Adv}_h^\pi(s, o) \\ &= d_h^\pi(s)\pi_h(o \mid s)\Bigg(\underbrace{\ell_h^\pi(s)}_{\substack{\text{local action-selection} \\ \text{loss}}} + \underbrace{\sum_{a\in\mathcal{A}} \pi_h(a \mid s) P_h e_{h+1}^\pi(s, a)}_{\substack{\text{average propagated} \\ \text{error}}} - \underbrace{P_h e_{h+1}^\pi(s, o)}_{\substack{\text{downstream Bellman} \\ \text{error}}}\Bigg). \end{aligned} \tag{9}$$

Let $\Delta_h(s, a) = V_h^\star(s) - Q_h^\star(s, a) \geq 0$ denote the Bellman action gap. The local action-selection loss is $\ell_h^\pi(s) = \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\Delta_h(s, a)$. Let $e_h^\pi(s) = V_h^\star(s) - V_h^\pi(s) \geq 0$ denote the value error, with $e_{H+1}^\pi \equiv 0$. The transition operator acts on value errors via $P_h e(s, a) = \sum_{x\in\mathcal{S}_{h+1}} P_h(x \mid s, a)e(x)$. Weighting $\ell_h^\pi(s)$ by the state occupancy $d_h^\pi(s)$ yields the stage and total regret,

$$\mathcal{R}_h(\pi) = \sum_{s\in\mathcal{S}_h} d_h^\pi(s)\ell_h^\pi(s), \qquad \mathcal{R}(\pi) = \sum_{h=1}^H \mathcal{R}_h(\pi).$$

We aim to control $\mathcal{R}_h(\pi)$ through the mean updates $g_h(\theta; s, o)$ of optimal actions. In (9), the first two terms are nonnegative, but the final downstream error term can make the drift negative.

The proof of Robertson et al. (2025, proof of Theorem 5.2) waits until downstream value errors become sufficiently small relative to action gaps before establishing upstream action-value separation. However, the regret accumulated before then still needs control. Our key observation is that downstream stage regrets bound the adverse contribution of downstream value errors to the drift, enabling compensation across stages. We define this aggregate contribution as the downstream **Bellman debt**,

$$D_h(\pi) = \sum_{s\in\mathcal{S}_h} d_h^\pi(s) \sum_{o\in\mathcal{A}_h^\star(s)} \pi_h(o \mid s) P_h e_{h+1}^\pi(s, o).$$

This debt is bounded by

$$\begin{aligned} D_h(\pi) &\leq \sum_{s\in\mathcal{S}_h} d_h^\pi(s) \sum_{a\in\mathcal{A}} \pi_h(a \mid s) P_h e_{h+1}^\pi(s, a) \\ &= \sum_{x\in\mathcal{S}_{h+1}} d_{h+1}^\pi(x) e_{h+1}^\pi(x) = \sum_{j=h+1}^H \mathcal{R}_j(\pi). \end{aligned} \tag{10}$$

This bound follows from the nonnegativity of value errors ($e_{h+1}^\pi \geq 0$) and occupancy and Bellman recursions under the same policy. Assume a uniform floor on the total probability of optimal actions, $\pi_h(\mathcal{A}_h^\star(s) \mid s) \geq \mu_0$, for all states and stages, with $\mu_0 \in (0, 1/A)$. Summing the mean updates in (9) over $o \in \mathcal{A}_h^\star(s)$ and $s \in \mathcal{S}_h$, dropping the nonnegative average propagated error, and applying (10) yields the stagewise drift bound

$$\sum_{s\in\mathcal{S}_h} g_h(\theta; s, \mathcal{A}_h^\star(s)) \geq \mu_0 \mathcal{R}_h(\pi) - \sum_{j=h+1}^H \mathcal{R}_j(\pi).$$

To compensate for upstream deductions, we define the stage weights and linear potential as

$$w_h(\mu_0) = \left(1 + \mu_0^{-1}\right)^{h-1}, \quad h \in [H], \qquad Y_{\mu_0}(\theta) = \sum_{h=1}^H w_h(\mu_0) \sum_{s\in\mathcal{S}_h} \theta_h(s, \mathcal{A}_h^\star(s)).$$

This cancellation of cross-stage debt, named *Bellman compensation*, yields the lemma upon summing the weighted drift bounds over stages.

Lemma 4.1 (Drift of the Linear Potential). *Let $\mu_0 \in (0, 1/A)$. For any logit vector θ, if $\pi_{\theta,h}(\mathcal{A}_h^\star(s) \mid s) \geq \mu_0$ for all $h \in [H]$ and $s \in \mathcal{S}_h$, then*

$$\langle g(\theta), \nabla Y_{\mu_0} \rangle \geq \mu_0 \mathcal{R}(\pi_\theta).$$

The algebraic verification of Lemma 4.1 is detailed in Appendix A (specifically Lemmas A.7 and A.5). Consider stochastic policy gradient updates $\theta_{k+1} = \theta_k + \eta \widehat{g}_k$. Linearity of the potential gives the exact sample increment $Y_{\mu_0}(\theta_{k+1}) - Y_{\mu_0}(\theta_k) = \eta \langle \widehat{g}_k, \nabla Y_{\mu_0} \rangle$. Assuming conditional unbiasedness $\mathbb{E}[\widehat{g}_k \mid \mathcal{F}_{k-1}] = g(\theta_k)$, taking conditional expectations yields

$$\mathbb{E}\left[Y_{\mu_0}(\theta_{k+1}) - Y_{\mu_0}(\theta_k) \mid \mathcal{F}_{k-1}\right] \geq \eta \mu_0 \mathcal{R}_k \tag{11}$$

whenever π_k satisfies the floor on the total probability of optimal actions across all states and stages. The floor on the total probability of optimal actions ensures that the expected regret drift (11) holds for all episodes before a stopping time τ. In Appendix A, τ is defined as the first episode where either the probability floor or logit localization fails, capped at $K + 1$. Logit localization together with the zero-sum invariant bounds the endpoint difference $Y_{\mu_0}(\theta_\tau) - Y_{\mu_0}(\theta_1)$ (Lemma A.6). Stopped telescoping therefore controls the expected regret accumulated before τ, with the analysis of exceptional episodes deferred to Appendix A. The probability floor is established in the subsequent subsection.

4.2 Endogenous Visitation and Control of the Probability of Optimal Actions

The potential drift in (11) controls regret in each episode whenever the policy maintains a floor μ_0 on the total probability of optimal actions across all states and stages. Lemma 4.2 establishes this bound.

Lemma 4.2 (Probability floor for optimal actions). *Fix any finite MDP satisfying the conditions of Section 2 with $\mathcal{I}_+ \neq \varnothing$. Let $\mu_0 \in (0, 1/A)$ and $\bar{\eta} \in (0, 1)$ be the constants from Section 3. For every (K, η) in the certified step-size regime (4), the REINFORCE update (2) satisfies*

$$\mathbb{P}\left(\min_{1 \leq k \leq K} \min_{h \in [H],\, s \in \mathcal{S}_h} \pi_{k,h}(\mathcal{A}_h^\star(s) \mid s) \leq \mu_0 \right) \leq \frac{1}{K}.$$

We establish Lemma 4.2 by analyzing the deterministic mean flow (3), proceeding backward from stage H to 1. Stage readiness (Definition 2.1, Section 2) requires each suboptimal inverse coordinate to exceed the smallest optimal inverse coordinate by at least $L_h^{\mathrm{sep}}+2$, where L_h^{sep} controls continuation errors and $+2$ is an entry buffer for discrete tracking. When all downstream stages $j > h$ are ready, their L_j^{sep} thresholds bound continuation value errors and yield the current-policy gap $Q_h^{\bar{\pi}_u}(s, a^+) - Q_h^{\bar{\pi}_u}(s, b) \geq \Delta_{\min}/2$ for all optimal actions $a^+ \in \mathcal{A}_h^\star(s)$ and suboptimal actions $b \notin \mathcal{A}_h^\star(s)$ (Lemma D.3, Appendix D). At stage H, future value terms vanish identically, so $Q_H^{\bar{\pi}_u}(s, a^+) - Q_H^{\bar{\pi}_u}(s, b) \geq \Delta_H(s, b) \geq \Delta_{\min}$ holds immediately. In any ready block, at least one optimal action attains the maximal logit, guaranteeing $\bar{\pi}_{u,h}(\mathcal{A}_h^\star(s) \mid s) \geq 1/A$. The core obstruction is that unready upstream policies may temporarily throttle downstream visitation.

To show how state occupancy drives readiness at stage h, fix a reachable block (h, s), suppress block indices on inverse coordinates $\bar{X}_i(u)$, and define the local normalizer $\bar{Z}(u) := \sum_{i \in \mathcal{A}} e^{\bar{\theta}_i(u)}$. For any optimal action $a^+ \in \mathcal{A}_h^\star(s)$ and suboptimal action $b \notin \mathcal{A}_h^\star(s)$, we define the local clock χ. Provided all downstream stages remain ready over a time window beginning at $U \geq 0$, the pairwise coordinate derivative satisfies

$$d\chi(u) := \frac{d_h^{\bar{\pi}_u}(s)}{\bar{Z}(u)}\, du, \qquad \frac{d(\bar{X}_b - \bar{X}_{a^+})}{d\chi} = Q_h^{\bar{\pi}_u}(s, a^+) - Q_h^{\bar{\pi}_u}(s, b) \geq \frac{\Delta_{\min}}{2}.$$

For duration $t \geq 0$, let $\mathfrak{a}_{h,s}(U, t) := \int_U^{U+t} d_h^{\bar{\pi}_u}(s)\, du$ denote cumulative state occupancy, $C(U) := 1 + HU/A$ the coordinate bound at time U, and $q := \chi(U + t) - \chi(U)$ the elapsed clock increment. Bounding $|d\bar{X}_i/d\chi| \leq H$ and normalizer growth yields the occupancy–clock conversion

$$\mathfrak{a}_{h,s}(U, t) \leq \frac{(C(U) + Hq)^A - C(U)^A}{H}.$$

Because the initial difference satisfies $\bar{X}_b(U) - \bar{X}_{a^+}(U) \geq -C(U)$, reaching the clock increment $q \geq 2(C(U) + L_h^{\mathrm{sep}} + 2)/\Delta_{\min}$ guarantees the inequality $\bar{X}_b - \bar{X}_{a^+} \geq L_h^{\mathrm{sep}} + 2$. The local clock

thus converts sufficient cumulative occupancy into readiness. Our remaining deterministic task is guaranteeing sufficient cumulative state occupancy despite unready upstream stages.

Our key observation is that suppressing an action requires accumulating use of that same action. Along the mean flow, the probability factor bounds the magnitude of logit drift by $Hd_h^{\bar{\pi}_u}(s)\bar{\pi}_{u,h}(a \mid s)$. For any action $a \in \mathcal{A}$, we define cumulative state–action occupancy by $w_a(U,t) := \int_U^{U+t} d_h^{\bar{\pi}_u}(s)\bar{\pi}_{u,h}(a \mid s)\, du$. The action probability satisfies

$$\log \frac{1}{\bar{\pi}_{U+t,h}(a \mid s)} \le \log A + A \log C(U) + AH w_a(U,t). \tag{12}$$

Thus, a smaller endpoint probability increases the required lower bound on cumulative use $w_a(U,t)$ relative to the baseline $\log A + A \log C(U)$. This suppression bound holds for arbitrary orderings of Q-values under the current policy (Lemma D.4).

To derive (12) (Appendix D), let V and Q_i denote state and action values under the current policy, with $|V - Q_i| \le H$, and reparametrize time by cumulative action use $w(u) := w_a(U, u-U)$ with inverse $u(w)$. Differentiating inverse coordinates yields the exact cancellation

$$\frac{d\bar{X}_i}{dw} = \bar{X}_a(V - Q_i).$$

Integrating with $\bar{X}_i(U) \le C(U)$ bounds $\bar{X}_a(w) \le C(U)e^{Hw}$, which subsequently bounds all $\bar{X}_i(w) \le C(U)e^{Hw}$. The zero-sum logit invariant $\sum_i \bar{\theta}_i = 0$ guarantees $\prod_{i\in\mathcal{A}} \bar{X}_i = 1$, yielding

$$\frac{1}{\bar{\pi}_a(u(w))} = \bar{X}_a(w) \sum_{i\in\mathcal{A}} \prod_{c\neq i} \bar{X}_c(w) \le AC(U)^A e^{AHw}.$$

Integrating the inverse probability bound over action use and inverting the resulting inequality gives a lower bound on action use in terms of state occupancy (Appendix D). Combining this bound with the transition recursion $\mathfrak{a}_{h+1,x}(U,t) \ge P_h(x \mid s,a) w_a(U,t)$ for any child state $x \in \mathcal{S}_{h+1}$ with transition probability $P_h(x \mid s,a) > 0$ transmits cumulative occupancy along the edge $(s,a) \to x$:

$$w_a(U,t) \ge \frac{1}{AH} \log\left(1 + \frac{H\mathfrak{a}_{h,s}(U,t)}{C(U)^A}\right), \tag{13}$$

$$\mathfrak{a}_{h+1,x}(U,t) \ge \frac{P_h(x \mid s,a)}{AH} \log\left(1 + \frac{H\mathfrak{a}_{h,s}(U,t)}{(1+HU/A)^A}\right). \tag{14}$$

Because cumulative root occupancy grows as $\mathfrak{a}_{1,s_1}(U,t) = \rho(s_1)t$, composing (14) along positive-probability paths delivers every finite occupancy target to every reachable state in finite time (Lemma D.4 and Theorem D.5). Applying this guarantee backward from stage H to 1 closes the induction: once downstream stages are ready, transmitted occupancy drives $\bar{X}_b - \bar{X}_{a^+} \ge L_h^{\mathrm{sep}} + 2$ for all optimal $a^+ \in \mathcal{A}_h^\star(s)$ and suboptimal $b \notin \mathcal{A}_h^\star(s)$. Provided downstream readiness continues, positive pairwise derivatives preserve this inequality for all subsequent times. Thus, all reachable blocks become and remain ready by a finite instance-dependent time u_{ready} independent of the episode budget K (Theorem D.1).

We transfer these guarantees to the REINFORCE iterates. On the tracking good event (Lemma C.1, Appendix C), iterates track the mean flow over the pre-readiness window, maintaining the probability floor μ_0. At entry, the difference between each suboptimal inverse coordinate and the minimum optimal inverse coordinate is greater than $L_h^{\mathrm{sep}} + 1$. If the budget ends within this window, tracking alone covers all episodes. For larger budgets, a post-readiness discrete barrier keeps that difference at least L_h^{sep} on its good event (Lemma B.2, Appendix B), ensuring optimal-action probability at least $1/A > \mu_0$. In regime (4), a union bound limits total failure probability to $1/K$. Unreachable blocks remain uniform, while blocks in which all actions are optimal trivially satisfy the floor, establishing Lemma 4.2.

Proof sketch of Theorem 3.1. Let τ be the minimum of $K+1$ and the first episode k where a stopping failure occurs: either the total optimal-action probability is at most μ_0 for some block at $k \le K$, or any logit is at or below $-L_K(\eta)$ at $k \le K+1$ (Lemma A.6). Prior to τ, the valid probability

floor bounds instantaneous regret by the conditional drift of $Y_{\mu_0}(\theta_k)$ via (11). Because single-step updates change logits by at most ηH, logits at stopping satisfy the one-step overshoot bound $\theta_{\tau,h}(s,a) \geq -L_K(\eta) - \eta H$. Combined with the zero-sum invariant $\sum_{a\in\mathcal{A}} \theta_{\tau,h}(s,a) = 0$, this controls optimal logit sums and gives an upper bound on the terminal potential $Y_{\mu_0}(\theta_\tau)$. Telescoping increments yields an expected regret of $O_{\mathcal{M}}(\log K/\eta)$ before stopping. By Lemmas 4.2 and A.6, premature stopping occurs with probability at most $2/K$, so post-stopping episodes contribute at most $(2/K) \cdot K \cdot H = 2H$ in expectation (Appendix A). □

5 Lower Bounds and the Scope of the Regret Guarantee

This section quantifies the potentially long initial learning cost accompanying the regret guarantee in Theorem 3.1.

Li et al. (2023, Theorem 1 and Section 4.2) analyze exact-gradient softmax policy gradient in discounted MDPs under a uniform initial-state distribution. They establish slow convergence caused by suppressed optimal-action probabilities and recursively amplified delays before other states improve. Our finite-horizon chain, with each episode starting at its root, exhibits an analogous mechanism, where unattractive continue actions diminish visitation to downstream states. Theorem 5.1 shows that under trajectory sampling, this bottleneck induces an extended period of linear expected regret.

Theorem 5.1 (Sampled Tower Delay)**.** *For every $H \geq 4$, there exists a finite-horizon MDP with pairwise disjoint state sets across stages, two actions at every state, and minimum positive Bellman gap $\Delta_{\min} = 1/4$. Its rewards and transitions are deterministic, and each episode starts at its root. There are positive constants U_H and η_H^{cap} such that the MDP and both constants depend only on H. For sufficiently large H, U_H is bounded below by an exponential tower with $H - O(\log H)$ exponentiation levels. Fix any scalar step cap $0 < \eta_{\max} \leq \eta_H^{\mathrm{cap}}$. Run REINFORCE from zero initialization. Each scalar step η_k may depend on the history before episode k, subject to $0 \leq \eta_k \leq \eta_{\max}$ almost surely. Then, for every deterministic integer $K \geq 1$ satisfying $\eta_{\max} K \leq U_H/2$,*

$$\mathbb{E}[\mathrm{Reg}_K] \geq K/8. \tag{15}$$

This sampled lower bound quantifies a long initial learning cost. For fixed instances, the $O_{\mathcal{M}}(\log^2 K)$ expected regret guarantee of Corollary 3.2 remains asymptotic in K under the certified constant-step tuning chosen for each budget. The construction, constants, proofs, and a separate mean-flow result are provided in Appendix E.

Under the certified tuning of Corollary 3.2, a terminal two-action instance gives an $\Omega(\log^2 K)$ regret lower bound for sufficiently large K (see Appendix E.4).

For each constant step size $\eta > 0$, Appendix E.5 constructs a fixed deterministic two-stage instance where expected regret grows as a positive power of K. Consequently, no single positive constant step size shared across all finite MDPs guarantees logarithmic expected regret.

6 Conclusion

In the stated finite-horizon setting, we establish that REINFORCE initialized from zero logits achieves expected regret of order $O_{\mathcal{M}}(\log^2 K)$. For an episode budget $K \geq 2$, this guarantee is obtained by choosing the step size $\eta = \bar{\eta}/\log K$ and holding it constant within the run. We bound Bellman debt by downstream stage regrets under the same policy, enabling geometric stage weights to compensate for downstream value errors in a linear potential. Along the mean flow, cumulative action use limits probability decay and yields quantitative visitation bounds that establish readiness in finite instance-dependent time. Pre-readiness tracking and the post-readiness barrier then ensure a floor on the total probability of optimal actions with high probability, allowing the potential to control cumulative regret. The bound has large instance-dependent constants. Extending the argument for endogenous visitation to time-varying steps in a single run without restarts and to function approximation remains an open direction.

AI USE STATEMENT

We used generative AI tools to help derive and check proof details, including intermediate arguments and the wording of proofs. These tools also helped identify potentially relevant literature, organize research notes, and draft and edit parts of the manuscript. We did not use them to generate research ideas. The authors carefully checked AI-generated content, including proof details and references, and take responsibility for the final mathematical statements, proofs, citations, and text.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL https://www.jmlr.org/papers/v22/19-736.html.

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 the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Dorian Baudry, Emmeran Johnson, Simon Vary, Ciara Pike-Burke, and Patrick Rebeschini. Does stochastic gradient really succeed for bandits? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/a4e683f0ce6b91e7fbdae9d32642d88f-Abstract-Conference.html.

Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. In *International Conference on Learning Representations, Workshop Track*, 2017. URL https://openreview.net/forum?id=Bk9mxlSFx.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008. ISBN 9780521515924.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1283–1294. PMLR, 2020. URL https://proceedings.mlr.press/v119/cai20d.html.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. Best of both worlds policy optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6968–7008. PMLR, 2023. URL https://proceedings.mlr.press/v202/dann23b.html.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9827–9869. PMLR, 2023. URL https://proceedings.mlr.press/v202/fatkhullin23a.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Yanwei Jia and Du Ouyang. Convergence and regret of the policy gradient for multi-armed bandits in diffusion environment, 2026. URL https://arxiv.org/abs/2607.29593. Version 1, 2026-07-31.

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1VeQ6VBbev.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Stochastic Modelling and Applied Probability*. Springer, 2 edition, 2003. doi: 10.1007/b97441. URL https://doi.org/10.1007/b97441.

Safwan Labbi, Daniil Tiapkin, Paul Mangold, and Eric Moulines. Beyond softmax and entropy: Convergence rates of policy gradients with f-softargmax parameterization & coupled regularization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O93c9H4SXc.

Tor Lattimore. A diffusion analysis of policy gradient for stochastic bandits, 2026a. URL https://arxiv.org/abs/2603.10219.

Tor Lattimore. A lyapunov analysis of softmax policy gradient for stochastic bandits, 2026b. URL https://arxiv.org/abs/2603.26547.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, 201:707–802, 2023. doi: 10.1007/s10107-022-01920-6. URL https://arxiv.org/abs/2102.11270. Conference announcement: COLT 2021.

Mingyi Li, Taira Tsuchiya, and Kenji Yamanishi. Policy optimization achieves data-dependent regret bounds in MDPs with unknown transitions, 2026. URL https://arxiv.org/abs/2606.31769. Version 1, 2026-06-30.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22931–22942, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/c1b8bf9e071c0dabb899e7a27f353762-Abstract.html.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. In *Advances in Neural Information Processing Systems*, volume 33, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/f1cf2a082126bf02de0b307778ce73a7-Abstract.html.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 2020b. URL https://proceedings.mlr.press/v119/mei20b.html.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/file/a12f69495f41bb3b637ba1b6238884d6-Paper.pdf.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. The role of baselines in policy gradient optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/718d02a76d69686a36eccc8cde3e6a41-Abstract-Conference.html.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvári, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24325–24360. PMLR, 2023. URL https://proceedings.mlr.press/v202/mei23a.html. Corrected author version: arXiv:2402.17235 (2024).

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994. doi: 10.1002/9780470316887. URL https://doi.org/10.1002/9780470316887.

Samuel Robertson, Thang D. Chu, Bo Dai, Dale Schuurmans, Csaba Szepesvári, and Jincheng Mei. REINFORCE converges to optimal policies with any learning rate. In *Advances in Neural Information Processing Systems*, volume 38, pp. 16010–16053, 2025. doi: 10.52202/

085713-0541. URL https://papers.neurips.cc/paper_files/paper/2025/hash/17a84369cf6e2803101294b5d74dd92c-Abstract-Conference.html.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8604–8613. PMLR, 2020. URL https://proceedings.mlr.press/v119/shani20a.html.

Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/10a5ab2db37feedfdeaab192ead4ac0e-Abstract.html.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. ISBN 9780262039246. URL https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. URL https://jmlr.org/papers/v23/22-0056.html.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular MDPs via adaptive multi-step bootstrap. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4438–4472. PMLR, 2021. URL https://proceedings.mlr.press/v134/xu21a.html.

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3332–3380. PMLR, 2022. URL https://proceedings.mlr.press/v151/yuan22a.html.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10887–10895, 2021. doi: 10.1609/aaai.v35i12.17300. URL https://doi.org/10.1609/aaai.v35i12.17300.

NOTATION AND STANDING SETUP FOR THE APPENDICES

The appendices use the model and notation below throughout. Appendix A proves the expected regret upper bound. Appendix B supplies the required probability floor for optimal actions, using pre-readiness tracking in Appendix C and readiness in Appendix D. Appendix E gives the separate lower-bound constructions. Appendices F and G provide the constant definitions, boundary cases, and elementary proofs.

For an integer $m \geq 1$, write $[m] := \{1, \ldots, m\}$; $K \vee 2 := \max\{K, 2\}$. The index h denotes a stage, k an episode, and u continuous mean-flow time.

MDP and Feedback Model. Let

$$M = \big(\{\mathcal{S}_h\}_{h=1}^{H}, \mathcal{A}, \rho, \{P_h\}_{h=1}^{H-1}, \{r_h\}_{h=1}^{H}\big)$$

be a finite-horizon tabular Markov decision process (MDP). The state sets $\mathcal{S}_1, \ldots, \mathcal{S}_H$ are disjoint, and the common action set $\mathcal{A}$ has cardinality A. The initial distribution is $\rho \in \Delta(\mathcal{S}_1)$; transitions satisfy $P_h(\cdot \mid s, a) \in \Delta(\mathcal{S}_{h+1})$, and rewards satisfy $r_h(s, a) \in [0, 1]$. We use the following information model throughout.

For $h < j$, stage h is upstream of stage j, and stage j is downstream of stage h. Relative to stage h, the downstream policy is $(\pi_j)_{j=h+1}^{H}$. If $P_h(x \mid s, a) > 0$ for some $a \in \mathcal{A}$, then s is a parent state of x. The state x is a child state of s along that transition.

The learner does not know or estimate the transition kernels. Reward functions $r_h : \mathcal{S}_h \times \mathcal{A} \to [0, 1]$ are bounded and deterministic. In episode k, the learner observes one on-policy state–action trajectory and the value $r_h(S_{k,h}, A_{k,h})$ at each visited pair. Actions and transitions are sampled.

A stage-indexed logit table θ induces the Markov policy

$$\pi_{\theta,h}(a \mid s) = \frac{\exp(\theta_h(s, a))}{Z_{h,s}(\theta)}, \qquad Z_{h,s}(\theta) := \sum_{b \in \mathcal{A}} \exp(\theta_h(s, b)). \tag{N.1}$$

Before episode k, the learner has logits θ_k and uses $\pi_k := \pi_{\theta_k}$. The trajectory is generated by

$$S_{k,1} \sim \rho, \qquad A_{k,h} \sim \pi_{k,h}(\cdot \mid S_{k,h}), \qquad S_{k,h+1} \sim P_h(\cdot \mid S_{k,h}, A_{k,h}), \quad h \in [H-1].$$

Let $\mathcal{F}_{k-1}$ be the sigma-field generated by the first $k-1$ complete episodes. Then θ_k is $\mathcal{F}_{k-1}$-measurable and the episode-k trajectory is sampled conditionally from π_k.

Define the exact return-to-go

$$G_{k,h} := \sum_{j=h}^{H} r_j(S_{k,j}, A_{k,j}) \in [0, \overline{H}_h], \qquad \overline{H}_h := H - h + 1, \tag{N.2}$$

and the REINFORCE estimator without a baseline

$$\widehat{g}_{k,h}(s, a) := \mathbf{1}\{S_{k,h} = s\}\big(\mathbf{1}\{A_{k,h} = a\} - \pi_{k,h}(a \mid s)\big)G_{k,h}. \tag{N.3}$$

Starting from zero logits, the REINFORCE update rule is

$$\theta_{1,h}(s, a) = 0, \qquad \theta_{k+1} = \theta_k + \eta \widehat{g}_k. \tag{N.4}$$

The proof of the upper bound uses a step size $\eta > 0$ that remains constant throughout each run. The sampled tower lower bound explicitly allows a predictable scalar step η_k in the same update.

Values, Gaps, Occupancies, and Regret. For any Markov policy π, define

$$Q_h^\pi(s, a) := r_h(s, a) + \begin{cases} \displaystyle\sum_{x \in \mathcal{S}_{h+1}} P_h(x \mid s, a) V_{h+1}^\pi(x), & h < H, \\ 0, & h = H, \end{cases} \tag{N.5}$$

$$V_h^\pi(s) := \sum_{a \in \mathcal{A}} \pi_h(a \mid s) Q_h^\pi(s, a), \qquad \mathrm{Adv}_h^\pi(s, a) := Q_h^\pi(s, a) - V_h^\pi(s), \tag{N.6}$$

$$J(\pi) := \mathbb{E}_{S_1\sim\rho}[V_1^\pi(S_1)]. \tag{N.7}$$

Let V_h^*, Q_h^* be the optimal value functions, and define

$$\mathcal{A}_h^\star(s) := \arg\max_{a\in\mathcal{A}} Q_h^*(s,a), \tag{N.8}$$

$$\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a) \geq 0, \tag{N.9}$$

$$\alpha_h^\pi(s) := \pi_h(\mathcal{A}_h^\star(s) \mid s) = \sum_{a\in\mathcal{A}_h^\star(s)} \pi_h(a \mid s). \tag{N.10}$$

Write $d_h^\pi(s) := \mathbb{P}_{M,\pi}(S_h = s)$ for the stage-h occupancy. The policy regret, its episode-k specialization, and its cumulative version are

$$\mathcal{R}(\pi) := J^* - J(\pi), \qquad \mathcal{R}_k := \mathcal{R}(\pi_k), \qquad \mathrm{Reg}_K := \sum_{k=1}^K \mathcal{R}_k, \qquad J^* := \max_\pi J(\pi). \tag{N.11}$$

A state $s \in \mathcal{S}_h$ is *structurally reachable* if some root-to-s path has strictly positive initial mass and strictly positive transition probability on every edge. Let $\mathcal{S}_h^+$ be the set of such states. Because the MDP is finite, the positive reachability scale

$$p_{\min} := \min\Big(\{\rho(s) : \rho(s) > 0\} \cup \{P_h(x \mid s,a) : s \in \mathcal{S}_h^+,\ P_h(x \mid s,a) > 0\}\Big) \tag{N.12}$$

is well defined and positive. Empty sets of transition probabilities, as in $H = 1$, are simply omitted. Let

$$\mathcal{I}_+ := \{(h,s,a) : s \in \mathcal{S}_h^+,\ \Delta_h(s,a) > 0\}, \qquad \Delta_{\min} := \min_{(h,s,a)\in\mathcal{I}_+} \Delta_h(s,a) \tag{N.13}$$

whenever $\mathcal{I}_+ \neq \varnothing$. The case $\mathcal{I}_+ = \varnothing$ is treated in Section F; its regret is identically zero.

For later use, define the local optimal Bellman loss, stage-h regret contribution, and value error

$$\ell_h^\pi(s) := \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\Delta_h(s,a), \tag{N.14}$$

$$\mathcal{R}_h(\pi) := \sum_{s\in\mathcal{S}_h} d_h^\pi(s)\ell_h^\pi(s), \tag{N.15}$$

$$e_h^\pi(s) := V_h^*(s) - V_h^\pi(s). \tag{N.16}$$

Mean Update and Mean Flow. The conditional mean of the trajectory estimator and its centered noise are

$$g_h(\theta; s,a) = d_h^{\pi_\theta}(s)\pi_{\theta,h}(a \mid s)\,\mathrm{Adv}_h^{\pi_\theta}(s,a), \qquad \varepsilon_k := \widehat{g}_k - g(\theta_k).$$

Thus $\mathbb{E}[\varepsilon_k \mid \mathcal{F}_{k-1}] = 0$; Appendix G proves the mean and moment identities. The mean flow satisfies

$$\dot{\bar\theta}(u) = g(\bar\theta(u)), \qquad \bar\theta(0) = 0, \qquad \bar\pi_u := \pi_{\bar\theta(u)}.$$

The bar marks the ODE solution. We compare the stochastic iterate θ_k with $\bar\theta(u_k)$ at $u_k = (k-1)\eta$.

Inverse Coordinates and Proof Constants. For the sampled iterates and the mean flow, respectively, write

$$X_{k,h,s,a} := e^{-\theta_{k,h}(s,a)}, \qquad \bar X_{h,s,a}(u) := e^{-\bar\theta_h(u;s,a)}.$$

Their softmax normalizers are $Z_{k,h,s} := \sum_a X_{k,h,s,a}^{-1}$ and $\bar Z_{h,s}(u) := \sum_a \bar X_{h,s,a}(u)^{-1}$. The minimum inverse coordinates among optimal actions are

$$X_{k,h,s}^{\star,\min} := \min_{o\in\mathcal{A}_h^\star(s)} X_{k,h,s,o}, \qquad \bar X_{h,s}^{\star,\min}(u) := \min_{o\in\mathcal{A}_h^\star(s)} \bar X_{h,s,o}(u).$$

All optimal actions are retained when ties occur. For an action set B, use $\theta_h(s,B) := \sum_{a\in B}\theta_h(s,a)$ and $g_h(\theta;s,B) := \sum_{a\in B} g_h(\theta;s,a)$. The structural parameters are $\xi := (H, A, (|\mathcal{S}_h|)_h, p_{\min})$. On the nondegenerate branch, μ_0 is the certified optimal-set probability floor and $\bar\eta$ the certified-step numerator. Both depend on ξ and $\Delta_{\min}$ and are independent of K. Appendix F specifies their full dependence.

A Proof of the Regret Bound

A.1 Main Theorem and Certified Step Selection

We restate the upper bound and its step-size corollary in full. Throughout this appendix, the model, algorithm, and notation are those in the standing setup. For the budget K, define

$$n_\theta := A\sum_{h=1}^{H}|\mathcal{S}_h|, \qquad L_K(\eta) := \log\big(n_\theta K[1+K(\eta H+\eta^2H^2)]\big)\,. \tag{A.1}$$

This threshold $L_K(\eta)$ is the logit localization scale appearing in Theorem 3.1. To verify the uniform logarithmic bound (5) stated in Section 3, note that for $K \geq 2$, (B.11) gives $\eta H \leq \bar{\eta}H/\log K \leq \Delta_{\min}/(64H) \leq 1/64$. Defining $\beta := \eta H + \eta^2H^2 \leq 65/4096 < 1$, we obtain $1 + K\beta \leq 2K$ for all $K \geq 2$. Consequently, $L_K(\eta) \leq \log(2n_\theta K^2) = 2\log K + \log(2n_\theta)$ as claimed.

Theorem A.1 (Regret bound)**.** *Fix an MDP with the model, feedback, and REINFORCE update specified in the standing setup, and suppose $\mathcal{I}_+ \neq \varnothing$. Initialize $\theta_1 = 0$. Set $\mu_0 := \mu_\xi(\Delta_{\min}) \in (0, 1/A)$ and $\bar{\eta} := \bar{\eta}_\xi(\Delta_{\min}) \in (0, 1)$. Appendices B and F construct these instance constants independently of K. For every announced episode budget and constant within-run step satisfying*

$$K \geq 2, \qquad 0 < \eta \leq \frac{\bar{\eta}}{\log(K \vee 2)}, \tag{A.2}$$

the expected regret obeys

$$\mathbb{E}[\mathrm{Reg}_K] \leq \frac{A\big(L_K(\eta)+\eta H\big)}{\eta\mu_0}\sum_{h=1}^{H}|\mathcal{S}_h|(1+\mu_0^{-1})^{h-1} + 2H. \tag{A.3}$$

Consequently, for every fixed instance,

$$\mathbb{E}[\mathrm{Reg}_K] = O_{\mathcal{M}}\left(\frac{\log K}{\eta}\right). \tag{A.4}$$

Corollary A.2 (Largest certified step)**.** *Under the conditions of Theorem A.1, let $\delta := \Delta_{\min}$ and define the reciprocal constants*

$$m_\xi(\delta) := \mu_\xi(\delta)^{-1}, \qquad M_\xi(\delta) := \bar{\eta}_\xi(\delta)^{-1}.$$

Their full expansions are given in Appendix F. Define

$$W_\xi(\delta) := \sum_{h=1}^{H}|\mathcal{S}_h|(1+m_\xi(\delta))^{h-1}. \tag{A.5}$$

Choose the largest certified step

$$\eta_K = \frac{1}{M_\xi(\delta)\log(K \vee 2)}. \tag{A.6}$$

Then the bound is

$$\mathbb{E}[\mathrm{Reg}_K] \leq Am_\xi(\delta)\left[M_\xi(\delta)\log(K \vee 2)L_K\left(\frac{1}{M_\xi(\delta)\log(K \vee 2)}\right) + H\right]W_\xi(\delta) + 2H. \tag{A.7}$$

In particular, set

$$\mathfrak{F}_\xi(\delta) := m_\xi(\delta)M_\xi(\delta)W_\xi(\delta), \qquad C_\xi := A\left(3 + \frac{\log(2n_\theta)}{\log 2}\right). \tag{A.8}$$

Both C_ξ and the additive term below are independent of δ, and

$$\mathbb{E}[\mathrm{Reg}_K] \leq C_\xi\mathfrak{F}_\xi(\Delta_{\min})\log^2 K + 2H. \tag{A.9}$$

This establishes the tuned regret bound with leading constant $C_{\mathcal{M}} := C_\xi\mathfrak{F}_\xi(\Delta_{\min})$ defined in (A.8). The constant $C_{\mathcal{M}}$ is positive, finite, and independent of K. It depends solely on the fixed MDP instance, matching Corollary 3.2 and the Introduction.

Remark A.3 (Budget-only step family with fixed ε). Fix an instance $\mathcal{M}$ and $\varepsilon > 0$. For every announced episode budget $K \geq 2$, setting $\eta_K = (1+\log K)^{-(1+\varepsilon)}$ and holding it constant throughout the run guarantees

$$\mathbb{E}[\mathrm{Reg}_K] = O_{\mathcal{M},\varepsilon}\big((1+\log K)^{2+\varepsilon}\big)\,. \tag{A.10}$$

Proof. Fix the instance $\mathcal{M}$ and $\varepsilon > 0$, and let $\bar{\eta}$ be the instance step threshold from Theorem A.1. For any $K \geq 2$, $\eta_K \log K \leq (1+\log K)^{-\varepsilon} \to 0$ as $K \to \infty$. Thus there exists a deterministic threshold $K_0 = K_0(\mathcal{M}, \varepsilon)$ such that $\eta_K \leq \bar{\eta}/\log K$ for all $K \geq K_0$. By (A.4), there is a finite constant $C_1(\mathcal{M})$ such that for all $K \geq K_0$,

$$\mathbb{E}[\mathrm{Reg}_K] \leq C_1(\mathcal{M})\frac{\log K}{\eta_K} \leq C_1(\mathcal{M})(1+\log K)^{2+\varepsilon}.$$

For the finitely many earlier budgets $2 \leq K < K_0$, the pathwise bound $0 \leq \mathrm{Reg}_K \leq HK$ holds because stage rewards are bounded in $[0,1]$. Setting $C(\mathcal{M},\varepsilon) := \max\{C_1(\mathcal{M}), HK_0\}$ gives $\mathbb{E}[\mathrm{Reg}_K] \leq C(\mathcal{M},\varepsilon)(1+\log K)^{2+\varepsilon}$ for all $K \geq 2$. □

Remark A.4 (Anytime regret via deterministic restarts). Run REINFORCE in epochs $j = 0, 1, 2, \ldots$ of length $N_j = 2^{2^j}$, resetting all logits to zero at the start of each epoch. Setting $\eta_j = (1+\log N_j)^{-2}$ in epoch j achieves $\mathbb{E}[\mathrm{Reg}_T] = O_{\mathcal{M}}((1+\log T)^3)$ for every deterministic evaluation budget $T \geq 1$. When the instance parameter $\bar{\eta}$ is supplied and each epoch sets $\eta_j = \bar{\eta}/\log N_j$, the bound improves to $\mathbb{E}[\mathrm{Reg}_T] = O_{\mathcal{M}}((1+\log T)^2)$.

Proof. If $\mathcal{I}_+ = \varnothing$, every policy regret vanishes. Otherwise, define scheduled cumulative epoch lengths $S_0 := 0$ and $S_m := \sum_{j=0}^{m-1} N_j$ for $m \geq 1$. For any deterministic evaluation budget $T \geq 1$, let $m = m(T) \geq 0$ be the unique index satisfying $S_m < T \leq S_{m+1}$, so that the execution reaches epoch m with prefix length $L := T - S_m \in [1, N_m]$.

Write $\mathrm{Reg}_n^{(j)}$ for regret in the first n episodes of epoch j. Conditional on the preceding epochs, the reset run starts from zero on the fixed instance $\mathcal{M}$ and has deterministic full duration $N_j \geq 2$. Thus Remark A.3 (with $\varepsilon = 1$) provides a shared complete-epoch constant $C_{\mathcal{M}}$ such that $\mathbb{E}[\mathrm{Reg}_{N_j}^{(j)}] \leq C_{\mathcal{M}}(1+\log N_j)^3$ for all $j \geq 0$. Coupling the L-episode prefix of epoch m to a full N_m-episode run from the same reset, and using that policy regret is nonnegative in every episode, we have $\mathrm{Reg}_L^{(m)} \leq \mathrm{Reg}_{N_m}^{(m)}$ pathwise. By linearity of expectation,

$$\mathbb{E}[\mathrm{Reg}_T] \leq \sum_{j=0}^{m} \mathbb{E}[\mathrm{Reg}_{N_j}^{(j)}] \leq C_{\mathcal{M}} \sum_{j=0}^{m} (1+\log N_j)^3.$$

For $m = 0$, $T \in \{1, 2\}$, and the first epoch regret is bounded by the fixed constant $2H$. For $m \geq 1$, $T > S_m \geq N_{m-1}$, which implies $\log N_m = 2\log N_{m-1} \leq 2\log T$. Since $\log N_j = 2^j \log 2$, the geometric sum satisfies $\sum_{j=0}^{m}(1+\log N_j)^3 \leq C_3'(1+\log N_m)^3 \leq C_3''(1+\log T)^3$ for absolute constants $C_3', C_3'' > 0$. Combining with the base case yields $\mathbb{E}[\mathrm{Reg}_T] = O_{\mathcal{M}}((1+\log T)^3)$ for every deterministic $T \geq 1$.

When $\bar{\eta}$ is supplied, setting $\eta_j = \bar{\eta}/\log N_j$ within epoch j yields the shared complete-epoch bound $\mathbb{E}[\mathrm{Reg}_{N_j}^{(j)}] \leq C_{\mathcal{M}}'(1+\log N_j)^2$ by Corollary A.2. The identical prefix domination and geometric sum over squares yield $\mathbb{E}[\mathrm{Reg}_T] = O_{\mathcal{M}}((1+\log T)^2)$ for every deterministic $T \geq 1$. □

A.2 Proof Strategy and Potential

To control cumulative regret, we sum the expected increments of a linear potential over optimal logits. To make its drift charge the current regret, we weight stages geometrically so that downstream Bellman debt cancels. This step requires a positive lower bound on the total probability of optimal actions, which Appendix B proves. To bound the total increase of the potential, we control the minimum logit and use the zero-sum property within each block. We stop when either control fails, telescope the potential before that time, and bound the remaining expected regret by $2H$.

Write $\alpha_h^\pi(s) := \pi_h(\mathcal{A}_h^\star(s) \mid s)$ for the total optimal action mass. Here $\mathcal{R}_h(\pi) = \sum_s d_h^\pi(s)\ell_h^\pi(s)$ and $e_h^\pi = V_h^\star - V_h^\pi$ are the stage contribution and value error defined in the standing setup. For a logit

table θ, use $\theta_h(s, B) := \sum_{a \in B} \theta_h(s, a)$ for an action set B. The stage-h optimal-logit sum and its geometrically weighted aggregate are

$$\sum_{s \in \mathcal{S}_h} \theta_h(s, \mathcal{A}_h^\star(s)) = \sum_{s \in \mathcal{S}_h} \sum_{o \in \mathcal{A}_h^\star(s)} \theta_h(s, o), \tag{A.11}$$

$$w_h(\mu) := (1 + \mu^{-1})^{h-1}, \qquad Y_\mu(\theta) := \sum_{h=1}^{H} w_h(\mu) \sum_{s \in \mathcal{S}_h} \theta_h(s, \mathcal{A}_h^\star(s)). \tag{A.12}$$

For the budget K, also set

$$n_\theta := A \sum_{h=1}^{H} |\mathcal{S}_h|, \qquad \beta_\eta := \eta H + \eta^2 H^2, \qquad B_K^{\text{logit}} := L_K(\eta) = \log\big(n_\theta K[1 + K\beta_\eta]\big). \tag{A.13}$$

A.3 Three Inputs to the Regret Bound

The probability-floor lemma in Appendix B states that, under the step-size condition above,

$$\mathbb{P}\left(\min_{1 \le k \le K} \min_{h,s} \alpha_h^{\pi_k}(s) \le \mu_0 \right) \le K^{-1}.$$

We use it together with the following two lemmas. Their proofs appear in the remaining subsections of this appendix.

Lemma A.5 (Weighted Linear Potential). *Let $\mu > 0$. Whenever every optimal-set mass is at least μ,*

$$\langle g(\theta), \nabla Y_\mu(\theta) \rangle \ge \mu \mathcal{R}(\pi_\theta). \tag{A.14}$$

For the exact discrete update,

$$Y_\mu(\theta_{k+1}) - Y_\mu(\theta_k) = \eta \, \langle \widehat{g}_k, \nabla Y_\mu \rangle \, . \tag{A.15}$$

Lemma A.6 (Logit Localization). *Suppose $\mathcal{I}_+ \neq \varnothing$ and the step-size condition of Lemma B.1 holds. Then*

$$\mathbb{P}\left(\min_{1 \le k \le K+1} \min_{h,s,a} \theta_{k,h}(s, a) \le -B_K^{\text{logit}} \right) \le \frac{1}{K}. \tag{A.16}$$

A.4 Proof of the Main Bound and Step-Size Corollary

Proof of Theorem A.1. The theorem is stated on the nondegenerate branch $\mathcal{I}_+ \neq \varnothing$, in which all constants below are defined. The zero-regret branch $\mathcal{I}_+ = \varnothing$ is handled separately in Appendix F.

Let τ be the minimum of $K + 1$ and the following two stopping indices:

1. the first $k \le K$ for which $\min_{h,s} \alpha_h^{\pi_k}(s) \le \mu_0$;
2. the first $k \le K + 1$ for which $\min_{h,s,a} \theta_{k,h}(s, a) \le -B_K^{\text{logit}}$.

For every episode $k < \tau$, the probability floor is still valid. Hence Lemma A.5 and Equation (A.15) give

$$\mathbb{E}[Y_{\mu_0}(\theta_{k+1}) - Y_{\mu_0}(\theta_k) \mid \mathcal{F}_{k-1}] \ge \eta \mu_0 \mathcal{R}_k. \tag{A.17}$$

Because $\{k < \tau\} \in \mathcal{F}_{k-1}$, multiplying this inequality by $\mathbf{1}\{k < \tau\}$ preserves its conditional direction. Summing from $k = 1$ to K therefore yields

$$\mathbb{E}\left[\sum_{k=1}^{\tau-1} \mathcal{R}_k \right] \le \frac{\mathbb{E}[Y_{\mu_0}(\theta_\tau) - Y_{\mu_0}(\theta_1)]}{\eta \mu_0}. \tag{A.18}$$

At the stopped endpoint, every logit is at least $-B_K^{\text{logit}} - \eta H$; the additional ηH accounts for one-step overshoot. In any state–stage block, the logits sum to zero by Equation (G.10). The technical

subset-sum bound Lemma G.4 therefore converts the common lower bound on all coordinates into an upper bound for the sum over the optimal action set. Applying it with $C = B_K^{\text{logit}} + \eta H$ gives

$$Y_{\mu_0}(\theta_\tau) \le A(B_K^{\text{logit}} + \eta H) \sum_{h=1}^{H} |\mathcal{S}_h|(1+\mu_0^{-1})^{h-1}. \tag{A.19}$$

The initialization is zero, so $Y_{\mu_0}(\theta_1) = 0$. Combining Equations (A.18) and (A.19) gives

$$\mathbb{E}\left[\sum_{k=1}^{\tau-1} \mathcal{R}_k\right] \le \frac{A(B_K^{\text{logit}} + \eta H)}{\eta \mu_0} \sum_{h=1}^{H} |\mathcal{S}_h|(1+\mu_0^{-1})^{h-1}. \tag{A.20}$$

The instantaneous regret is at most H. By Lemmas A.6 and B.1, the two stopping events each have probability at most $1/K$. Because

$$\sum_{k=\tau}^{K} \mathcal{R}_k \le HK\mathbf{1}\{\tau \le K\}, \qquad \mathbb{P}(\tau \le K) \le \frac{2}{K}, \tag{A.21}$$

all episodes at or after a premature stop contribute at most $2H$ in expectation. Thus

$$\boxed{\mathbb{E}[\text{Reg}_K] \le \frac{A(B_K^{\text{logit}} + \eta H)}{\eta \mu_0} \sum_{h=1}^{H} |\mathcal{S}_h|(1+\mu_0^{-1})^{h-1} + 2H.} \tag{A.22}$$

For a fixed finite instance, $B_K^{\text{logit}} = O(\log K)$, proving $\mathbb{E}[\text{Reg}_K] = O(\log K/\eta)$. Choosing $\eta = \bar{\eta}_\xi(\Delta_{\min})/\log(K \vee 2)$ and substituting the gap-explicit reciprocal constants gives corollary A.2. □

Proof of corollary A.2. We first evaluate (A.3) with the reciprocal floor $m_\xi(\delta) = \mu_0^{-1}$ and weight $W_\xi(\delta) = \sum_{h=1}^{H} |\mathcal{S}_h|(1+m_\xi(\delta))^{h-1}$. Substituting the largest certified step $\eta_K = \frac{1}{M_\xi(\delta)\log(K\vee 2)}$ yields (A.7) directly.

To establish the simplified rate (A.9), recall that the step threshold condition in Equations (B.3) and (B.11) ensures $\eta_K H \le 1/64$. Hence,

$$1 + K(\eta_K H + \eta_K^2 H^2) \le 1 + 2K\eta_K H \le 2K,$$

which implies that the localization scale satisfies

$$L_K(\eta_K) \le \log(2n_\theta K^2) = 2\log K + \log(2n_\theta).$$

For $K \ge 2$, we have $\log(2n_\theta) \le \frac{\log(2n_\theta)}{\log 2}\log K$, so

$$L_K(\eta_K) \le \left(2 + \frac{\log(2n_\theta)}{\log 2}\right)\log K.$$

Since $M_\xi(\delta) \ge \frac{64H}{\log 2}$, for $K \ge 2$ we have $M_\xi(\delta)\log^2 K \ge 64H\log 2 > H$. The additive H term in (A.7) is therefore bounded by $H \le M_\xi(\delta)\log^2 K$, giving

$$M_\xi(\delta)\log(K\vee 2)L_K(\eta_K) + H \le M_\xi(\delta)\log^2 K\left(3 + \frac{\log(2n_\theta)}{\log 2}\right).$$

Multiplying by $Am_\xi(\delta)W_\xi(\delta)$ and substituting $\mathfrak{F}_\xi(\delta) := m_\xi(\delta)M_\xi(\delta)W_\xi(\delta)$ and $C_\xi := A(3 + \frac{\log(2n_\theta)}{\log 2})$ proves (A.9). □

It remains to prove the auxiliary bounds used above. Appendix B proves the probability floor. The rest of this appendix derives the linear potential drift and the logit localization bound.

A.5 From Stage Regret to a Weighted Linear Potential

For $h < H$, define the downstream Bellman error

$$b_h^\pi(s,a) := \sum_{x \in \mathcal{S}_{h+1}} P_h(x \mid s,a) e_{h+1}^\pi(x), \tag{A.23}$$

and set $b_H^\pi \equiv 0$. Also let

$$\bar{b}_h^\pi(s) := \sum_a \pi_h(a \mid s) b_h^\pi(s,a), \qquad b_h^{\pi,*}(s) := \frac{1}{\alpha_h^\pi(s)} \sum_{o \in \mathcal{A}_h^\star(s)} \pi_h(o \mid s) b_h^\pi(s,o). \tag{A.24}$$

The second quantity is well defined because finite softmax policies assign positive mass to every action.

Write $g_h(\theta; s, B) := \sum_{a \in B} g_h(\theta; s, a)$ for the mean update summed over an action set.

Lemma A.7 (Stagewise Drift with Downstream Debt)**.** *If $\alpha_h^{\pi_\theta}(s) \geq \mu$ for every h, s, where $\mu > 0$, then*

$$\sum_s g_h(\theta; s, \mathcal{A}_h^\star(s)) \geq \mu \mathcal{R}_h(\pi_\theta) - \sum_{j=h+1}^{H} \mathcal{R}_j(\pi_\theta). \tag{A.25}$$

Proof. Because

$$Q_h^\pi(s,a) = Q_h^*(s,a) - b_h^\pi(s,a), \tag{A.26}$$

the error recursion gives $e_h^\pi(s) = \ell_h^\pi(s) + \bar{b}_h^\pi(s)$. For each $o \in \mathcal{A}_h^\star(s)$, $Q_h^*(s,o) = V_h^*(s)$, so

$$\mathrm{Adv}_h^\pi(s,o) = e_h^\pi(s) - b_h^\pi(s,o) = \ell_h^\pi(s) + \bar{b}_h^\pi(s) - b_h^\pi(s,o).$$

Multiplying by $d_h^\pi(s)\pi_h(o \mid s)$ and summing over every optimal action gives

$$g_h(\theta; s, \mathcal{A}_h^\star(s)) = d_h^\pi(s) \left[\alpha_h^\pi(s)(\ell_h^\pi(s) + \bar{b}_h^\pi(s)) - \sum_{o \in \mathcal{A}_h^\star(s)} \pi_h(o \mid s) b_h^\pi(s,o) \right].$$

A further sum over states therefore gives

$$\sum_s g_h(\theta; s, \mathcal{A}_h^\star(s)) = \sum_{s \in \mathcal{S}_h} d_h^\pi(s) \alpha_h^\pi(s) \big(\ell_h^\pi(s) + \bar{b}_h^\pi(s) - b_h^{\pi,*}(s) \big). \tag{A.27}$$

The discarded $\bar{b}_h^\pi$ term is nonnegative because $e_{h+1}^\pi \geq 0$. Define the downstream Bellman debt by

$$D_h(\pi) := \sum_s d_h^\pi(s) \sum_{o \in \mathcal{A}_h^\star(s)} \pi_h(o \mid s) b_h^\pi(s,o). \tag{A.28}$$

This quantity retains the probability of actually reaching the state and selecting the optimal action. If $h = H$, both the restricted downstream flux and the debt sum in Equation (A.25) are zero. If $h < H$, the restricted flux is bounded by the full next-stage flux:

$$\sum_s d_h^\pi(s) \alpha_h^\pi(s) b_h^{\pi,*}(s) \leq \sum_s d_h^\pi(s) \sum_a \pi_h(a \mid s) b_h^\pi(s,a) \tag{A.29}$$

$$= \sum_x d_{h+1}^\pi(x) e_{h+1}^\pi(x) = \sum_{j=h+1}^{H} \mathcal{R}_j(\pi). \tag{A.30}$$

The first equality is exactly the occupancy recursion Equation (G.3); the final equality is the suffix identity Equation (G.4) at stage $h+1$. These equalities use the same policy π throughout, including at downstream states. Finally, $\alpha_h^\pi(s)\ell_h^\pi(s) \geq \mu \ell_h^\pi(s)$, which proves Equation (A.25). □

Proof of Lemma A.5. The geometric weights satisfy

$$\sum_{j<h} w_j(\mu) = \mu(w_h(\mu) - 1). \tag{A.31}$$

Multiply Equation (A.25) by $w_h(\mu)$ and sum over h. Interchanging the two finite sums in the debt term gives

$$\begin{aligned}\langle g(\theta), \nabla Y_\mu \rangle &\geq \mu \sum_{h=1}^{H} w_h(\mu) \mathcal{R}_h(\pi_\theta) - \sum_{h=1}^{H} w_h(\mu) \sum_{j>h} \mathcal{R}_j(\pi_\theta) \\ &= \sum_{j=1}^{H} \left[\mu w_j(\mu) - \sum_{h<j} w_h(\mu) \right] \mathcal{R}_j(\pi_\theta).\end{aligned}$$

The coefficient remaining on each stage contribution is

$$\mu w_h(\mu) - \sum_{j<h} w_j(\mu) = \mu,$$

which proves Equation (A.14) together with Equation (G.2). Because the potential is linear in the logits, substituting Equation (N.4) proves Equation (A.15) exactly. □

Within the class of certificates that uses only Equation (A.25), normalizes $w_1 = 1$, and requires a residual coefficient of at least μ on every $\mathcal{R}_h$, these weights are coordinatewise minimal. Setting equality in $\mu w_h - \sum_{j<h} w_j \geq \mu$ recursively yields Equation (A.12).

A.6 Logit Localization

Proof of Lemma A.6. Fix a coordinate $i = (h, s, a)$, write $X_{k,i} = e^{-\theta_{k,i}}$, and set $D_{k,i} := \eta \widehat{g}_{k,i}$. The admissible-step condition implies $|D_{k,i}| \leq 1$. Hence $e^{-D_{k,i}} \leq 1 - D_{k,i} + D_{k,i}^2$. Furthermore,

$$-g_i(\theta_k) \leq \overline{H}_h d_h^{\pi_k}(s) \pi_{k,h}(a \mid s).$$

Multiplying the exponential inequality by $X_{k,i}$ and taking the conditional expectation gives

$$\begin{aligned}\mathbb{E}[X_{k+1,i} - X_{k,i} \mid \mathcal{F}_{k-1}] &\leq -\eta X_{k,i} g_i(\theta_k) + \eta^2 X_{k,i} \mathbb{E}[\widehat{g}_{k,i}^2 \mid \mathcal{F}_{k-1}] \\ &\leq \left(\eta \overline{H}_h + \eta^2 \overline{H}_h^2\right) \frac{d_h^{\pi_k}(s)}{Z_{k,h,s}} \\ &\leq \eta H + \eta^2 H^2 = \beta_\eta.\end{aligned}$$

Here Equation (G.7) controls the second term, the product gauge gives $Z_{k,h,s} \geq A \geq 1$, and $d_h^{\pi_k}(s) \leq 1$. Thus

$$\mathbb{E}[X_{k+1,i} \mid \mathcal{F}_{k-1}] \leq X_{k,i} + \beta_\eta. \tag{A.32}$$

Consequently,

$$M_{k,i} := X_{k,i} + (K + 1 - k)\beta_\eta, \qquad 1 \leq k \leq K + 1,$$

is a nonnegative supermartingale relative to $(\mathcal{F}_{k-1})$, with $M_{1,i} = 1 + K\beta_\eta$. To see the maximal bound directly, stop at the first $k \leq K + 1$ with $M_{k,i} \geq e^{B_K^{\text{logit}}}$, or at $K + 1$ if this never happens. Bounded optional stopping and nonnegativity give

$$e^{B_K^{\text{logit}}} \mathbb{P}\left(\max_{k \leq K+1} M_{k,i} \geq e^{B_K^{\text{logit}}} \right) \leq \mathbb{E}[M_{1,i}] = 1 + K\beta_\eta.$$

Since $X_{k,i} \leq M_{k,i}$, a union bound over the n_θ coordinates yields

$$\mathbb{P}\left(\max_{k \leq K+1} \max_i X_{k,i} \geq e^{B_K^{\text{logit}}} \right) \leq n_\theta [1 + K\beta_\eta] e^{-B_K^{\text{logit}}} = \frac{1}{K},$$

which is equivalent to Equation (A.16). □

B Proof of the Probability Floor for Optimal Actions

B.1 Target and Proof Strategy

The regret proof needs the following probability bound. Its constants $\mu_\xi(\Delta_{\min})$ and $\bar{\eta}_\xi(\Delta_{\min})$ are defined below and expanded in Appendix F; they are independent of K.

Lemma B.1 (Probability floor for optimal actions)**.** *Fix a finite MDP with $\mathcal{I}_+ \neq \varnothing$ under the standing model and feedback assumptions. Use REINFORCE from zero initialization. Set $\mu_0 := \mu_\xi(\Delta_{\min}) \in (0, 1/A)$ and $\bar{\eta} := \bar{\eta}_\xi(\Delta_{\min}) \in (0, 1)$. For every announced budget $K \geq 2$ and constant within-run step $0 < \eta \leq \bar{\eta}/\log(K \vee 2)$, the REINFORCE update (N.4) satisfies*

$$\mathbb{P}\left(\min_{1\leq k\leq K}\min_{h\in[H],\, s\in\mathcal{S}_h} \pi_{k,h}(\mathcal{A}_h^\star(s) \mid s) \leq \mu_0\right) \leq \frac{1}{K}. \tag{B.1}$$

To preserve optimal-set probability from initialization, we divide the analysis at a deterministic readiness index. Before that index, the tracking lemma in Appendix C keeps the stochastic iterates close to the mean flow and gives a positive probability bound for every action. At readiness, the iterate has strict separation between suboptimal inverse coordinates and the minimum optimal inverse coordinate. To preserve this separation afterward, we construct a negative-power supermartingale and bound the probability of leaving the joint safe region. Within that region, at least one optimal action has probability at least $1/A$. Combining the two failure bounds proves the stated floor.

If $\mathcal{I}_+ = \varnothing$, the floor is immediate as recorded in Appendix F. In the probability-floor and readiness arguments, we therefore treat the nondegenerate branch $\mathcal{I}_+ \neq \varnothing$ and abbreviate $\Delta := \Delta_{\min}$.

B.2 Two Proof Obligations and Their Combination

The pre-readiness module in Appendix C supplies an event $\mathcal{G}_{\mathrm{pre}}$. On this event, every action has a fixed positive mass up to k_{ready}, and the iterate enters the safe region with a strict margin. For the post-readiness phase, write

$$X_{k,h,s,a} := e^{-\theta_{k,h}(s,a)}, \qquad Z_{k,h,s} := \sum_{a\in\mathcal{A}} X_{k,h,s,a}^{-1}. \tag{B.2}$$

For the discrete iterate, use the corresponding shorthand

$$X_{k,h,s}^{\star,\min} := \min_{o\in\mathcal{A}_h^\star(s)} X_{k,h,s,o}.$$

Let $\mathcal{J}$ be the finite set of reachable competitor indices $j = (h, s, b)$ with $s \in \mathcal{S}_h^+$ and $b \notin \mathcal{A}_h^\star(s)$.

Set

$$\kappa_0 := \frac{\Delta}{64H^2}, \qquad \nu_\eta := \frac{\kappa_0}{\eta}, \tag{B.3}$$

and assume temporarily that $0 < \eta \leq \kappa_0$. Since $\Delta \leq H$, this implies $\eta H \leq 1/64$.

For $j = (h, s, b)$, define the actual shifted separation

$$B_{j,k} := X_{k,h,s,b} - X_{k,h,s}^{\star,\min} + 2. \tag{B.4}$$

The joint safe region is $B_{j,k} \geq L_h^{\mathrm{sep}} + 2$ for all $j \in \mathcal{J}$. Let

$$\tau_{\mathrm{safe}} := \inf\left\{k \geq k_{\mathrm{ready}} : \exists j = (h, s, b) \in \mathcal{J}, B_{j,k} < L_h^{\mathrm{sep}} + 2\right\}. \tag{B.5}$$

The barrier lemma below uses the exponent

$$\kappa_{\mathrm{post}} := \kappa_0 \log\left(\frac{L_{\max}^{\mathrm{sep}} + 3}{L_{\max}^{\mathrm{sep}} + 2}\right). \tag{B.6}$$

Define the final floor and step-size constants by

$$\mu_0 := \frac{1}{2A}\exp(-2B_H U^\dagger - 2\varepsilon_0), \tag{B.7}$$

$$\eta_{\max} := \min\{1, \eta_{\text{Euler}}, \kappa_0\}, \tag{B.8}$$

$$\kappa_* := \min\{\kappa_{\text{pre}}, \kappa_{\text{post}}\}, \tag{B.9}$$

$$C_\theta := 1 + \frac{\log(3n_\theta)}{\log 2}, \tag{B.10}$$

$$\bar{\eta} := \min\left\{\eta_{\max}\log 2, \frac{\kappa_*}{C_\theta}\right\}. \tag{B.11}$$

These constants μ_0 and $\bar{\eta}$ are precisely the fixed instance constants used in Theorem 3.1 and Lemma 4.2. All constituent ingredients are positive and finite. Since $B_H > 0$, $U^\dagger \geq 1$, and $\varepsilon_0 > 0$, the exponential construction in (B.7) yields the strict floor $0 < \mu_0 < 1/(2A) < 1/A$. Similarly, because $\eta_{\max} \leq \kappa_0 = \Delta_{\min}/(64H^2) \leq 1/64$, $\kappa_* > 0$, and $C_\theta \geq 1$, (B.11) yields $0 < \bar{\eta} \leq \eta_{\max}\log 2 \leq \Delta_{\min}\log 2/(64H^2) < 1$. Both constants depend solely on the fixed MDP instance and are independent of K and η. Appendix F expands μ_0^{-1} and $\bar{\eta}^{-1}$ as explicit functions of $\Delta_{\min}$ and the instance parameters.

Proof of Lemma B.1. On $\mathcal{G}_{\text{pre}}$, Equation (C.9) puts every individual action probability above $2\mu_0$ up to k_{ready}. Thus the desired optimal-set floor holds throughout the first phase. At readiness, Equation (C.10) places every reachable competitor strictly inside the safe region.

After readiness and before τ_{safe}, the minimum optimal inverse coordinate is also the minimum over its whole action block. If o is such a minimum, then $X_a \geq X_o$ for every action a, and hence

$$Z_{k,h,s} = \sum_a X_a^{-1} \leq \frac{A}{X_o}, \qquad \pi_{k,h}(o \mid s) = \frac{1}{X_o Z_{k,h,s}} \geq \frac{1}{A} > \mu_0.$$

Consequently, a violation of the floor must occur either because the pre-readiness tube fails or because the joint safe region is exited. By Lemmas B.2 and C.1,

$$\mathbb{P}\left(\min_{1\leq k\leq K}\min_{h,s}\alpha_h^{\pi_k}(s) \leq \mu_0\right) \leq 2n_\theta e^{-\kappa_{\text{pre}}/\eta} + n_\theta e^{-\kappa_{\text{post}}/\eta} \tag{B.12}$$

$$\leq 3n_\theta e^{-\kappa_*/\eta}. \tag{B.13}$$

If $0 < \eta \leq \bar{\eta}/\log(K \vee 2)$, then $\eta \leq \eta_{\max}$ and

$$\frac{\kappa_*}{\eta} \geq C_\theta \log K \geq \log(3n_\theta K).$$

Hence the failure probability in Equation (B.13) is at most $1/K$. Structurally unreachable blocks remain uniform, while a block in which all actions are optimal has optimal-set mass one. The same pre-readiness bound covers budgets shorter than k_{ready}, completing the proof for all blocks and all $K \geq 2$. □

B.3 The Barrier after Readiness

Lemma B.2 (Barrier after readiness)**.** *The event $\mathcal{G}_{\text{pre}}$ is $\mathcal{F}_{k_{\text{ready}}-1}$-measurable. For every $j = (h, s, b) \in \mathcal{J}$, the process defined by*

$$\mathcal{M}_{j,k} := \begin{cases} B_{j,k\wedge\tau_{\text{safe}}}^{-\nu_\eta}, & \text{on } \mathcal{G}_{\text{pre}}, \\ 0, & \text{on } \mathcal{G}_{\text{pre}}^c, \end{cases} \qquad k \geq k_{\text{ready}} \tag{B.14}$$

is a nonnegative supermartingale relative to $(\mathcal{F}_{k-1})$, including the first exit step. Consequently,

$$\mathbb{P}\left(\mathcal{G}_{\text{pre}} \cap \{\tau_{\text{safe}} < \infty\}\right) \leq n_\theta \exp\left(-\frac{\kappa_{\text{post}}}{\eta}\right), \tag{B.15}$$

where κ_{post} is defined in Equation (B.6).

Proof. Fix a pre-exit iterate k and a competitor $j = (h, s, b)$. Choose, using a fixed deterministic tie-breaking rule, a current optimal minimizer

$$o \in \arg\min_{a\in\mathcal{A}_h^\star(s)} X_{k,h,s,a}.$$

On the safe region this o minimizes the inverse coordinate over the full action block. The product gauge therefore gives $X_o \le 1$, and

$$X_b + X_o \le X_b - X_o + 2 =: B. \tag{B.16}$$

Suppress the block and time indices. Let $D_i := \eta \widehat{g}_{k,h}(s,i)$, so $X_i^+ = X_i e^{-D_i}$, and define a surrogate increment by keeping the current minimizer fixed:

$$R := X_b(e^{-D_b} - 1) - X_o(e^{-D_o} - 1). \tag{B.17}$$

If the optimal minimizer changes after the update, then

$$X_b^+ - \min_{a \in \mathcal{A}_h^\star(s)} X_a^+ \ge X_b^+ - X_o^+. \tag{B.18}$$

Thus the actual next barrier is at least $B + R$; because the transform below is decreasing, switching is favorable.

Let

$$\upsilon_{k,h,s} := \frac{d_h^{\pi_k}(s)}{Z_{k,h,s}}. \tag{B.19}$$

The identity $X_i \pi_{k,h}(i \mid s) = 1/Z_{k,h,s}$ and Equation (G.6) give the exact first-order cancellation

$$\mathbb{E}[X_o \widehat{g}_{k,h}(s,o) - X_b \widehat{g}_{k,h}(s,b) \mid \mathcal{F}_{k-1}] = \upsilon_{k,h,s}\big(Q_h^{\pi_k}(s,o) - Q_h^{\pi_k}(s,b)\big). \tag{B.20}$$

This scalar projection retains the shared return and the complete trajectory law.

For $|x| \le 1/64$,

$$e^{-x} - 1 \ge -x, \qquad e^{-x} - 1 \le -x + x^2, \qquad |e^{-x} - 1| \le 2|x|.$$

Using Equations (B.16), (B.20), (D.12) and (G.7) yields the following bounds. First, applying the lower exponential bound to the competitor and the upper bound to the old optimal minimizer gives

$$R \ge -\eta X_b \widehat{g}_{k,h}(s,b) + \eta X_o \widehat{g}_{k,h}(s,o) - \eta^2 X_o \widehat{g}_{k,h}(s,o)^2.$$

On the joint safe region, the downstream stages meet the premise of Lemma D.3. Hence the conditional expectation of the first two terms is at least $3\eta\Delta\upsilon_{k,h,s}/4$, while

$$\eta^2 X_o \mathbb{E}[\widehat{g}_{k,h}(s,o)^2 \mid \mathcal{F}_{k-1}] \le \eta^2 H^2 \upsilon_{k,h,s} \le \frac{\eta\Delta}{64}\upsilon_{k,h,s}.$$

This is stronger than the first inequality below. Next, $|R| \le 2\eta H(X_b + X_o) \le 2\eta H B$. Finally,

$$\begin{aligned}\mathbb{E}[R^2 \mid \mathcal{F}_{k-1}] &\le 8\eta^2 \mathbb{E}\left[X_b^2 \widehat{g}_{k,h}(s,b)^2 + X_o^2 \widehat{g}_{k,h}(s,o)^2 \mid \mathcal{F}_{k-1}\right] \\ &\le 8\eta^2 H^2 \upsilon_{k,h,s}(X_b + X_o) \le 8\eta^2 H^2 \upsilon_{k,h,s} B,\end{aligned}$$

where $X_i^2 d_h^{\pi_k}(s)\pi_{k,h}(i \mid s) = \upsilon_{k,h,s} X_i$. Therefore

$$\mathbb{E}[R \mid \mathcal{F}_{k-1}] \ge \frac{\eta\Delta}{2}\upsilon_{k,h,s}, \tag{B.21}$$

$$|R| \le 2\eta H B, \tag{B.22}$$

$$\mathbb{E}[R^2 \mid \mathcal{F}_{k-1}] \le 8\eta^2 H^2 \upsilon_{k,h,s} B. \tag{B.23}$$

In particular, $B + R \ge (1 - 2\eta H)B > 0$. This inequality is proved using only the pre-update safe region. It therefore remains valid when the new iterate is the first unsafe iterate. Since the actual next barrier is at least $B + R$, every stopped barrier on $\mathcal{G}_{\mathrm{pre}}$ is at least $2(1 - 2\eta H) > 0$. Its negative power is consequently bounded by a finite constant for each fixed admissible η.

Take $f(x) := x^{-\nu_\eta}$. Along the segment between B and $B + R$,

$$f''(x) \le \nu_\eta(\nu_\eta + 1)(1 - 2\eta H)^{-\nu_\eta - 2} B^{-\nu_\eta - 2}.$$

Using $-\log(1 - y) \le y/(1 - y)$,

$$\log((1 - 2\eta H)^{-\nu_\eta - 2}) \le \frac{2\kappa_0 H + 4\eta H}{1 - 2\eta H} \le \frac{3}{31} < \log 2.$$

Taylor's theorem and Equations (B.21) and (B.23) therefore give

$$\mathbb{E}[f(B+R) - f(B) \mid \mathcal{F}_{k-1}] \le \nu_\eta \eta \upsilon_{k,h,s} B^{-\nu_\eta - 1} \left[-\frac{\Delta}{2} + 8(\nu_\eta + 1)\eta H^2 \right] \le 0,$$

because $(\nu_\eta + 1)\eta = \kappa_0 + \eta \le \Delta/(32H^2)$. Since f is decreasing and the actual next barrier is at least $B + R$, Equation (B.18) gives the same conditional inequality for the actual process. After τ_{safe}, the stopped coordinate is constant. Multiplication by the already-measurable readiness event, with the process defined as zero on its complement, proves Equation (B.14). If the state clock is zero, the state is visited with conditional probability zero, so its sampled update vanishes almost surely and the conclusion is immediate.

On $\mathcal{G}_{\mathrm{pre}}$, Equation (C.10) gives $B_{j,k_{\mathrm{ready}}} > L_h^{\mathrm{sep}} + 3$. For a finite cap m, let $A_{j,m}$ be the event that the first joint exit occurs by m and coordinate j is one of the exiting coordinates. On $A_{j,m}$, the stopped barrier satisfies $B_{j,\tau_{\mathrm{safe}}} < L_h^{\mathrm{sep}} + 2$. Conditional on a readiness history in $\mathcal{G}_{\mathrm{pre}}$, optional stopping at $m \wedge \tau_{\mathrm{safe}}$ therefore gives

$$(L_h^{\mathrm{sep}} + 2)^{-\nu_\eta} \mathbb{P}(A_{j,m} \mid \mathcal{F}_{k_{\mathrm{ready}}-1}) \le \mathbb{E}[B_{j,m\wedge\tau_{\mathrm{safe}}}^{-\nu_\eta} \mid \mathcal{F}_{k_{\mathrm{ready}}-1}] \le B_{j,k_{\mathrm{ready}}}^{-\nu_\eta}.$$

The entry margin thus gives

$$\mathbb{P}(A_{j,m} \mid \mathcal{F}_{k_{\mathrm{ready}}-1}) \le \left(\frac{L_h^{\mathrm{sep}} + 2}{L_h^{\mathrm{sep}} + 3} \right)^{\kappa_0/\eta} \quad \text{on } \mathcal{G}_{\mathrm{pre}}.$$

Letting $m \to \infty$ and taking a union over at most n_θ competitors handles single and simultaneous exits and proves Equations (B.6) and (B.15). □

C Pre-Readiness Tracking

This appendix discharges the pre-readiness obligation used in Appendix B. Its output is the event $\mathcal{G}_{\mathrm{pre}}$: up to a deterministic readiness index, the REINFORCE iterates stay close to the mean flow; at that index they inherit a strict inverse-coordinate margin. To obtain this event, we bound the maximum of a martingale, the accumulated Euler error, and the amplification of both errors through the Lipschitz mean field. We state the complete interface first and derive those three bounds afterward.

Assume $\mathcal{I}_+ \neq \varnothing$ and write $\Delta := \Delta_{\min}$. Appendix D supplies a finite time u_{ready} such that, for all $u \ge u_{\mathrm{ready}}$, every reachable suboptimal action satisfies

$$\bar{X}_{h,s,b}(u) - \bar{X}_{h,s}^{\star,\min}(u) \ge L_h^{\mathrm{sep}} + 2.$$

Here $L_1^{\mathrm{sep}} = 0$ and, when $H \ge 2$, $L_h^{\mathrm{sep}} = 4A(H-1)(H-h+1)/\Delta$ for $2 \le h \le H$. This separation is the deterministic margin that the tracking estimate must transfer to the sampled iterate.

Let

$$n_\theta := A \sum_{h=1}^{H} |\mathcal{S}_h|, \qquad L_{\max}^{\mathrm{sep}} := \max_{1\le h\le H} L_h^{\mathrm{sep}} = \begin{cases} 0, & H = 1, \\ 4A(H-1)^2/\Delta, & H \ge 2, \end{cases} \tag{C.1}$$

and define

$$\varepsilon_0 := \min\left\{ \frac{1}{8}, \frac{1}{16(L_{\max}^{\mathrm{sep}} + 5)} \right\}, \tag{C.2}$$

$$U^\dagger := u_{\mathrm{ready}} + 1, \tag{C.3}$$

$$a_0 := \frac{\varepsilon_0 e^{-\Lambda U^\dagger}}{2\sqrt{n_\theta}}, \tag{C.4}$$

$$\eta_{\mathrm{Euler}} := \frac{\varepsilon_0 e^{-\Lambda U^\dagger}}{\Lambda B_H U^\dagger}, \tag{C.5}$$

$$\kappa_{\mathrm{pre}} := \frac{a_0^2}{2\big(B_H^2 U^\dagger + 2B_H a_0/3\big)}. \tag{C.6}$$

Here Λ and B_H are defined in Equations (G.8) and (G.16).

Lemma C.1 (Euler–Freedman Readiness Tube). *Suppose* $0 < \eta \le \min\{1, \eta_{\text{Euler}}\}$. *Let*

$$k_{\text{ready}} := 1 + \lceil u_{\text{ready}}/\eta \rceil, \qquad u_k := (k-1)\eta. \tag{C.7}$$

Then

$$\mathbb{P}\left(\max_{1 \le k \le k_{\text{ready}}} \left\| \theta_k - \bar{\theta}(u_k) \right\|_2 \ge \varepsilon_0 \right) \le 2 n_\theta \exp\left(-\frac{\kappa_{\text{pre}}}{\eta} \right). \tag{C.8}$$

On the complementary event, denoted $\mathcal{G}_{\text{pre}}$*, every action probability before readiness satisfies*

$$\pi_{k,h}(a \mid s) \ge \frac{1}{A} \exp(-2B_H U^\dagger - 2\varepsilon_0), \qquad k \le k_{\text{ready}}, \tag{C.9}$$

and at $k = k_{\text{ready}}$*, every reachable competitor obeys the strict margin*

$$e^{-\theta_{k,h}(s,b)} - \min_{o \in \mathcal{A}_h^\star(s)} e^{-\theta_{k,h}(s,o)} > L_h^{\text{sep}} + 1. \tag{C.10}$$

Proof. Define the grid error $E_k := \theta_k - \bar{\theta}(u_k)$ and the local ODE remainder

$$r_k := \bar{\theta}(u_{k+1}) - \bar{\theta}(u_k) - \eta g(\bar{\theta}(u_k)).$$

The integral form of the ODE gives

$$r_k = \int_{u_k}^{u_{k+1}} \left[g(\bar{\theta}(v)) - g(\bar{\theta}(u_k)) \right] \mathrm{d}v.$$

Since $\|\bar{\theta}(v) - \bar{\theta}(u_k)\|_2 \le B_H(v - u_k)$, its norm is bounded by $\int_0^\eta \Lambda B_H v \, \mathrm{d}v$. Thus

$$\left\| \bar{\theta}(u_{k+1}) - \bar{\theta}(u_k) - \eta g(\bar{\theta}(u_k)) \right\|_2 \le \frac{\Lambda B_H}{2} \eta^2. \tag{C.11}$$

For coordinate i, define the martingale over the tracking window by

$$\mathcal{N}_{\ell,i} := \eta \sum_{k=1}^{\ell-1} \varepsilon_{k,i}, \qquad 1 \le \ell \le k_{\text{ready}}. \tag{C.12}$$

By Lemma G.3, its increments have magnitude at most $2\eta B_H$. For each coordinate, centering reduces the conditional second moment:

$$\mathbb{E}[\varepsilon_{k,i}^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\widehat{g}_{k,i}^2 \mid \mathcal{F}_{k-1}] - g_i(\theta_k)^2 \le B_H^2.$$

Therefore, its predictable quadratic variation through this window is at most $(k_{\text{ready}} - 1)\eta^2 B_H^2 \le \eta B_H^2 U^\dagger$. Applying Freedman's maximal inequality (Freedman, 1975) to both $\mathcal{N}_{\ell,i}$ and $-\mathcal{N}_{\ell,i}$ yields

$$\mathbb{P}\left(\max_{\ell \le k_{\text{ready}}} |\mathcal{N}_{\ell,i}| \ge a_0 \right) \le 2 \exp(-\kappa_{\text{pre}}/\eta).$$

Taking a union over the n_θ coordinates shows that, outside an event with probability at most the right-hand side of Equation (C.8),

$$\max_{\ell \le k_{\text{ready}}} \|\mathcal{N}_\ell\|_2 < \sqrt{n_\theta} a_0 = \frac{\varepsilon_0 e^{-\Lambda U^\dagger}}{2}. \tag{C.13}$$

The total deterministic remainder satisfies

$$\sum_{k < k_{\text{ready}}} \|r_k\|_2 \le \frac{\Lambda B_H}{2} \eta U^\dagger \le \frac{\varepsilon_0 e^{-\Lambda U^\dagger}}{2}.$$

Summing the difference recursion, with $E_1 = 0$, yields

$$E_\ell = \mathcal{N}_\ell - \sum_{k<\ell} r_k + \eta \sum_{k<\ell} \left[g(\theta_k) - g(\bar{\theta}(u_k)) \right].$$

Set $c = \max_\ell \|\mathcal{N}_\ell\|_2 + \sum_k \|r_k\|_2$. The last display implies $\|E_\ell\|_2 \le c + \eta\Lambda \sum_{k<\ell} \|E_k\|_2$. Induction gives $\|E_\ell\|_2 \le c(1+\eta\Lambda)^{\ell-1} \le c e^{\Lambda U^\dagger}$. On the complement of the coordinatewise threshold events, $c < \varepsilon_0 e^{-\Lambda U^\dagger}$, and hence

$$\max_{k \le k_{\mathrm{ready}}} \left\|\theta_k - \bar\theta(u_k)\right\|_2 < e^{\Lambda U^\dagger} \left(\frac{\varepsilon_0 e^{-\Lambda U^\dagger}}{2} + \frac{\varepsilon_0 e^{-\Lambda U^\dagger}}{2} \right) = \varepsilon_0.$$

This proves Equation (C.8).

The mean flow satisfies $\left\|\bar\theta(u)\right\|_2 \le B_H u$. Hence, on the tube, every logit difference before readiness is at most $2B_H U^\dagger + 2\varepsilon_0$, which yields Equation (C.9) from the softmax formula.

Finally, $u_{k_{\mathrm{ready}}} \ge u_{\mathrm{ready}}$, so Theorem D.1 gives deterministic inverse-coordinate separation at least $L_h^{\mathrm{sep}} + 2$. The tube implies the multiplicative comparison

$$e^{-\varepsilon_0} e^{-\bar\theta_h(u_k;s,a)} \le e^{-\theta_{k,h}(s,a)} \le e^{\varepsilon_0} e^{-\bar\theta_h(u_k;s,a)}.$$

Fix the block at readiness, put $L = L_h^{\mathrm{sep}}$, and abbreviate its deterministic minimum by $\bar X^{\star,\min}$. Readiness makes this the minimum over the whole action block, so the product gauge gives $\bar X^{\star,\min} \le 1$. The multiplicative comparison therefore yields

$$\begin{aligned} X_{k,h,s,b} - X^{\star,\min}_{k,h,s} &\ge e^{-\varepsilon_0}(\bar X_{h,s,b} - \bar X^{\star,\min}) - (e^{\varepsilon_0} - e^{-\varepsilon_0})\bar X^{\star,\min} \\ &\ge e^{-\varepsilon_0}(L+2) - (e^{\varepsilon_0} - e^{-\varepsilon_0}) \\ &\ge L + 2 - \varepsilon_0(L+5) \ge L + 2 - \frac{1}{16} > L + 1. \end{aligned}$$

Here $e^{-x} \ge 1 - x$ and $e^x - e^{-x} \le 3x$ for $0 \le x \le 1/8$, and $L \le L_{\max}^{\mathrm{sep}}$. This proves Equation (C.10). □

If the budget ends before k_{ready}, only Equation (C.9) is used. We may nevertheless continue the same REINFORCE updates beyond the budget for the purpose of bounding $\mathcal{G}_{\mathrm{pre}}$; the first K policies remain the same. Structurally unreachable blocks have zero update and remain uniform exactly.

D Quantitative Endogenous Visitation and Readiness

Assume throughout this section that $\mathcal{I}_+ \neq \varnothing$, and abbreviate

$$\Delta := \Delta_{\min}, \qquad p := p_{\min}.$$

We analyze the mean flow

$$\dot{\bar\theta}(u) = g(\bar\theta(u)), \qquad \bar\theta(0) = 0, \tag{D.1}$$

where u is scaled time. Let $\bar\pi_u := \pi_{\bar\theta(u)}$. This section proves that all reachable state–stage blocks become ready after a finite, explicit time. For the inverse coordinates defined below, write

$$\bar X^{\star,\min}_{h,s}(u) := \min_{o \in \mathcal{A}^\star_h(s)} \bar X_{h,s,o}(u)$$

for the minimum inverse coordinate among optimal actions.

D.1 Target and Backward Dependency Chain

We state this target before developing the local-clock machinery that proves it. The separation thresholds L_h^{sep} and the explicit time u_{ready} are defined in Equations (D.10) and (D.30) below.

Theorem D.1 (Readiness in Finite Time). *For every $u \ge u_{\mathrm{ready}}$, every reachable $s \in \mathcal{S}_h^+$ and every $b \notin \mathcal{A}_h^\star(s)$ satisfy*

$$\bar X_{h,s,b}(u) - \bar X^{\star,\min}_{h,s}(u) \ge L_h^{\mathrm{sep}} + 2. \tag{D.2}$$

The separation is persistent for all later mean-flow times.

The proof proceeds backward from stage H to stage 1. At a generic stage, it uses the following implication chain:

$$\begin{aligned} &\text{downstream separation} \\ &\quad\Longrightarrow \text{a positive current-policy action gap} \\ &\quad\Longrightarrow \text{inverse-coordinate growth in local clock} \\ &\quad\Longrightarrow \text{current-stage separation.} \end{aligned} \tag{D.3}$$

It remains only to show that every reachable state receives sufficient local clock. We establish this property by giving a lower bound on the local clock along a reachable path. The next three subsections establish these modules, and the final subsection combines them according to the displayed chain.

D.2 Inverse Coordinates and Local Clocks

For one block (h, s), define

$$\bar{X}_{h,s,a}(u) := e^{-\bar{\theta}_h(u;s,a)}, \qquad \bar{Z}_{h,s}(u) := \sum_{a\in\mathcal{A}} \bar{X}_{h,s,a}(u)^{-1}. \tag{D.4}$$

The block gauge in Equation (G.10) is preserved by the mean flow, so

$$\prod_{a\in\mathcal{A}} \bar{X}_{h,s,a}(u) = 1, \qquad \bar{\pi}_{u,h}(a \mid s) = \frac{1}{\bar{X}_{h,s,a}(u)\bar{Z}_{h,s}(u)}. \tag{D.5}$$

Lemma D.2 (Exact Inverse-Clock Identity)**.** *For a block (h, s) with positive occupancy, introduce the nondecreasing local clock*

$$\mathrm{d}\chi_{h,s}(u) := \frac{d_h^{\bar{\pi}_u}(s)}{\bar{Z}_{h,s}(u)}\,\mathrm{d}u. \tag{D.6}$$

Then, on the support of this clock,

$$\frac{\mathrm{d}\bar{X}_{h,s,a}}{\mathrm{d}\chi_{h,s}} = V_h^{\bar{\pi}_u}(s) - Q_h^{\bar{\pi}_u}(s,a), \tag{D.7}$$

$$\frac{\mathrm{d}(\bar{X}_{h,s,b} - \bar{X}_{h,s,o})}{\mathrm{d}\chi_{h,s}} = Q_h^{\bar{\pi}_u}(s,o) - Q_h^{\bar{\pi}_u}(s,b). \tag{D.8}$$

On intervals where the local clock does not advance, the block remains constant.

Proof. By Equations (D.1) and (G.6), suppressing (h, s, u),

$$\dot{\bar{\theta}}_a = d\,\pi_a(Q_a - V).$$

Since $\bar{X}_a\pi_a = 1/\bar{Z}$,

$$\dot{\bar{X}}_a = -\bar{X}_a\dot{\bar{\theta}}_a = \frac{d}{\bar{Z}}(V - Q_a).$$

Dividing by the clock density gives Equation (D.7); subtracting two coordinates gives Equation (D.8). □

The gauge also gives $\bar{Z}_{h,s}(u) \geq A$. Because $|V_h^{\bar{\pi}_u}(s) - Q_h^{\bar{\pi}_u}(s,a)| \leq \overline{H}_h$, the preceding proof implies

$$\left|\frac{\mathrm{d}\bar{X}_{h,s,a}}{\mathrm{d}u}\right| \leq \frac{\overline{H}_h}{A}, \qquad \bar{X}_{h,s,a}(U) \leq X_h^{\mathrm{ub}}(U) := 1 + \frac{\overline{H}_h U}{A}. \tag{D.9}$$

D.3 Bellman-Tail Control

For $H \geq 2$, set

$$\varepsilon_{\mathrm{tail}} := \frac{\Delta}{4(H-1)}, \qquad L_1^{\mathrm{sep}} := 0, \qquad L_h^{\mathrm{sep}} := \frac{4A(H-1)\overline{H}_h}{\Delta} \quad (2 \leq h \leq H). \tag{D.10}$$

The following implication is algebraic in the current policy and its inverse coordinates. We state it in mean-flow notation, but the same statement holds for any softmax logit table satisfying the zero-sum block gauge, including a discrete iterate.

Lemma D.3 (Safe Separation and Bellman Tail Control)**.** *Suppose that for every reachable state $s \in \mathcal{S}_j^+$ in stages $j = h+1, \dots, H$, every $b \notin \mathcal{A}_j^\star(s)$ satisfies*

$$\bar{X}_{j,s,b} - \bar{X}_{j,s}^{\star,\min} \geq L_j^{\mathrm{sep}}. \tag{D.11}$$

Then, for every reachable $s \in \mathcal{S}_h^+$, $o \in \mathcal{A}_h^\star(s)$, and $b \notin \mathcal{A}_h^\star(s)$,

$$Q_h^{\bar{\pi}_u}(s,o) - Q_h^{\bar{\pi}_u}(s,b) \geq \frac{3\Delta}{4}. \tag{D.12}$$

At $h = H$, the same conclusion holds with Δ on the right and requires no downstream assumption.

Proof. Suppose first that $h < H$. We bound the downstream local losses, propagate that bound through the Bellman recursion, and then compare current-policy action values.

Downstream local loss. Fix $j \in \{h+1, \dots, H\}$ and $s \in \mathcal{S}_j^+$, and set

$$\bar{X}_{j,s}^{\star,\min} := \min_{o \in \mathcal{A}_j^\star(s)} \bar{X}_{j,s,o}.$$

Since $j \geq 2$, one has $L_j^{\mathrm{sep}} > 0$. Thus Equation (D.11) implies that every nonoptimal inverse coordinate is strictly larger than $\bar{X}_{j,s}^{\star,\min}$, so

$$\bar{X}_{j,s}^{\star,\min} = \min_{a \in \mathcal{A}} \bar{X}_{j,s,a}.$$

All inverse coordinates are positive, and the product identity in Equation (D.5) therefore gives

$$\bar{X}_{j,s}^{\star,\min} \leq \left(\prod_{a \in \mathcal{A}} \bar{X}_{j,s,a} \right)^{1/A} = 1.$$

Because $\bar{Z}_{j,s}$ contains the term $(\bar{X}_{j,s}^{\star,\min})^{-1}$, one has $\bar{Z}_{j,s} \geq (\bar{X}_{j,s}^{\star,\min})^{-1}$. Hence, for every $b \notin \mathcal{A}_j^\star(s)$,

$$\begin{aligned} \bar{\pi}_{u,j}(b \mid s) = \frac{1}{\bar{X}_{j,s,b}\bar{Z}_{j,s}} &\leq \frac{\bar{X}_{j,s}^{\star,\min}}{\bar{X}_{j,s,b}} \leq \frac{\bar{X}_{j,s}^{\star,\min}}{\bar{X}_{j,s}^{\star,\min} + L_j^{\mathrm{sep}}} \\ &\leq \frac{1}{1 + L_j^{\mathrm{sep}}}, \end{aligned}$$

where the last inequality uses $\bar{X}_{j,s}^{\star,\min} \leq 1$. Since rewards lie in $[0,1]$, the optimality gaps obey $0 \leq \Delta_j(s,a) \leq \overline{H}_j$, and they vanish on $\mathcal{A}_j^\star(s)$. Consequently,

$$\begin{aligned} \ell_j^{\bar{\pi}_u}(s) &= \sum_{b \notin \mathcal{A}_j^\star(s)} \bar{\pi}_{u,j}(b \mid s) \Delta_j(s,b) \\ &\leq \frac{A\overline{H}_j}{1 + L_j^{\mathrm{sep}}} \leq \frac{A\overline{H}_j}{L_j^{\mathrm{sep}}} = \varepsilon_{\mathrm{tail}}. \end{aligned} \tag{D.13}$$

If every action in the block is optimal, the sum defining $\ell_j^{\bar{\pi}_u}(s)$ is empty, so $\ell_j^{\bar{\pi}_u}(s) = 0 \leq \varepsilon_{\mathrm{tail}}$.

Bellman tail. For $j \in \{h+1, \dots, H\}$, define

$$M_j := \max_{x \in \mathcal{S}_j^+} e_j^{\bar{\pi}_u}(x).$$

At the terminal stage, Equation (G.1) reduces to $e_H^{\bar{\pi}_u} = \ell_H^{\bar{\pi}_u}$, so Equation (D.13) yields $M_H \leq \varepsilon_{\mathrm{tail}}$. If $s \in \mathcal{S}_j^+$ and $P_j(x \mid s,a) > 0$, then appending the edge (s,a,x) to a positive-probability witness path for s shows that $x \in \mathcal{S}_{j+1}^+$. Therefore the same recursion and Equation (D.13) give, for $h+1 \leq j < H$,

$$M_j \leq \varepsilon_{\mathrm{tail}} + M_{j+1}.$$

Backward induction now gives

$$M_{h+1} \leq (H-h)\varepsilon_{\mathrm{tail}} = \frac{H-h}{H-1}\frac{\Delta}{4} \leq \frac{\Delta}{4}. \tag{D.14}$$

Current-policy action gap. For any $a \in \mathcal{A}$, the optimal and current-policy Bellman equations imply

$$Q_h^{\bar{\pi}_u}(s,a) = Q_h^*(s,a) - \sum_{x \in \mathcal{S}_{h+1}} P_h(x \mid s,a) e_{h+1}^{\bar{\pi}_u}(x).$$

Thus, for $o \in \mathcal{A}_h^\star(s)$ and $b \notin \mathcal{A}_h^\star(s)$,

$$\begin{aligned} Q_h^{\bar{\pi}_u}(s,o) - Q_h^{\bar{\pi}_u}(s,b) = \Delta_h(s,b) - \sum_x P_h(x \mid s,o) e_{h+1}^{\bar{\pi}_u}(x) \\ + \sum_x P_h(x \mid s,b) e_{h+1}^{\bar{\pi}_u}(x). \end{aligned}$$

The value errors are nonnegative, and every child state reached from s with positive probability is structurally reachable. Hence Equation (D.14) gives

$$Q_h^{\bar{\pi}_u}(s,o) - Q_h^{\bar{\pi}_u}(s,b) \geq \Delta_h(s,b) - M_{h+1} \geq \Delta - \frac{\Delta}{4} = \frac{3\Delta}{4},$$

where $\Delta_h(s,b) \geq \Delta$ follows from $(h,s,b) \in \mathcal{I}_+$. This proves Equation (D.12).

Finally, if $h = H$, then $Q_H^{\bar{\pi}_u} = r_H = Q_H^*$, and therefore

$$Q_H^{\bar{\pi}_u}(s,o) - Q_H^{\bar{\pi}_u}(s,b) = \Delta_H(s,b) \geq \Delta,$$

with no downstream assumption. □

We use the conservative recovery gap

$$g_{\mathrm{rec}} := \frac{\Delta}{2}. \tag{D.15}$$

D.4 Quantitative Occupancy Transmission

For a fixed window beginning at $U \geq 0$, define cumulative state occupancy by

$$\mathfrak{a}_{h,s}(U,t) := \int_U^{U+t} d_h^{\bar{\pi}_u}(s)\,\mathrm{d}u, \qquad t \geq 0. \tag{D.16}$$

This is an integral of mean-flow occupancy probabilities. The policy at U continues the mean flow initialized at time zero. The next two results also hold when $\mathcal{I}_+ = \varnothing$, with $p = p_{\min}$.

Lemma D.4 (Action Suppression and One-Edge Occupancy Transmission)**.** *Fix j, $s \in \mathcal{S}_j^+$, $a \in \mathcal{A}$, and $U \geq 0$. Put*

$$C_j(U) := 1 + \frac{\overline{H}_j U}{A}, \qquad w(u) := \int_U^u d_j^{\bar{\pi}_q}(s) \bar{\pi}_{q,j}(a \mid s)\,\mathrm{d}q.$$

For every finite $u \geq U$,

$$\log \frac{1}{\bar{\pi}_{u,j}(a \mid s)} \leq \log A + A \log C_j(U) + A\overline{H}_j w(u). \tag{D.17}$$

If $j < H$ and $P_j(x \mid s,a) > 0$, then every $t \geq 0$ satisfies

$$\mathfrak{a}_{j+1,x}(U,t) \geq \frac{p}{A\overline{H}_j} \log\left(1 + \frac{\overline{H}_j \mathfrak{a}_{j,s}(U,t)}{C_j(U)^A}\right). \tag{D.18}$$

Replacing $\overline{H}_j$ throughout this display by H, and $C_j(U)$ by $C(U) := 1 + HU/A$, also gives a valid bound.

Proof. The globally Lipschitz and bounded mean field in Lemmas G.3 and G.5 has a unique solution at every finite time. Every finite logit gives a strictly positive softmax probability. Along a positive-probability witness path, the occupancy recursion therefore gives $d_j^{\bar{\pi}_u}(s) > 0$. Thus $w'(u) > 0$ on each finite interval, and we may use w as a variable on its attained range.

Suppress (j, s), and write V, Q_i for the values at the current global time. The inverse-coordinate identity gives

$$\frac{\mathrm{d}\bar{X}_i}{\mathrm{d}w} = \frac{d(V - Q_i)/\bar{Z}}{d/(\bar{X}_a\bar{Z})} = \bar{X}_a(V - Q_i). \tag{D.19}$$

The same $\bar{X}_a$ multiplies every coordinate. Since $|V - Q_i| \le \overline{H}_j$ and $\bar{X}_i(U) \le C_j(U)$, first taking $i = a$ yields

$$\frac{\mathrm{d}}{\mathrm{d}w}\left(e^{-\overline{H}_j w}\bar{X}_a\right) \le 0, \qquad \bar{X}_a(w) \le C_j(U)e^{\overline{H}_j w}.$$

For any other i, integration of Equation (D.19) then gives

$$\bar{X}_i(w) \le C_j(U) + \int_0^w \overline{H}_j C_j(U)e^{\overline{H}_j z}\,\mathrm{d}z = C_j(U)e^{\overline{H}_j w}.$$

The product identity is needed to control the whole softmax denominator:

$$\bar{X}_i^{-1} = \prod_{c\neq i}\bar{X}_c, \qquad \bar{X}_a\bar{Z} = \bar{X}_a\sum_i\prod_{c\neq i}\bar{X}_c \le AC_j(U)^A e^{A\overline{H}_j w}.$$

Taking logarithms proves Equation (D.17).

Set $w_t = w(U + t)$. Integrating the inverse action probability gives

$$\begin{aligned}\mathfrak{a}_{j,s}(U,t) &= \int_0^{w_t}\frac{1}{\bar{\pi}_{u(w),j}(a \mid s)}\,\mathrm{d}w \\ &\le \int_0^{w_t} AC_j(U)^A e^{A\overline{H}_j w}\,\mathrm{d}w = \frac{C_j(U)^A}{\overline{H}_j}\left(e^{A\overline{H}_j w_t} - 1\right).\end{aligned} \tag{D.20}$$

Meanwhile, the occupancy recursion gives

$$\mathfrak{a}_{j+1,x}(U,t) \ge P_j(x \mid s, a)w_t \ge pw_t.$$

Solving the previous display for w_t proves Equation (D.18). The proof with the common range bound H is identical, starting with $|V - Q_i| \le H$ and $\bar{X}_i(U) \le C(U)$. At $t = 0$ both sides of the occupancy bound are zero. □

Theorem D.5 (Explicit Occupancy Time Along Reachable Paths)**.** *Let* $\exp^{\circ m}$ *denote* m *compositions of the exponential function, with* $\exp^{\circ 0}(z) = z$*, and define*

$$\beta_U := \log\frac{A}{p} + A\log\left(1 + \frac{HU}{A}\right), \qquad \mathsf{T}_h^{\mathrm{occ}}(U, D) := \frac{\exp^{\circ(h-1)}(AHD/p + \beta_U) - \beta_U}{AH}. \tag{D.21}$$

For every $U \ge 0$*,* $D > 0$*, and* $h \in [H]$*, simultaneously for all* $s \in \mathcal{S}_h^+$*,*

$$\mathfrak{a}_{h,s}\left(U, \mathsf{T}_h^{\mathrm{occ}}(U, D)\right) \ge D. \tag{D.22}$$

Proof. Fix a witness path $(s_1, a_1, \ldots, s_{h-1}, a_{h-1}, s_h = s)$ and a common time window $[U, U + t]$. Write $y_j = (AH/p)\mathfrak{a}_{j,s_j}(U, t)$ and $C(U) = 1 + HU/A$. The common-H version of Equation (D.18) implies

$$y_{j+1} \ge \log\left(1 + \frac{py_j}{AC(U)^A}\right), \qquad y_j \le e^{\beta_U}(e^{y_{j+1}} - 1).$$

Since $\beta_U \ge 0$ and $\beta_U \le e^{\beta_U}$,

$$y_j + \beta_U \le e^{y_{j+1}+\beta_U} - e^{\beta_U} + \beta_U \le e^{y_{j+1}+\beta_U}.$$

The root occupancy is $\rho(s_1)t \ge pt$, so $y_1 \ge AHt$. Composing the preceding inequalities along the path gives

$$AHt + \beta_U \le \exp^{\circ(h-1)}(y_h + \beta_U).$$

At the time in Equation (D.21), strict monotonicity of the composed exponential implies $y_h \ge AHD/p$, proving the target. For $h = 1$, the formula reduces to D/p and the root calculation applies directly. The time bound and the argument hold uniformly over the choice of witness. All witnesses use the same policy and time window, so the conclusion holds simultaneously over the finitely many reachable states at that stage. □

The displayed time has $h - 1$ exponential compositions for one specified occupancy target. The readiness schedule below applies occupancy delivery repeatedly with targets depending on previous phases.

D.5 Occupancy along a Reachable Path

The next lemma quantifies how much wall-clock time suffices to deliver a prescribed amount of occupancy to every reachable state.

Lemma D.6 (Occupancy across One Edge)**.** *Fix a starting time U, a reachable state $s \in \mathcal{S}_j^+$, an action $a \in \mathcal{A}$, and a target $D > 0$. Define*

$$v_{s,a}(u) := p \int_U^u d_j^{\bar{\pi}_q}(s)\bar{\pi}_{q,j}(a \mid s)\,\mathrm{d}q. \tag{D.23}$$

An additional state occupancy

$$F_j(U, D) := \frac{AD}{p}\left[\left(1 + \frac{\overline{H}_j U}{A}\right)\exp\left(\frac{\overline{H}_j D}{p}\right)\right]^A \tag{D.24}$$

forces $v_{s,a}$ to reach D. If $P_j(x \mid s, a) > 0$, reaching $v_{s,a} = D$ delivers at least D additional occupancy to x.

Proof. Write $C_j(U) = 1 + \overline{H}_j U/A$. The integrated bound Equation (D.20) shows that the smaller parent-occupancy target

$$F_j^{\mathrm{int}}(U, D) := \frac{C_j(U)^A}{\overline{H}_j}\left(e^{A\overline{H}_j D/p} - 1\right) \tag{D.25}$$

forces $w(u) \geq D/p$, or equivalently $v_{s,a}(u) \geq D$. In detail, $v_{s,a}(u) < D$ would make the right side of Equation (D.20) strictly smaller than $F_j^{\mathrm{int}}(U, D)$, contradicting attainment of that parent-occupancy target.

For $z = A\overline{H}_j D/p > 0$,

$$\frac{F_j^{\mathrm{int}}(U, D)}{F_j(U, D)} = \frac{1 - e^{-z}}{z} \leq 1,$$

since $1 - e^{-z} = \int_0^z e^{-q}\,\mathrm{d}q \leq z$. Thus the target $F_j(U, D)$ in the statement also suffices. Finally,

$$\int_U^u d_{j+1}^{\bar{\pi}_q}(x)\,\mathrm{d}q \geq P_j(x \mid s, a)\int_U^u d_j^{\bar{\pi}_q}(s)\bar{\pi}_{q,j}(a \mid s)\,\mathrm{d}q \geq v_{s,a}(u),$$

which proves the child-state guarantee. □

We retain F_j in the readiness constants to give one fixed recursion throughout the theorem. The integrated target is smaller on each edge; the closed-form time in Equation (D.21) additionally replaces each stage range by H and enlarges the composition.

For a target D at stage h, define recursively

$$q_h = D, \qquad q_j = F_j(U, q_{j+1}) \quad (j = h-1, \ldots, 1), \qquad \mathsf{D}_h(U, D) := \frac{q_1}{p}. \tag{D.26}$$

Every positive root mass is at least p. Applying Lemma D.6 along a positive-probability witness path gives the following explicit induction. Fix $s_h \in \mathcal{S}_h^+$ and a witness $(s_1, a_1, s_2, \ldots, a_{h-1}, s_h)$. Over the common interval $[U, U + q_1/p]$, the root occupancy obeys

$$\int_U^{U+q_1/p} d_1^{\bar{\pi}_u}(s_1)\,\mathrm{d}u = \rho(s_1)\frac{q_1}{p} \geq q_1 = F_1(U, q_2).$$

The one-edge lemma therefore delivers at least q_2 occupancy to s_2. Inductively, if the same interval has delivered $q_j = F_j(U, q_{j+1})$ occupancy to s_j, applying the lemma to (s_j, a_j) delivers at least q_{j+1} occupancy to s_{j+1}. At the last state this is $q_h = D$. All applications use the same start time U and the same enclosing wall-time interval. Since the same occupancy bounds apply to each witness, this argument holds simultaneously for the finitely many fixed witnesses. Thus the wall-time increment $\mathsf{D}_h(U, D) = q_1/p$ delivers at least D additional occupancy to every state in $\mathcal{S}_h^+$, although individual hitting times may differ.

D.6 Explicit Backward Schedule

Assume first that $H \geq 2$. Starting from $U_{H+1} := 0$, define for $h = H, H-1, \ldots, 1$

$$C_h := 1 + \frac{\overline{H}_h U_{h+1}}{A}, \tag{D.27}$$

$$\omega_h := \frac{C_h + L_h^{\mathrm{sep}} + 2}{g_{\mathrm{rec}}}, \tag{D.28}$$

$$T_h := \frac{(C_h + \overline{H}_h \omega_h)^A - C_h^A}{\overline{H}_h}, \tag{D.29}$$

$$U_h := U_{h+1} + \mathsf{D}_h(U_{h+1}, T_h), \qquad u_{\mathrm{ready}} := U_1. \tag{D.30}$$

To convert state occupancy bounds into local clock accumulation quantitatively, let $q = \chi_{h,s}(U + t) - \chi_{h,s}(U)$ and $C = 1 + HU/A$. The inverse-coordinate flow (D.7) bounds coordinates by $\bar{X}_i \leq C + Hq$, giving $\bar{Z} \leq A(C + Hq)^{A-1}$. Integrating $d\mathfrak{a}_{h,s}/dq = \bar{Z}$ yields the cumulative occupancy conversion

$$\mathfrak{a}_{h,s}(U, t) \leq \frac{(C + Hq)^A - C^A}{H}. \tag{D.31}$$

Consequently, delivering cumulative occupancy at least $[(C + H\omega)^A - C^A]/H$ guarantees that the accumulated local clock reaches $q \geq \omega$.

Proof of Theorem D.1. We prove the stronger statement that, for every $u \geq U_h$, the separation holds for each fixed $o \in \mathcal{A}_h^\star(s)$ and $b \notin \mathcal{A}_h^\star(s)$. This avoids differentiating a time-varying minimizer. Proceed backward in h. At the start of phase h, stages $h+1, \ldots, H$ are ready and persistent by induction. Hence Lemma D.3 gives a current-policy gap of at least g_{rec} throughout the phase for every optimal–nonoptimal pair at stage h.

For a fixed stage-h state s, reset its phase clock at U_{h+1}:

$$\chi_s^{[h]}(u) := \int_{U_{h+1}}^{u} \frac{d_h^{\bar{\pi}_q}(s)}{\bar{Z}_{h,s}(q)}\, \mathrm{d}q.$$

Before this clock reaches ω_h, Equations (D.7) and (D.27) imply $\bar{X}_{h,s,a} \leq C_h + \overline{H}_h \chi_s^{[h]}$ for every a. The gauge then gives

$$\bar{Z}_{h,s} \leq A(C_h + \overline{H}_h \chi_s^{[h]})^{A-1}.$$

If $\mathfrak{a}_s := \int_{U_{h+1}}^{u} d_h^{\bar{\pi}_q}(s)\, \mathrm{d}q$ is the added state occupancy, integrating $\mathrm{d}\mathfrak{a}_s = \bar{Z}_{h,s}\, \mathrm{d}\chi_s^{[h]}$ shows that, while $\chi_s^{[h]} \leq \omega_h$,

$$\mathfrak{a}_s \leq \frac{(C_h + \overline{H}_h \chi_s^{[h]})^A - C_h^A}{\overline{H}_h}.$$

Consequently, $\mathfrak{a}_s \geq T_h$ forces $\chi_s^{[h]} \geq \omega_h$. The multi-hop delivery time in Equation (D.30) supplies T_h to every reachable stage-h state.

At the start of the phase, any pair obeys $\bar{X}_b - \bar{X}_o \geq -C_h$. By Equations (D.8) and (D.28), at that block's own clock hitting time,

$$\bar{X}_b - \bar{X}_o \geq -C_h + g_{\mathrm{rec}}\omega_h = L_h^{\mathrm{sep}} + 2.$$

The transient cap above is used only until the individual block completes. After its completion, the already-ready downstream tail and Lemma D.3 keep every pairwise derivative nonnegative, so the separation persists. The induction starts at $h = H$, where action values are the fixed terminal rewards, and ends at $h = 1$. □

When $H = 1$, set $L_1^{\mathrm{sep}} = 0$, $\omega_1 = 2/\Delta$, $T_1 = (1 + \omega_1)^A - 1$, and $u_{\mathrm{ready}} = T_1/p$. The root occupancy supplies at least T_1, so the preceding integral bound with $C_1 = 1$ and $\overline{H}_1 = 1$ gives a local clock of at least ω_1. Each initial inverse-coordinate difference is exactly zero, and the terminal

action-value gap is at least Δ. Therefore each difference is at least $\Delta\omega_1 = 2$ at u_{ready}, and remains nondecreasing thereafter.

All constants above are explicit and finite. Repeated composition of $F_j(U, D)$ first along witness paths and then across backward phases can make u_{ready} an iterated exponential tower in H and p^{-1}. We use these constants as a qualitative existence certificate.

E Lower Bounds and Diagnostic Counterexamples

This appendix proves the horizon-dependent lower bounds first, followed by the terminal-bandit and fixed-step lower bounds. It ends with two examples that explain why the upper-bound proof needs downstream readiness and optimal-set aggregation. Throughout, K counts episodes, θ_k and π_k are the logits and policy used in episode k, and $\theta_1 = 0$. All constructions use deterministic rewards and deterministic transitions, within the setting of assumption G.1. The sampled update and regret are those in Equations (N.4) and (N.11); the mean flow is identified separately by barred notation.

E.1 A Common Chain for the Mean-Flow and Sampled Lower Bounds

We first prove the two claims needed by both tower lower bounds: continuing has an unfavorable current-policy value until the terminal policy improves, and this period incurs a constant instantaneous regret. We then derive the mean-flow occupancy inequality. The sampled proof replaces monotonicity of the logits by an explicit martingale bound and derives its own occupancy inequality.

Our construction shares the mechanism identified by Li et al. (2023), where optimal-action probabilities remain suppressed until other states improve, generating recursively amplified delays. Here a chain starting at its root in a finite-horizon MDP yields a lower bound on regret.

Fix $H \geq 4$. There is one active state s_h at each stage. At s_h, for $h < H$, action a^{c} continues to s_{h+1} with zero immediate reward, while action a^{s} gives reward

$$v_h = \frac{5}{8} + \frac{H - 1 - h}{8(H - 2)} \tag{E.1}$$

and enters a zero-reward tail. At s_H, the two actions give rewards one and zero. The optimal policy always continues and then selects the rewarding terminal action. This MDP has deterministic transitions, two actions at every state, and $\Delta_{\min} = 1/4$.

Let p_h denote the continue probability, with p_H denoting the rewarding terminal-action probability. Until p_H reaches $9/16$, the current value of every continue action is below its safe alternative. Define

$$\delta_H := \frac{1}{8(H - 2)}, \qquad \lambda_H := \frac{\delta_H}{2}, \qquad \ell := \log(9/7). \tag{E.2}$$

The disadvantage of continuing is at least δ_H, although its *optimal* action value is larger. The regret remains at least $1/4$ throughout this period.

Formally, each stage after the first also has an absorbing state. Every action at an absorbing state has reward zero and moves to the next absorbing state. A safe action at s_h moves to that zero-reward tail after paying v_h. Every state has the same two-action set, and the initial distribution is concentrated on s_1. A trajectory receives at most one nonzero reward, so every return-to-go lies in $[0, 1]$. All nonzero transition probabilities equal one. The optimal value at every active state is one; the gap of a safe action is $1 - v_h \in [1/4, 3/8]$, and the terminal zero-reward action has gap one. Thus the minimum positive Bellman gap among reachable actions is exactly $1/4$.

For a policy π, write p_h for its continue probability at s_h, or its rewarding-action probability when $h = H$. Its active-state occupancy and values satisfy

$$d_h = \prod_{j<h} p_j, \qquad V_H = p_H, \qquad V_h = (1 - p_h)v_h + p_h V_{h+1} \quad (h < H). \tag{E.3}$$

If $p_H < 9/16$, then $v_{H-1} - V_H > 1/16 \geq \delta_H$, since $H \geq 4$. Backward induction gives $V_h \leq v_h$ for $h < H$ and

$$v_h - V_{h+1} \geq \delta_H \quad (h < H), \qquad \mathcal{R}(\pi) = 1 - V_1 \geq 1 - v_1 = \frac{1}{4}. \tag{E.4}$$

For $h \le H-2$, the induction uses $v_h - v_{h+1} = \delta_H$. These inequalities hold for every policy below the terminal threshold.

E.2 Proof of the Mean-Flow Tower Delay

Theorem E.1 (Mean-Flow Tower Delay)**.** *On the chain above, let $\bar\theta$ solve $d\bar\theta/du = g(\bar\theta)$ from zero initialization, and let $\bar\tau$ be the first time its terminal rewarding-action probability reaches $9/16$. Define*

$$\mathsf{B}_H = 2\ell, \qquad \mathsf{B}_h = \frac{2}{\lambda_H}\big(e^{\lambda_H \mathsf{B}_{h+1}} - 1\big), \quad h = H-1, \dots, 1. \tag{E.5}$$

Then $\bar\tau \ge \mathsf{B}_1$, with the inequality understood to include $\bar\tau = \infty$, and

$$\mathcal{R}(\bar\pi_u) \ge \frac{1}{4} \quad (0 \le u < \mathsf{B}_1), \qquad \int_0^U \mathcal{R}(\bar\pi_u)\, du \ge \frac{U}{4} \quad (0 \le U \le \mathsf{B}_1). \tag{E.6}$$

For sufficiently large H, B_1 is bounded below by an exponential tower with $H - O(\log H)$ exponentiation levels.

To obtain a lower bound on the time of terminal improvement, we first bound continue probabilities in terms of cumulative state occupancy. Each edge then limits the occupancy delivered to its child. Inverting these bounds backward along the chain gives the required root time.

Proof of Theorem E.1. Let $\bar p_h(u)$ be the active action probabilities of $\bar\pi_u = \pi_{\bar\theta(u)}$, let $\bar d_h(u) = \prod_{j<h} \bar p_j(u)$, and let $\bar z_h(u)$ be the continue-minus-safe logit difference, or the rewarding-minus-zero-reward logit difference at the terminal state. The exact softmax mean-gradient identity gives, before $\bar\tau$,

$$\frac{d\bar z_h}{du} = 2\bar d_h \bar p_h (1-\bar p_h)(V_{h+1}^{\bar\pi_u} - v_h) \le -2\delta_H \bar d_h \bar p_h (1-\bar p_h), \qquad h < H, \tag{E.7}$$

$$\frac{d\bar z_H}{du} = 2\bar d_H \bar p_H (1-\bar p_H) \le \frac{1}{2}\bar d_H. \tag{E.8}$$

In particular, each nonterminal $\bar p_h$ decreases from $1/2$. Define the cumulative occupancy

$$\bar T_h(u) := \int_0^u \bar d_h(q)\, dq. \tag{E.9}$$

This occupancy integral measures time available at the active state; it is different from the inverse-coordinate local clock $\chi_{h,s}$ used in the readiness proof.

To bound how much occupancy reaches the next stage, define $\bar q_h = (1-\bar p_h)/\bar p_h = e^{-\bar z_h}$ for $h < H$. Using Equation (E.7) and $\bar p_h \le 1/2$ gives

$$\frac{d\bar q_h}{du} \ge 2\delta_H \bar d_h (1-\bar p_h)^2 \ge \lambda_H \bar d_h. \tag{E.10}$$

Since $\bar q_h(0) = 1$, integration implies $\bar p_h(u) \le (2 + \lambda_H \bar T_h(u))^{-1}$. The identity $\bar d_{h+1} = \bar d_h \bar p_h$ therefore yields

$$\begin{aligned} \bar T_{h+1}(u) &\le \int_0^u \frac{\bar d_h(q)}{2 + \lambda_H \bar T_h(q)}\, dq \\ &= \frac{1}{\lambda_H} \log\left(1 + \frac{\lambda_H \bar T_h(u)}{2}\right). \end{aligned} \tag{E.11}$$

The occupancy-delivery bound holds for every $u < \bar\tau$, and by continuity at every finite $\bar\tau$.

If $\bar\tau < \infty$, then $\bar z_H(\bar\tau) = \ell$. Integrating Equation (E.8) gives $\bar T_H(\bar\tau) \ge 2\ell = \mathsf{B}_H$. Inverting Equation (E.11) successively gives $\bar T_h(\bar\tau) \ge \mathsf{B}_h$ for $h = H-1, \dots, 1$. At the root, $\bar d_1 = 1$, so $\bar\tau = \bar T_1(\bar\tau) \ge \mathsf{B}_1$. If $\bar\tau = \infty$, the same hitting-time conclusion is immediate. Equation (E.4) now proves Equation (E.6).

For completeness, the recursion has a quantitative tower lower bound. Set $y_h = \lambda_H \mathsf{B}_h$. Then $y_H = 2\lambda_H \ell$ and $y_h = 2(e^{y_{h+1}} - 1) \geq 2y_{h+1}$. Let

$$m_H^{\mathrm{flow}} := \left\lceil \log_2 \frac{1}{\lambda_H \ell} \right\rceil, \qquad r_H^{\mathrm{flow}} := H - 1 - m_H^{\mathrm{flow}}. \tag{E.12}$$

After m_H^{flow} predecessor recursions the scaled value is at least two. For $y \geq 2$, $2(e^y - 1) \geq e^y$. Thus, whenever $r_H^{\mathrm{flow}} \geq 0$, defining $E_0 = 2$ and $E_{r+1} = e^{E_r}$ gives

$$\mathsf{B}_1 \geq \frac{E_{r_H^{\mathrm{flow}}}}{\lambda_H}. \tag{E.13}$$

Since $m_H^{\mathrm{flow}} = O(\log H)$, this proves the claimed tower height. □

E.3 Proof of the Sampled Tower Delay

For the sampled lower bound, define the step cap and required learning time by

$$\begin{aligned} c_* &:= \frac{1}{1 + e^{1/4}}, \qquad D_* := e^{-5/16}, \\ f(x) &:= x + D_*(e^x - 1), \qquad U_H := \frac{f^{\circ(H-1)}(\delta_H c_* \ell)}{\delta_H}, \\ \eta_H^{\mathrm{cap}} &:= \min\left\{ \frac{\delta_H}{32 \log(4H)}, \frac{\ell}{12 \log 4} \right\}. \end{aligned} \tag{E.14}$$

Here $f^{\circ m}$ denotes m successive applications of f.

Theorem E.2 (Sampled Tower Delay)**.** *Fix the chain in Section E.1, and let $0 < \eta_{\max} \leq \eta_H^{\mathrm{cap}}$. Run the REINFORCE update from zero initialization, allowing the scalar step η_k to be chosen using the history before episode k, subject to $0 \leq \eta_k \leq \eta_{\max}$ almost surely. For every deterministic integer $K \geq 1$ with $\eta_{\max} K \leq U_H/2$,*

$$\mathbb{E}[\mathrm{Reg}_K] \geq \frac{K}{8}. \tag{E.15}$$

The MDP and U_H depend only on H, and are fixed before the step cap, scheduling rule, and episode budget are chosen.

The target is to show that terminal improvement requires root learning time at least U_H on an event of probability at least $1/2$. The proof has three parts. A nonterminal martingale bound controls upward fluctuations of each continue logit. A terminal martingale bound supplies a positive amount of required terminal occupancy. A pathwise relation then propagates this requirement backward to the root.

E.3.1 Stopped Iterates and an Exponential Martingale Inequality

Use the standing episode indexing: θ_k and π_k are used to sample episode k, and $\theta_1 = 0$. Write $p_{k,h}$, $d_{k,h}$, and $z_{k,h}$ for the corresponding active probabilities, occupancies, and logit differences. Let $\mathcal{F}_{k-1}$ contain the history before episode k, together with any independent random seed used by the step rule. The step η_k is $\mathcal{F}_{k-1}$-measurable. Define the number of completed updates at the first terminal crossing by

$$\tau := \inf\{n \geq 0 : p_{n+1,H} \geq 9/16\}. \tag{E.16}$$

In particular, $\{k \leq \tau\}$ is known before episode k. For $n \geq 0$, define

$$\begin{aligned} \widetilde{z}_h(n) &:= z_{(n \wedge \tau)+1,h}, \\ T_h(n) &:= \sum_{k=1}^{n} \eta_k \mathbf{1}\{k \leq \tau\} d_{k,h}, \\ C_h(n) &:= \sum_{k=1}^{n} \eta_k \mathbf{1}\{k \leq \tau\} d_{k,h} p_{k,h}(1 - p_{k,h}). \end{aligned} \tag{E.17}$$

The stopped coordinate includes the update that first crosses the threshold. All subsequent increments in these definitions vanish.

We record the concentration argument to make the constants explicit. Suppose D_k are martingale differences with $|D_k| \le B$, and set $M_n = \sum_{k=1}^{n} D_k$ and $V_n = \sum_{k=1}^{n} \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}]$. For $0 < \xi < 3/B$, put

$$\psi_B(\xi) := \frac{\xi^2}{2(1 - \xi B/3)}. \tag{E.18}$$

The power series of the exponential and $m! \ge 2 \cdot 3^{m-2}$ for $m \ge 2$ give

$$\mathbb{E}[e^{\xi D_k} \mid \mathcal{F}_{k-1}] \le 1 + \psi_B(\xi)\mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}] \le \exp\big(\psi_B(\xi)\mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}]\big). \tag{E.19}$$

Consequently $\exp(\xi M_n - \psi_B(\xi)V_n)$ is a nonnegative supermartingale starting at one. Stopping at its first crossing of e^t and at a deterministic finite horizon gives crossing probability at most e^{-t}; increasing the horizon proves the time-uniform bound. In particular, choosing $\xi = a/(v + Ba/3)$ yields, for $a, v > 0$,

$$\mathbb{P}(\exists n : M_n \ge a,\ V_n \le v) \le \exp\left(-\frac{a^2}{2(v + Ba/3)}\right). \tag{E.20}$$

This derivation will be applied under the actual trajectory law.

E.3.2 The Nonterminal Logit Boundary

Fix $h < H$ and let $Y_{k,h} := \widetilde{z}_h(k) - \widetilde{z}_h(k-1)$. If stage h is visited before stopping, a continue action gives logit increment $2\eta_k(1 - p_{k,h})G_{k,h}$, while a safe action gives $-2\eta_k p_{k,h} G_{k,h}$. Here $0 \le G_{k,h} \le 1$ is the observed return-to-go. The increment is zero on all other events. The conditional mean of the return after continuing is $V_{h+1}^{\pi_k}$, whereas the safe return is v_h. Thus Equation (E.4) gives

$$\mathbb{E}[Y_{k,h} \mid \mathcal{F}_{k-1}] \le -2\delta_H \eta_k \mathbf{1}\{k \le \tau\} d_{k,h} p_{k,h}(1 - p_{k,h}), \tag{E.21}$$

$$\mathbb{E}[Y_{k,h}^2 \mid \mathcal{F}_{k-1}] \le 4\eta_k^2 \mathbf{1}\{k \le \tau\} d_{k,h} p_{k,h}(1 - p_{k,h}). \tag{E.22}$$

For the second line, condition separately on the two actions and use $p(1-p)^2 + (1-p)p^2 = p(1-p)$.

Let $M_h(n)$ be the sum of the centered increments $Y_{k,h}$. Since $|Y_{k,h}| \le 2\eta_{\max}$ and the absolute conditional mean is at most $\eta_{\max}/2$, its jumps are bounded by $B = 5\eta_{\max}/2$. Its predictable variance satisfies

$$\langle M_h \rangle_n \le 4\eta_{\max} C_h(n). \tag{E.23}$$

Choose $\xi = \delta_H/(8\eta_{\max})$ in Equation (E.18). Direct substitution, using $\delta_H \le 1/16$, gives $4\eta_{\max}\psi_B(\xi) \le \xi\delta_H$. The supermartingale crossing bound therefore implies, for every $b > 0$,

$$\mathbb{P}\big(\exists n : M_h(n) > \delta_H C_h(n) + b\big) \le \exp\left(-\frac{\delta_H b}{8\eta_{\max}}\right). \tag{E.24}$$

Zero initialization and Equation (E.21) give $\widetilde{z}_h(n) \le M_h(n) - 2\delta_H C_h(n)$. Hence, except on the event bounded in Equation (E.24),

$$\widetilde{z}_h(n) \le b - \delta_H C_h(n) \qquad \text{for every } n \ge 0. \tag{E.25}$$

E.3.3 A Terminal Occupancy Requirement

At stage H, the stopped increment of the logit difference between the rewarding and zero-reward actions has conditional mean $2\eta_k \mathbf{1}\{k \le \tau\} d_{k,H} p_{k,H}(1 - p_{k,H})$, which is at most $\eta_k \mathbf{1}\{k \le \tau\} d_{k,H}/2$. Its conditional second moment is $4\eta_k^2 \mathbf{1}\{k \le \tau\} d_{k,H} p_{k,H}(1 - p_{k,H})^2$, which is at most $\eta_k^2 \mathbf{1}\{k \le \tau\} d_{k,H}$. Thus the centered terminal martingale M_H has jumps bounded by $5\eta_{\max}/2$ and satisfies

$$\langle M_H \rangle_n \le \eta_{\max} T_H(n), \qquad \widetilde{z}_H(n) \le \tfrac{1}{2} T_H(n) + M_H(n). \tag{E.26}$$

If $\tau < \infty$ and $T_H(\tau) < \ell$, then $\widetilde{z}_H(\tau) \ge \ell$ and $M_H(\tau) > \ell/2$, while $\langle M_H \rangle_\tau < \eta_{\max}\ell$. Apply Equation (E.20) with $a = \ell/2$, $v = \eta_{\max}\ell$, and $B = 5\eta_{\max}/2$. This gives

$$\mathbb{P}(\tau < \infty,\ T_H(\tau) < \ell) \le e^{-3\ell/(34\eta_{\max})} \le e^{-\ell/(12\eta_{\max})}. \tag{E.27}$$

E.3.4 Backward Propagation of the Occupancy Requirement

Set

$$b := \frac{8\eta_{\max}}{\delta_H}\log(4H), \qquad c_b := \frac{1}{1+e^b}, \qquad c := \delta_H c_b, \qquad D := e^{-b-c\eta_{\max}}. \tag{E.28}$$

Let $\mathcal{G}$ be the event that Equation (E.25) holds at every nonterminal stage and that $T_H(\tau) \geq \ell$ whenever $\tau < \infty$. A union bound over $H-1$ stages, followed by Equation (E.27), gives

$$\mathbb{P}(\mathcal{G}) \geq 1 - \frac{H-1}{4H} - e^{-\ell/(12\eta_{\max})} \geq \frac{1}{2}. \tag{E.29}$$

The last inequality uses $\eta_{\max} \leq \eta_H^{\mathrm{cap}}$.

On $\mathcal{G}$, for every $k \leq \tau$, $z_{k,h} \leq b - \delta_H C_h(k-1) \leq b$. It follows that $1 - p_{k,h} \geq c_b$. Using $d_{k,h+1} = d_{k,h}p_{k,h}$ in the clock sums therefore gives

$$C_h(n) \geq c_b T_{h+1}(n), \qquad \frac{1}{p_{k,h}} \geq 1 + \exp(-b + cT_{h+1}(k-1))\,. \tag{E.30}$$

Suppose now that $\tau < \infty$. Put $t_k = T_{h+1}(k)$, and $\Delta t_k = t_k - t_{k-1}$ for $1 \leq k \leq \tau$. The mesh satisfies $0 \leq \Delta t_k \leq \eta_{\max}$, and the exact occupancy identity gives

$$\begin{aligned} T_h(\tau) &= \sum_{k=1}^{\tau} \frac{\Delta t_k}{p_{k,h}} \\ &\geq T_{h+1}(\tau) + e^{-b}\sum_{k=1}^{\tau} e^{ct_{k-1}}\Delta t_k. \end{aligned} \tag{E.31}$$

For each interval $[t_{k-1}, t_k]$, $e^{ct_{k-1}} \geq e^{-c\eta_{\max}}e^{ct}$. Integration on that interval and summation consequently yield

$$\sum_{k=1}^{\tau} e^{ct_{k-1}}\Delta t_k \geq e^{-c\eta_{\max}} \int_0^{T_{h+1}(\tau)} e^{ct}\,dt. \tag{E.32}$$

The factor $e^{-c\eta_{\max}}$ accounts for the increasing integrand. Combining the last two displays gives the pathwise recursion

$$cT_h(\tau) \geq cT_{h+1}(\tau) + D\big(e^{cT_{h+1}(\tau)} - 1\big), \qquad h < H. \tag{E.33}$$

Proof of Theorem E.2. The step restriction ensures $b \leq 1/4$, $c_b \geq c_*$, and $c \leq \delta_H/2$. Also $\eta_{\max} \leq 1$ and $\delta_H \leq 1/16$, so $c\eta_{\max} \leq 1/16$ and $D \geq e^{-5/16} = D_*$. On $\mathcal{G}$, every finite hit therefore satisfies $cT_H(\tau) \geq \delta_H c_* \ell$. Apply Equation (E.33) successively, using that f in Equation (E.14) is increasing. The result is

$$cT_1(\tau) \geq f^{\circ(H-1)}(\delta_H c_* \ell) = \delta_H U_H, \qquad T_1(\tau) \geq U_H. \tag{E.34}$$

Here the second inequality uses $c \leq \delta_H$. Since root occupancy is one, a hit at $\tau \leq K$ would instead imply $T_1(\tau) = \sum_{k=1}^{\tau}\eta_k \leq \eta_{\max}K \leq U_H/2$. Thus $\tau > K$ on $\mathcal{G}$ for every budget in the theorem. All policies used in these K episodes have regret at least $1/4$ by Equation (E.4). Multiplying their total $K/4$ by Equation (E.29) proves $\mathbb{E}[\mathrm{Reg}_K] \geq K/8$.

To verify the tower height, let $\rho = 1 + D_* > 1$ and define

$$m_H^{\mathrm{samp}} := \left\lceil \frac{\log(2/(\delta_H c_* \ell))}{\log \rho} \right\rceil, \qquad r_H^{\mathrm{samp}} := H - 1 - m_H^{\mathrm{samp}}. \tag{E.35}$$

Since $e^x - 1 \geq x$ for $x \geq 0$, $f(x) \geq \rho x$. After m_H^{samp} applications, the value is at least two. For $x \geq 2$, $f(x) \geq e^{x/2}$: for example, $D_* > 1/2$ and $e^x - 1 \geq 2e^{x/2}$ in that range already imply the bound. Define $F_0 = 2$ and $F_{r+1} = e^{F_r/2}$. Whenever $r_H^{\mathrm{samp}} \geq 0$, we obtain

$$U_H \geq \frac{F_{r_H^{\mathrm{samp}}}}{\delta_H}. \tag{E.36}$$

For an explicit comparison to ordinary exponentiation, $F_3 > 6$. Let $E_0 = 2$ and $E_{j+1} = e^{E_j}$. Inductively, $F_{j+3} \geq 2E_j + 2$: the base case is $F_3 > 6$, and the induction step uses $e^{E_j+1} \geq 2e^{E_j} + 2$ for $E_j \geq 2$. Consequently $F_r \geq E_{r-3}$ for $r \geq 3$. Since $r_H^{\mathrm{samp}} = H - O(\log H)$, the same tower-height statement holds with ordinary exponential maps. □

Corollary E.3 (A Tower Window for a Budget-Tuned Constant Step)**.** *Fix $0 < \beta \le 1$. For all sufficiently large H, set*

$$Z_H := U_H/\beta, \qquad K_H := \left\lfloor \frac{Z_H \log Z_H}{4} \right\rfloor, \qquad \eta_k \equiv \frac{\beta}{\log K_H}. \tag{E.37}$$

Then the step is at most η_H^{cap}, $K_H \eta_k \le U_H/2$, and $\mathbb{E}[\mathrm{Reg}_{K_H}] \ge K_H/8$.

Proof of corollary E.3. Fix $0 < \beta \le 1$ and put $Z_H = U_H/\beta$. For sufficiently large H, Equation (E.36) ensures $Z_H \ge 16$ and $\log Z_H \ge 2\beta/\eta_H^{\mathrm{cap}}$. Since $Z_H \log Z_H/4 \ge 2$, rounding down loses at most half:

$$K_H \ge \frac{Z_H \log Z_H}{8}, \qquad \log K_H \ge \log Z_H + \log\log Z_H - \log 8 \ge \tfrac{1}{2}\log Z_H. \tag{E.38}$$

Hence $\eta = \beta/\log K_H \le \eta_H^{\mathrm{cap}}$, and

$$K_H \eta \le \frac{Z_H \log Z_H}{4} \frac{2\beta}{\log Z_H} = \frac{U_H}{2}. \tag{E.39}$$

Apply Theorem E.2 with $\eta_{\max} = \eta$. This proves the tuned-step lower bound for the stated budgets. □

E.4 A Terminal Two-Action Lower Bound

Theorem E.4 (Terminal-Bandit Lower Bound)**.** *Consider a one-stage MDP with two actions and deterministic rewards 1 and 0. Run REINFORCE from zero initialization using (N.4). For every $\eta > 0$ and every $K \ge 1$, its regret obeys the pathwise bound*

$$\mathrm{Reg}_K \ge \frac{1}{2\eta e^{2\eta}} \log\bigl(1 + \eta e^{2\eta} K\bigr)\,. \tag{E.40}$$

For $0 < \eta \le 1$, this is $\Omega(\log(1+\eta K)/\eta)$. The same lower scale embeds in the terminal stage of any fixed horizon.

Proof of Theorem E.4. Consider a terminal state that is visited in every episode and has two actions, $a^\star$ and a^0, with deterministic rewards 1 and 0, respectively. We suppress the fixed stage and state indices in this subsection and write

$$p_k := \pi_k(a^\star), \qquad q_k := \pi_k(a^0) = 1 - p_k, \qquad \omega_k := \exp\bigl(\theta_k(a^\star) - \theta_k(a^0)\bigr) = \frac{p_k}{q_k}. \tag{E.41}$$

The zero initialization gives $\omega_1 = 1$, and the regret in episode k is exactly

$$J^\star - J(\pi_k) = q_k = \frac{1}{1+\omega_k}. \tag{E.42}$$

If a^0 is sampled, its reward is zero and the REINFORCE update is zero. If $a^\star$ is sampled, the two logit updates give

$$\log \frac{\omega_{k+1}}{\omega_k} = 2\eta q_k = \frac{2\eta}{1+\omega_k}. \tag{E.43}$$

For every $x \ge 0$, $e^x - 1 \le x e^x$. Hence, on an optimal-action update,

$$\begin{aligned} \omega_{k+1} - \omega_k &= \omega_k \left[\exp\left(\frac{2\eta}{1+\omega_k}\right) - 1\right] \\ &\le \omega_k \frac{2\eta}{1+\omega_k} \exp\left(\frac{2\eta}{1+\omega_k}\right) \le 2\eta e^{2\eta}. \end{aligned} \tag{E.44}$$

The same upper bound is immediate when the update is zero. Therefore, on every sample path,

$$\omega_k \le 1 + 2\eta e^{2\eta}(k-1), \qquad k \ge 1. \tag{E.45}$$

Combining Equations (E.42) and (E.45) and using the integral lower bound for a decreasing sequence yields

$$\begin{aligned}\mathrm{Reg}_K &= \sum_{k=1}^{K} \frac{1}{1+\omega_k} \\ &\geq \sum_{j=0}^{K-1} \frac{1}{2+2\eta e^{2\eta} j} \\ &\geq \int_0^K \frac{dt}{2+2\eta e^{2\eta} t} = \frac{1}{2\eta e^{2\eta}} \log\bigl(1+\eta e^{2\eta} K\bigr)\,. \end{aligned} \tag{E.46}$$

This proves the claimed pathwise bound.

For an arbitrary horizon H, prepend $H-1$ zero-reward stages with one state per stage and deterministic action-independent transitions into this terminal state. All actions in the prepended stages are optimal, the terminal state is still visited in every episode, and its logits have exactly the dynamics above. Thus Equation (E.46) applies at every finite horizon. □

The exact expression also fixes the relevant regimes. If $0 < \eta \leq 1$, then

$$\mathrm{Reg}_K \geq \frac{1}{2e^2\eta} \log(1+\eta K) = \Omega\left(\frac{\log(1+\eta K)}{\eta}\right). \tag{E.47}$$

It is of order $\log K/\eta$ whenever $\log(1+\eta K) = \Theta(\log K)$, for example when $\eta \geq K^{-1/2}$. In particular, for any fixed $c > 0$, the choice $\eta = c/\log K$ gives $\Omega(\log^2 K/c)$ for all sufficiently large K. Indeed, when $\eta \ll 1/K$, the right-hand side of Equation (E.46) is of order K, whereas $\log K/\eta$ can exceed the trivial upper bound K.

E.5 A Fixed-Step Lower Bound at Horizon Two

Theorem E.5 (Fixed-step lower bound)**.** *Fix any $\eta > 0$ and choose an integer action count $A > \max\{2, 1+2/\eta\}$. There exists a finite MDP with $H = 2$, pairwise disjoint state sets across stages, and the same A-element action set at every state. Its rewards and transitions are deterministic. From zero initialization, REINFORCE satisfies, for every $K \geq 1$,*

$$\mathbb{E}[\mathrm{Reg}_K] \geq \frac{K}{2} \exp\left[-\frac{1}{A-1}\left(1+\frac{2}{\eta}\log(1+\eta K/2)\right)\right]. \tag{E.48}$$

Consequently,

$$\mathbb{E}[\mathrm{Reg}_K] = \Omega_{\eta,A}\left(K^{1-2/[\eta(A-1)]}\right), \tag{E.49}$$

where the exponent is strictly positive.

Proof of Theorem E.5. Fix any $\eta > 0$ and choose an integer

$$A > \max\{2, 1+2/\eta\}. \tag{E.50}$$

We construct one deterministic two-stage MDP with the same A-element action set $\mathcal{A}$ at every state. At the root state s_1, distinguish two actions a^{hid} and a^{dec}. Action a^{hid} gives immediate reward zero and moves to a terminal state s_x. Action a^{dec} gives immediate reward $c := 1/2$ and moves to a zero-reward terminal state s_0. Every other root action has zero reward and also moves to s_0. At s_x, one action $a^\star \in \mathcal{A}$ has reward one and the other $A-1$ actions have reward zero. Every action at s_0 has reward zero. All logits are initialized to zero.

The optimal policy selects a^{hid} and then $a^\star$, so its root value is one. Let E_K be the event that the pair $(a^{\mathrm{hid}}, a^\star)$ is never executed in the first K episodes. On E_K, every observed visit to s_x has zero return. The policy at s_x therefore remains uniform, and the current value of the hidden root branch remains $1/A < c$. The decoy branch has value c, while all other root branches have value zero. Consequently, in every episode on E_K the current root value is at most c, and hence

$$\mathrm{Reg}_K \geq \frac{K}{2} \qquad \text{on } E_K. \tag{E.51}$$

It remains to give a lower bound on the probability of E_K. Introduce an auxiliary law $\mathbb{Q}$ under which the reward after a^{hid} is identically zero, including the reward of $a^\star$ at s_x. Under $\mathbb{Q}$, only a selection of a^{dec} produces a nonzero root update. All non-decoy root logits remain equal. After exactly j selections of the decoy action, define the decoy-to-hidden policy odds

$$W_j := \frac{\pi(a^{\mathrm{dec}} \mid s_1)}{\pi(a^{\mathrm{hid}} \mid s_1)}, \qquad W_0 = 1. \tag{E.52}$$

At that time,

$$\pi(a^{\mathrm{dec}} \mid s_1) = \frac{W_j}{W_j + A - 1}, \qquad \pi(a^{\mathrm{hid}} \mid s_1) = \frac{1}{W_j + A - 1}. \tag{E.53}$$

A decoy selection has return c and changes the decoy-minus-hidden logit gap by

$$\eta c \left[1 - \pi(a^{\mathrm{dec}} \mid s_1) + \pi(a^{\mathrm{hid}} \mid s_1)\right] = \frac{\eta c A}{W_j + A - 1}. \tag{E.54}$$

Thus

$$W_{j+1} = W_j \exp\left(\frac{\eta c A}{W_j + A - 1}\right). \tag{E.55}$$

Since $W_j \geq 1$ and $e^x - 1 \geq x$ for $x \geq 0$,

$$W_{j+1} - W_j \geq \frac{\eta c A W_j}{W_j + A - 1} \geq \eta c, \qquad \text{and hence} \qquad W_j \geq 1 + j\eta c. \tag{E.56}$$

Let $M_{\mathrm{hid}}(K)$ be the number of hidden-branch selections in the first K episodes, and let D_{k-1} be the number of decoy selections before episode k. Conditional on the pre-episode history under $\mathbb{Q}$,

$$\mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}\{A_{k,1} = a^{\mathrm{hid}}\} \mid \mathcal{F}_{k-1}\right] = \mathbb{E}_{\mathbb{Q}}\left[\frac{\mathbf{1}\{A_{k,1} = a^{\mathrm{dec}}\}}{W_{D_{k-1}}} \,\middle|\, \mathcal{F}_{k-1}\right]. \tag{E.57}$$

Each decoy selection uses the next index $j = 0, 1, \ldots$. Summing Equation (E.57), then using Equation (E.56), gives

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[M_{\mathrm{hid}}(K)] &= \mathbb{E}_{\mathbb{Q}}\left[\sum_{j=0}^{M_{\mathrm{dec}}(K)-1} \frac{1}{W_j}\right] \\ &\leq \sum_{j=0}^{K-1} \frac{1}{1 + j\eta c} \leq 1 + \frac{1}{\eta c}\log(1 + \eta c K).\end{aligned} \tag{E.58}$$

Under $\mathbb{Q}$, the policy at s_x is always uniform. Conditional on the root trajectory, every hidden-branch visit therefore avoids $a^\star$ with probability $1 - 1/A$. On every path in E_K, the true process and the auxiliary process have identical rewards, logits, and action probabilities. A pathwise coupling, or equivalently a change-of-measure calculation restricted to E_K, consequently gives

$$\mathbb{P}(E_K) = \mathbb{E}_{\mathbb{Q}}\left[\left(1 - \frac{1}{A}\right)^{M_{\mathrm{hid}}(K)}\right]. \tag{E.59}$$

The function $x \mapsto (1 - 1/A)^x$ is convex and decreasing. Jensen's inequality, Equation (E.58), and $\log(1 - 1/A) \geq -1/(A - 1)$ imply

$$\mathbb{P}(E_K) \geq \exp\left[-\frac{1}{A-1}\left(1 + \frac{1}{\eta c}\log(1 + \eta c K)\right)\right]. \tag{E.60}$$

Combining Equations (E.51) and (E.60) and substituting $c = 1/2$ proves the explicit expectation bound

$$\boxed{\mathbb{E}[\mathrm{Reg}_K] \geq \frac{K}{2}\exp\left[-\frac{1}{A-1}\left(1 + \frac{2}{\eta}\log(1 + \eta K/2)\right)\right].} \tag{E.61}$$

For $K \geq 1$, $1 + \eta K/2 \leq (1 + \eta/2)K$. Therefore

$$\mathbb{E}[\mathrm{Reg}_K] \geq \frac{e^{-1/(A-1)}}{2(1+\eta/2)^{2/[\eta(A-1)]}} K^{\,1-2/[\eta(A-1)]}. \tag{E.62}$$

By Equation (E.50), the exponent $\gamma_\eta := 1 - 2/[\eta(A-1)]$ is strictly positive. The quantifiers are thus

$$\forall \eta > 0\ \exists A < \infty\ \exists \mathcal{M}_\eta, C_\eta, \gamma_\eta > 0\ \forall K \geq 1: \qquad \mathbb{E}_{\mathcal{M}_\eta}[\mathrm{Reg}_K] \geq C_\eta K^{\gamma_\eta}. \tag{E.63}$$

The MDP may depend on the proposed fixed step η, but after it is chosen it is fixed for every budget K. □

There is a useful, but deliberately limited, consequence for triangular schedules. Suppose the same instance-independent step η_K is held constant within each K-episode run and is required to cover every finite action count. If $\eta_K \not\to 0$, then some $\eta_0 > 0$ and an infinite subsequence satisfy $\eta_K \geq \eta_0$. Choose one fixed $A > \max\{2, 1 + 2/\eta_0\}$. For fixed K, the function

$$\eta \longmapsto \frac{1}{\eta} \log(1 + \eta K/2) \tag{E.64}$$

is decreasing, because its derivative has the sign of $x/(1+x) - \log(1+x) \leq 0$, where $x = \eta K/2$. Hence Equation (E.61) with $\eta_K \geq \eta_0$ is bounded below by the polynomial expression obtained with η_0, on the same fixed MDP and along the entire subsequence. Thus such an instance-independent schedule must satisfy $\eta_K \to 0$ if it is to guarantee logarithmic regret over all finite action counts.

E.6 Why Downstream Readiness and Optimal-Set Aggregation Are Needed

The following two deterministic $H = 2$ examples rule out tempting intermediate monotonicity claims.

E.6.1 Bellman Reversal Before Downstream Readiness

At the root, let actions a^{d} and a^{s} lead to distinct terminal states, with zero immediate reward. In the deep terminal state, one action pays one, all other actions pay zero, and the current policy assigns probability q to the rewarding action. In the safe terminal state, every action pays v, where

$$0 < q < v < 1. \tag{E.65}$$

Then

$$Q_1^\star(s_1, a^{\mathrm{d}}) = 1 > v = Q_1^\star(s_1, a^{\mathrm{s}}), \qquad Q_1^\pi(s_1, a^{\mathrm{d}}) = q < v = Q_1^\pi(s_1, a^{\mathrm{s}}). \tag{E.66}$$

Thus the optimal root action can have the smaller current-policy action value. If the two root logits are equal, the mean REINFORCE drift of $\theta(a^{\mathrm{d}}) - \theta(a^{\mathrm{s}})$ is exactly $(q - v)/2 < 0$. The example shows that an upstream sign argument must first control the downstream value error. This is the reason for the backward readiness order used in the upper-bound proof.

E.6.2 A Fixed Optimal Logit Pair Can Drift Backward

Consider three root actions o_1, o_2, b. Both o_1 and o_2 are optimal with optimal value one, while b has optimal value zero. For some $0 < \varepsilon < 1$, choose realizable downstream policies such that

$$Q_1^\pi(s_1, \cdot) = (1 - \varepsilon, 1, 0), \qquad Q_1^\star(s_1, \cdot) = (1, 1, 0). \tag{E.67}$$

For example, o_1 may lead to a terminal block that is not yet ready, o_2 to a terminal block in which every action pays one, and b to a zero-reward block. Let the root policy satisfy

$$\pi(o_1 \mid s_1) = \pi(o_2 \mid s_1) = \frac{1 - \beta}{2}, \qquad \pi(b \mid s_1) = \beta. \tag{E.68}$$

Writing $V^\pi = (1 - \beta)(1 - \varepsilon/2)$ and using the tabular softmax policy-gradient identity from Equation (G.6), direct substitution gives

$$g_1(\theta; s_1, o_1) - g_1(\theta; s_1, b) = \frac{1 - \beta}{2} \left[3\beta \left(1 - \frac{\varepsilon}{2} \right) - \frac{\varepsilon}{2} \right]. \tag{E.69}$$

This drift is negative whenever

$$0 < \beta < \frac{\varepsilon}{6(1-\varepsilon/2)}, \tag{E.70}$$

even though $Q_1^\pi(s_1, o_1) - Q_1^\pi(s_1, b) = 1 - \varepsilon > 0$. Hence the logit difference between a fixed optimal action and a fixed suboptimal action need not increase. The upper-bound proof instead uses inverse-coordinate comparisons with a minimum optimal inverse coordinate and charges the sum of all optimal logits. Minimizer switching and multiple optimal actions are then handled at the level of the optimal action set.

F Constant Ledger, Boundary Cases, and Scope

This appendix completes the constant definitions used in Theorem A.1 and corollary A.2. We list their dependency order, expand the probability-floor and step-size reciprocals as functions of the Bellman gap, and handle the cases excluded by $\mathcal{I}_+ \neq \varnothing$. The final subsection records the scope of the resulting upper bound.

F.1 Acyclic Constant Ledger

All quantities are determined in the following order from the fixed instance parameters

$$H, \qquad A, \qquad (|\mathcal{S}_h|)_{h=1}^H, \qquad \Delta_{\min}, \qquad p_{\min}.$$

Quantity	Definition	Role
B_H, Λ, n_θ	Equations (C.1), (G.8) and (G.16)	gradient, Lipschitz, and union bounds
$L_h^{\mathrm{sep}}, g_{\mathrm{rec}}$	Equations (D.10) and (D.15)	margins ensuring Bellman-tail control
F_j, D_h	Equations (D.24) and (D.26)	multi-hop occupancy delivery
U_h, u_{ready}	Equations (D.27) and (D.30)	readiness
$\varepsilon_0, U^\dagger, a_0, \eta_{\mathrm{Euler}}$	Equations (C.2) and (C.5)	discrete tracking tube
κ_{pre}	Equation (C.6)	pre-readiness failure exponent
$\kappa_0, \kappa_{\mathrm{post}}$	Equations (B.3) and (B.6)	post-readiness barrier exponent
$\mu_0, \bar{\eta}$	Equations (B.7) and (B.11)	probability floor and certified-step numerator
B_K^{logit}	Equation (A.13)	finite-K logit localization

The delivery map F_j is repeatedly composed along a path and across backward phases. Consequently, the displayed proof can make u_{ready} tower-sized, μ_0 extremely small, and $\bar{\eta}$ extremely conservative.

F.2 Gap-Explicit Evaluation of the Final Constants

To display the Bellman-gap dependence explicitly, fix the gap-free parameter tuple

$$\xi := \left(H, A, (|\mathcal{S}_h|)_{h=1}^H, p_{\min}\right), \qquad n_\theta = A\sum_{h=1}^H |\mathcal{S}_h|, \tag{F.1}$$

and evaluate the preceding ledger at a variable gap $\delta \in (0, H]$. Write

$$U_\xi^\dagger(\delta) := u_{\mathrm{ready},\xi}(\delta) + 1, \tag{F.2}$$

$$L_{\max,\xi}^{\mathrm{sep}}(\delta) := \begin{cases} 0, & H = 1, \\ 4A(H-1)^2/\delta, & H \geq 2, \end{cases} \tag{F.3}$$

$$\varepsilon_\xi(\delta) := \min\left\{\frac{1}{8}, \frac{1}{16(L_{\max,\xi}^{\mathrm{sep}}(\delta) + 5)}\right\}. \tag{F.4}$$

Here $u_{\mathrm{ready},\xi}(\delta)$ is exactly the finite recursion in Equations (D.24), (D.26), (D.27) and (D.30), with $\Delta = \delta$ and $p = p_{\min}$. Thus these definitions retain, rather than suppress, every occurrence of the gap inside the repeated delivery maps.

Define the reciprocal probability floor

$$m_\xi(\delta) := 2A\exp\Big(2B_H U_\xi^\dagger(\delta) + 2\varepsilon_\xi(\delta)\Big) = \mu_\xi(\delta)^{-1} \tag{F.5}$$

and the reciprocal certified-step numerator

$$M_\xi(\delta) := \max\Bigg\{ \frac{\Lambda B_H U_\xi^\dagger(\delta) e^{\Lambda U_\xi^\dagger(\delta)}}{\varepsilon_\xi(\delta)\log 2}, \frac{64H^2}{\delta \log 2}, C_\theta \left[\frac{8n_\theta B_H^2 U_\xi^\dagger(\delta) e^{2\Lambda U_\xi^\dagger(\delta)}}{\varepsilon_\xi(\delta)^2} + \frac{8B_H\sqrt{n_\theta} e^{\Lambda U_\xi^\dagger(\delta)}}{3\varepsilon_\xi(\delta)} \right], \frac{64H^2 C_\theta}{\delta \log\left(\frac{L^{\mathrm{sep}}_{\max,\xi}(\delta)+3}{L^{\mathrm{sep}}_{\max,\xi}(\delta)+2} \right)} \Bigg\} = \bar{\eta}_\xi(\delta)^{-1}. \tag{F.6}$$

These equalities are direct expansions. Indeed, $\kappa_0 < 1$ and Equations (B.8), (B.9) and (B.11) give

$$\bar{\eta} = \min\left\{ (\log 2)\eta_{\mathrm{Euler}}, (\log 2)\kappa_0, \frac{\kappa_{\mathrm{pre}}}{C_\theta}, \frac{\kappa_{\mathrm{post}}}{C_\theta} \right\}.$$

Substituting Equations (B.3), (B.6) and (C.4) to (C.6) and taking reciprocals gives Equation (F.6). At the actual gap,

$$\mu_0 = \mu_\xi(\Delta_{\min}), \qquad \bar{\eta} = \bar{\eta}_\xi(\Delta_{\min}). \tag{F.7}$$

F.2.1 Size of the Certified Constants

To expose how instance parameters govern the bound, consider a variable Bellman gap $\delta \in (0, H]$ and define the reciprocal quantities

$$m_\xi(\delta) := \mu_\xi(\delta)^{-1}, \qquad M_\xi(\delta) := \bar{\eta}_\xi(\delta)^{-1}.$$

These reciprocals govern the instance multiplier in the regret bound and the step-size restriction, respectively. We also define the state–stage weight

$$W_\xi(\delta) := \sum_{h=1}^{H} |\mathcal{S}_h| (1 + m_\xi(\delta))^{h-1}, \tag{F.8}$$

together with the aggregate rate factor and leading coefficient

$$\mathfrak{F}_\xi(\delta) := m_\xi(\delta) M_\xi(\delta) W_\xi(\delta), \qquad C_\xi := A\left(3 + \frac{\log(2n_\theta)}{\log 2}\right). \tag{F.9}$$

The constants are constructed via a deterministic mean-flow readiness time $u_{\mathrm{ready},\xi}(\delta)$ obtained from a backward recursion across stages in Appendix D. Readiness requires that optimal actions attain strictly higher current-policy action values than suboptimal actions at every reachable state, with sufficient margins to support the later probability-floor argument. We set $U := U_\xi^\dagger(\delta) = u_{\mathrm{ready},\xi}(\delta) + 1 \geq 1$. With structural constants $B_H := \sqrt{H(H+1)(2H+1)/3} = O(H^{3/2})$ and $\Lambda := H/2 + \sqrt{2H} B_H = O(H^2)$, the reciprocal floor and step scale satisfy

$$2A\exp(2B_H U) \leq m_\xi(\delta) \leq 2A\exp(2B_H U + 1/4),$$
$$\frac{64H^2}{\delta \log 2} \leq M_\xi(\delta) \lesssim_\xi (1+\delta^{-1})^2 U \exp(2\Lambda U), \tag{F.10}$$

where $\lesssim_\xi$ denotes an inequality up to a positive finite factor depending only on ξ, independently of δ and K. Because $m_\xi(\delta) \geq 1$, the state–stage weight satisfies

$$W_\xi(\delta) \leq \left(\sum_{h=1}^{H} |\mathcal{S}_h| \right) \big(2m_\xi(\delta)\big)^{H-1}, \tag{F.11}$$

which combines with (F.10) to bound the aggregate rate factor as

$$\mathfrak{F}_\xi(\delta) \lesssim_\xi (1+\delta^{-1})^2 U \exp\big((2HB_H + 2\Lambda)U\big). \tag{F.12}$$

These relations provide bounds on the constants supplied by this proof. The large instance dependence enters through U, where the backward recursion iteratively composes exponential delivery maps across paths and stages (Equations (D.24), (D.26), (D.27) and (D.30)). Even in the single-stage bandit setting ($H = 1$),

$$U = 1 + \frac{(1 + 2/\delta)^A - 1}{p_{\min}}$$

by (F.13). Substituting this expression for U places δ and $p_{\min}$ inside the exponents of $m_\xi(\delta)$ and $M_\xi(\delta)$ in (F.10). Exact expansions of $m_\xi(\delta)$ and $M_\xi(\delta)$ are given in (F.5) and (F.6).

The dependence in these functions is generally not a product of a gap-only factor and a factor containing all other parameters. Already when $H = 1$,

$$u_{\mathrm{ready},\xi}(\delta) = \frac{(1 + 2/\delta)^A - 1}{p_{\min}}, \tag{F.13}$$

so $p_{\min}^{-1}$ and δ^{-1} occur together inside the exponent in Equation (F.5) and inside still larger exponentials in Equation (F.6). More explicitly, for two positive reachability scales $p_1 \neq p_2$, the ratio of the corresponding floor reciprocals contains

$$\exp\big(2B_H\big[(1 + 2/\delta)^A - 1\big](p_1^{-1} - p_2^{-1})\big)\,,$$

which varies with δ. Thus even this factor cannot equal $C(A, p_{\min})F(\delta)$ exactly. This coupling persists throughout the general backward recursion. The present proof therefore supplies an explicit recursive gap function coupling δ with $H, A, p_{\min}$.

F.3 Occupancy Time and Readiness Time

The two delivery formulas play different roles. For a fixed window and target, Theorem D.5 gives the closed-form sufficient time $\mathsf{T}_h^{\mathrm{occ}}(U, D)$ with $h - 1$ exponential compositions. In the readiness ledger, $\mathsf{D}_h(U, D)$ instead denotes the recursion in Equation (D.26). Both are proved sufficient; the constants μ_0 and $\bar{\eta}$ in this paper use the latter throughout.

The one-edge targets satisfy the verified comparison

$$F_j^{\mathrm{int}}(U, D) \leq F_j(U, D).$$

The closed-form formula also replaces each stage range $\overline{H}_j$ by H and enlarges the path composition. Furthermore, the readiness recursion changes both the start time and the occupancy target at each backward phase.

The occupancy claims concern the deterministic integral $\mathfrak{a}_{h,s}(U, t)$. They supply the finite mean-flow interval used by the tracking lemma.

F.4 Boundary Cases

All reachable actions optimal. If $\mathcal{I}_+ = \varnothing$, every action at every structurally reachable state is optimal. By Equation (G.2), $\mathcal{R}(\pi) = 0$ for every policy generated by the algorithm, so $\mathrm{Reg}_K = 0$ pathwise. In this setting, reachable optimal-set masses equal one and unreachable blocks remain uniform. Thus the floor holds for every step size, and the gap-dependent tuning and all quantities that divide by $\Delta_{\min}$ are unnecessary.

One action. If $A = 1$, the only action is optimal in every state, so $\mathcal{I}_+ = \varnothing$ and the preceding zero-regret branch applies. The occupancy transmission theorem remains valid, using the empty-product convention in the inverse-coordinate identities.

Unreachable states and zero transition probabilities. A structurally unreachable block has zero update and remains uniform. Its optimal-set mass is at least $1/A$. The readiness proof uses only positive witness edges.

Multiple optimal actions. The optimal set $\mathcal{A}_h^\star(s)$ may contain multiple actions. The barrier chooses a predictable minimum inverse coordinate within the entire optimal set, and Equation (B.18) shows that changing minimizers is favorable. The fixed-pair counterexample in Section E.6 explains why this aggregation is necessary.

Rare state clocks. Before readiness, applying Lemma D.6 along a reachable path gives a lower bound on cumulative occupancy. After readiness, the same factor $d_h^{\pi_k}(s)/Z_{k,h,s}$ multiplies both the barrier drift and its conditional second moment. When this clock vanishes, the block has zero drift and zero noise.

Short budgets. If $K < k_{\mathrm{ready}}$, the proof uses only the pre-readiness tube and the action lower bound Equation (C.9).

F.5 Exact Theorem Scope

The established result has the following boundaries.

- The REINFORCE update in Equation (N.4) uses a separate logit block for every state and stage. All visited blocks are updated simultaneously.
- Each update uses one fresh on-policy trajectory and the estimator in Equation (N.4).
- Rewards follow assumption G.1.
- The result controls expected regret.
- The learning rate is constant within one run and is selected after an episode budget is announced.
- The gap dependence is the explicit recursive dependence in Equations (F.5) and (F.6).
- Theorem E.4 matches the selected tuning's K–η learning-time scale in its stated regime.
- Theorem E.5 rules out one universal positive fixed step over all finite action counts.

The proof of Theorem A.1 uses the mean flow and discrete estimates directly.

G Technical Foundations and Elementary Identities

This appendix proves the identities and uniform bounds used in Appendices A–D. The notation and algorithm are given in the standing setup at the start of the appendices. We first derive the Bellman and performance-difference identities, then the estimator moments and zero-sum logit property, and finally the Lipschitz bound used for tracking.

Assumption G.1 (Bounded Deterministic Rewards Along the Trajectory)**.** *The learner does not know or estimate the transition kernels. Reward functions $r_h : \mathcal{S}_h \times \mathcal{A} \to [0,1]$ are bounded and deterministic. In episode k, the learner observes one on-policy state–action trajectory and the value $r_h(S_{k,h}, A_{k,h})$ at each visited pair. Actions and transitions are sampled.*

G.1 Bellman and Performance-Difference Identities

Lemma G.2 (Bellman and Performance-Difference Identities)**.** *For every Markov policy π,*

$$e_h^\pi(s) = \ell_h^\pi(s) + \begin{cases} \sum_{a\in\mathcal{A}} \pi_h(a \mid s) \sum_{x\in\mathcal{S}_{h+1}} P_h(x \mid s,a) e_{h+1}^\pi(x), & h < H, \\ 0, & h = H, \end{cases} \tag{G.1}$$

$$\mathcal{R}(\pi) = \sum_{h=1}^{H} \mathcal{R}_h(\pi). \tag{G.2}$$

Proof. For $h < H$, add and subtract the policy average of the optimal action values:

$$\begin{aligned} e_h^\pi(s) &= \sum_a \pi_h(a \mid s)\big[V_h^*(s) - Q_h^*(s,a)\big] + \sum_a \pi_h(a \mid s)\big[Q_h^*(s,a) - Q_h^\pi(s,a)\big] \\ &= \ell_h^\pi(s) + \sum_a \pi_h(a \mid s) \sum_x P_h(x \mid s,a) e_{h+1}^\pi(x). \end{aligned}$$

The second term vanishes at $h = H$, proving Equation (G.1). Optimality also gives $e_h^\pi(s) \geq 0$, either by backward induction in the Bellman equations or from this recursion and $\ell_h^\pi \geq 0$.

The occupancy recursion is

$$d^{\pi}_{h+1}(x) = \sum_{s} d^{\pi}_{h}(s) \sum_{a} \pi_h(a \mid s) P_h(x \mid s, a). \tag{G.3}$$

Multiplying the error recursion by $d^{\pi}_{h}(s)$ and summing in s therefore gives

$$\sum_{s} d^{\pi}_{h}(s) e^{\pi}_{h}(s) = \mathcal{R}_h(\pi) + \sum_{x} d^{\pi}_{h+1}(x) e^{\pi}_{h+1}(x),$$

with the last sum absent at H. Iterating from any stage h proves the stronger suffix identity

$$\sum_{s} d^{\pi}_{h}(s) e^{\pi}_{h}(s) = \sum_{j=h}^{H} \mathcal{R}_j(\pi). \tag{G.4}$$

At $h = 1$, the left side is $\sum_s \rho(s) e^{\pi}_{1}(s) = J^* - J(\pi)$, which proves Equation (G.2). □

G.2 Conditional Mean, Moments, and the Block Gauge

Let

$$g(\theta) := \mathbb{E}[\widehat{g}_k \mid \theta_k = \theta] \tag{G.5}$$

denote the mean update field. $\widehat{g}_k$ is computed from one complete trajectory, so its components can be correlated across stages and state blocks.

Lemma G.3 (Exact Policy-Gradient Mean and Uniform Bounds)**.** *For every coordinate* (h, s, a)*,*

$$g_h(\theta; s, a) = d^{\pi_\theta}_{h}(s) \pi_{\theta,h}(a \mid s) \operatorname{Adv}^{\pi_\theta}_{h}(s, a) = \frac{\partial J(\pi_\theta)}{\partial \theta_h(s, a)}, \tag{G.6}$$

$$\mathbb{E}[\widehat{g}_{k,h}(s, a)^2 \mid \mathcal{F}_{k-1}] \le \overline{H}_h^2 d^{\pi_k}_{h}(s) \pi_{k,h}(a \mid s)(1 - \pi_{k,h}(a \mid s)). \tag{G.7}$$

With

$$B_H^2 := 2 \sum_{h=1}^{H} \overline{H}_h^2 = \frac{H(H+1)(2H+1)}{3}, \tag{G.8}$$

one has almost surely

$$\|\widehat{g}_k\|_2 \le B_H, \qquad \|g(\theta_k)\|_2 \le B_H, \qquad |\widehat{g}_{k,h}(s, a)| \le \overline{H}_h \le H. \tag{G.9}$$

Finally, every update has zero sum inside every state–stage block, so

$$\sum_{a \in \mathcal{A}} \theta_{k,h}(s, a) = 0 \quad \textit{for all } k, h, s. \tag{G.10}$$

Proof. Condition throughout on $\mathcal{F}_{k-1}$ and suppress the episode index on the policy. The Markov property and the return definition give

$$\mathbb{E}[G_{k,h} \mid S_{k,h} = s, A_{k,h} = a, \mathcal{F}_{k-1}] = Q^{\pi}_{h}(s, a).$$

Summing over the sampled action yields

$$\begin{aligned} \mathbb{E}[\widehat{g}_{k,h}(s, a) \mid \mathcal{F}_{k-1}] &= d^{\pi}_{h}(s) \left[\pi_h(a \mid s) Q^{\pi}_{h}(s, a) - \pi_h(a \mid s) \sum_{c} \pi_h(c \mid s) Q^{\pi}_{h}(s, c) \right] \\ &= d^{\pi}_{h}(s) \pi_h(a \mid s) \operatorname{Adv}^{\pi}_{h}(s, a). \end{aligned}$$

To identify this mean with $\partial J / \partial \theta_h(s, a)$, differentiate the finite trajectory sum for the total return. Its likelihood-ratio factor is precisely $\mathbf{1}\{S_h = s\}(\mathbf{1}\{A_h = a\} - \pi_h(a \mid s))$. Conditional on the history up to S_h, the score has mean zero, so every reward obtained before stage h contributes zero. The remaining reward sum is G_h, and the displayed calculation applies. This proves both equalities in Equation (G.6). Since $G_{k,h} \le \overline{H}_h$ and

$$\mathbb{E}[(\mathbf{1}\{A_{k,h} = a\} - \pi_{k,h}(a \mid s))^2 \mid S_{k,h} = s] = \pi_{k,h}(a \mid s)(1 - \pi_{k,h}(a \mid s)),$$

we obtain Equation (G.7). A trajectory visits one state block per stage and the Euclidean norm of $e_{A_{k,h}} - \pi_{k,h}(\cdot \mid S_{k,h})$ is at most $\sqrt{2}$. Summing the squared block norms proves Equation (G.9); Jensen's inequality gives the bound on g. The components of each softmax score sum to zero, which proves Equation (G.10) from the zero initialization. □

Lemma G.4 (Subset-Sum Bound Under the Block Gauge)**.** *Fix one state–stage block and suppose* $\sum_{a\in\mathcal{A}}\theta_a = 0$ *and* $\theta_a \geq -C$ *for every action, where* $C \geq 0$*. Then every subset* $O \subseteq \mathcal{A}$ *satisfies*

$$\left|\sum_{a\in O}\theta_a\right| \leq AC. \tag{G.11}$$

Proof. The coordinatewise lower bound gives $\sum_{a\in O}\theta_a \geq -|O|C \geq -AC$. Applying the zero-sum gauge also gives

$$\sum_{a\in O}\theta_a = -\sum_{a\notin O}\theta_a \leq |\mathcal{A}\setminus O|C \leq AC,$$

which proves the claim. □

G.3 Trajectory Moments

We record the first two moments of the REINFORCE update, retaining the covariance of the complete trajectory. The associated diffusion provides intuition for the proof quantities. Index a complete trajectory by $\tau = (s_1, a_1, \ldots, s_H, a_H)$, let $\iota_{h,s}$ embed an action vector into its state–stage block, and set $G_h(\tau) := \sum_{j=h}^{H} r_j(s_j, a_j)$. Its probability and REINFORCE atom are

$$p_\theta(\tau) := \rho(s_1)\prod_{h=1}^{H}\pi_{\theta,h}(a_h \mid s_h)\prod_{h=1}^{H-1}P_h(s_{h+1} \mid s_h, a_h), \tag{G.12}$$

$$\widehat{g}_\tau(\theta) := \sum_{h=1}^{H} G_h(\tau)\,\iota_{h,s_h}\big(e_{a_h} - \pi_{\theta,h}(\cdot \mid s_h)\big). \tag{G.13}$$

Consequently,

$$m(\theta) := \mathbb{E}_\theta[\widehat{g}_\tau] = \sum_\tau p_\theta(\tau)\widehat{g}_\tau(\theta) = g(\theta), \tag{G.14}$$

$$\begin{aligned} C(\theta) &:= \mathrm{Cov}_\theta(\widehat{g}_\tau) \\ &= \sum_\tau p_\theta(\tau)\big(\widehat{g}_\tau(\theta) - m(\theta)\big)\big(\widehat{g}_\tau(\theta) - m(\theta)\big)^\top. \end{aligned} \tag{G.15}$$

These finite sums retain action and transition randomness, shared returns, same-stage centering, and cross-stage correlations.

Lemma G.5 (Global Lipschitz Mean Field)**.** *The finite-dimensional vector field* g *is globally Lipschitz, and one may take*

$$\mathrm{Lip}(g) \leq \Lambda := \frac{H}{2} + \sqrt{2H}\,B_H. \tag{G.16}$$

Proof. For a trajectory with positive environment probability, let $S_\tau(\theta) := \nabla \log p_\theta(\tau)$. Each visited block contributes the score $e_{a_h} - \pi_{\theta,h}(\cdot \mid s_h)$, and distinct stages use distinct blocks. Therefore $\|S_\tau(\theta)\|_2^2 \leq 2H$. Trajectories with zero environment probability contribute identically zero and can be omitted.

For a softmax block π, its Jacobian is $J_\pi = \mathrm{diag}(\pi) - \pi\pi^\top$. The absolute row sum in row a is $2\pi_a(1-\pi_a) \leq 1/2$. Symmetry and the row-sum norm bound imply $\|J_\pi\|_{\mathrm{op}} \leq 1/2$. When a trajectory is fixed, its returns are fixed numbers, so the Jacobian of $\widehat{g}_\tau(\theta)$ is block diagonal, with visited block $-G_h(\tau)J_{\pi_h}$. Hence

$$\|D\widehat{g}_\tau(\theta)\|_{\mathrm{op}} \leq H/2.$$

Differentiating the finite expectation gives

$$Dg(\theta) = \sum_\tau p_\theta(\tau)\left[D\widehat{g}_\tau(\theta) + \widehat{g}_\tau(\theta)S_\tau(\theta)^\top\right].$$

Because the operator norm of an outer product equals the product of its Euclidean norms, the atom bound $\|\widehat{g}_\tau\|_2 \leq B_H$ yields $\|Dg(\theta)\|_{\mathrm{op}} \leq H/2 + B_H\sqrt{2H} = \Lambda$ uniformly in θ. Integrating this Jacobian along the line segment between two logit tables proves the claimed global Lipschitz inequality. □

For later martingale arguments, write

$$\varepsilon_k := \widehat{g}_k - g(\theta_k), \qquad \mathbb{E}[\varepsilon_k \mid \mathcal{F}_{k-1}] = 0. \tag{G.17}$$

H Related Work

Global convergence and iteration complexity for tabular softmax policy gradient have been characterized under exact-gradient and related oracle models (Mei et al., 2020b; Agarwal et al., 2021; Xiao, 2022). Stochastic policy-gradient convergence and sample complexity have also been studied under varied assumptions and update rules (Yuan et al., 2022; Fatkhullin et al., 2023; Klein et al., 2024; Labbi et al., 2026), alongside policy-optimization regret under unknown transitions (Li et al., 2026).

Existing regret guarantees in MDPs. Zhang et al. (2021) establish an anytime high-probability regret bound of $\widetilde{O}(N^{5/6})$ under a truncated-trajectory criterion for REINFORCE in discounted MDPs. Their algorithm employs phased doubling epochs with decreasing step sizes and trajectory truncation, using log-barrier regularization and policy postprocessing floors to sustain exploration. In unknown-transition MDPs, Shani et al. (2020) establish regret bounds via transition confidence sets and cost optimism. For episodic MDPs, Dann et al. (2023) achieve gap-dependent polylogarithmic regret in stochastic environments and square-root regret in adversarial settings using policy regularization with exploration bonuses.

Regret control from bandits to MDPs. In stochastic multi-armed bandits, Baudry et al. (2025) show that discrete stochastic gradient bandits achieve logarithmic regret for sufficiently small constant step sizes, whereas larger steps can incur polynomial regret. Related analyses study softmax optimization, stochasticity, and baselines (Mei et al., 2020a; 2021; 2022), as well as diffusion models for policy-gradient regret (Lattimore, 2026a; Jia & Ouyang, 2026). For discrete sampled bandits, Lattimore (2026b) shows that the conditional drift of the aggregate optimal-action logit coordinate equals the step size times the total optimal-action probability times regret. Telescoping a nonlinear Lyapunov potential of this logit sum bounds regret on a localized event under gap- and budget-dependent step sizes.

In MDPs, an inaccurate downstream policy can give an optimal upstream action negative advantage under the current policy, a challenge absent in bandits. For single-trajectory finite-horizon REINFORCE, Robertson et al. (2025, proof of Theorem 5.2 and Appendix E.3) analyze sampled-logit and visit-count dynamics. They establish infinite visitation of every reachable state–action pair almost surely and optimal-value convergence for any positive constant step size. After a random entry time, this yields an eventual pathwise average suboptimality rate of $O(\log T/T)$. Their backward induction waits for downstream errors to shrink, whereas bounding expected regret from initialization requires controlling expected cumulative transient loss and the expectations of random rate coefficients.

To address this challenge within REINFORCE, we introduce the Bellman debt, defined as the downstream value error incurred by optimal upstream actions, weighted by state occupancy and action probability. While downstream policies remain inaccurate, the occupancy recursion under the current policy directly bounds this debt by downstream stage regrets. In adversarial MDPs, Luo et al. (2021) compensate for estimation bias and variance under comparator occupancy by injecting a dilated Bellman bonus into the policy update. This bonus uses geometric cancellation to transfer error evaluation to learner occupancy. For REINFORCE, our geometric weighting is purely an analysis device, producing a linear potential within the Lyapunov analysis. This debt bound holds for every current policy, but using the linear potential to control regret requires a uniform floor on total optimal-action probability.

Quantitative implicit exploration. In exact-gradient discounted MDPs, Mei et al. (2020b) establish positive lower bounds on optimal-action probabilities along gradient trajectories under a full-support initial state distribution. Mei et al. (2023) provide quantitative bounds on softmax probability decay in stochastic bandits, showing that action probabilities cannot decay faster than an inverse-time rate, which implies infinite sampled visitation almost surely. For our deterministic mean flow from zero initialization, we consider an arbitrary finite window. Over this window, we establish a logarithmic lower bound on the cumulative use of an action in terms of the cumulative occupancy of its parent state. By controlling complete softmax probabilities through inverse coordinates and a zero-sum

invariant, this relation constrains probability decay and propagates occupancy along paths under arbitrary current-policy Q-value orderings. Finite-window tracking couples REINFORCE iterates to this mean flow, and together with a subsequent discrete barrier, establishes a high-probability floor on total optimal-action probability across the episode budget for REINFORCE.

Recursive policy suppression and slow convergence. Li et al. (2023) demonstrate slow last-iterate convergence for exact-gradient softmax policy gradient in discounted MDPs. Under stated step-size restrictions, a uniform initial-state distribution, and uniform policy initialization, this slow convergence is driven by recursive optimal-action suppression across states. We analyze this recursive suppression mechanism for root-started finite-horizon chains with two actions per state and deterministic transitions and rewards. In Sections 5 and E, we establish a mean-flow tower lower bound for deterministic continuous-time dynamics. For the stochastic setting, where step sizes may depend on history subject to a small uniform cap, REINFORCE initialized at zero incurs a long window of linear expected regret.

I Limitations and Open Directions

Additional reward noise. The model uses a deterministic bounded reward function $r_h(s, a)$ observed on visited state–action pairs. Conditionally zero-mean reward observation noise would add martingale increments to the tracking, barrier, and localization arguments; extending the analysis to random rewards remains open.

Budget-dependent tuning. For an announced budget $K \geq 2$, instance-dependent tuning in Corollary 3.2 sets $\eta_K = \bar{\eta}/\log K$ to achieve $O_{\mathcal{M}}(\log^2 K)$ expected regret. With budget-only tuning, Remark A.3 sets $\eta_K = (1 + \log K)^{-(1+\varepsilon)}$ for any fixed $\varepsilon > 0$ and achieves $\mathbb{E}[\mathrm{Reg}_K] = O_{\mathcal{M},\varepsilon}((1 + \log K)^{2+\varepsilon})$. The convenient choice $\varepsilon = 1$ gives the cubic rate in Corollary 3.3. For every deterministic episode budget $T \geq 1$, Remark A.4 gives anytime expected regret $O_{\mathcal{M}}((1 + \log T)^2)$ with $\bar{\eta}$ and $O_{\mathcal{M}}((1 + \log T)^3)$ with budget-only tuning; the constants depend on the instance. Obtaining these rates with time-varying steps in a single run that retains its logits remains open. Theorem E.5 rules out a single positive constant step that guarantees logarithmic expected regret across the finite class. Theorem E.1 establishes tower delay for the mean flow on the constructed chain. Theorem E.2 establishes a tower horizon obstruction for REINFORCE under its stated step-size cap and budget conditions. The necessary decay rate across unrestricted schedules and the performance of instance-dependent constant steps on a fixed instance remain open.

Parameter dependence. Corollary A.2 exposes the complete gap dependence of the current certificate through the recursive function $\mathfrak{F}_\xi(\Delta_{\min})$. The readiness recursion mixes $\Delta_{\min}$ with the horizon, action count, and positive reachability scale inside repeated exponentials, while the tracking and barrier constants add further dependence. The recursion can be an iterated-exponential tower. Our lower bounds (Theorem E.1 for the mean flow and Theorem E.2 for REINFORCE under the stated step-size cap and budget conditions) establish genuine tower delay on the constructed chain. The exact geometric Bellman weights and the full recursive constant $\mathfrak{F}_\xi(\Delta_{\min})$ reflect our specific linear certificate; their necessity remains open.

Model and policy class. Our analysis is tabular, uses a fixed finite horizon, and assigns a separate logit block to every state and stage. Shared stationary parameters, linear or nonlinear function approximation, continuous spaces, and growing horizons are outside scope. Extending the readiness argument and weighted linear potential to feature-coupled parameters remains open.

Performance criterion. We bound the expectation of regret under zero initialization. High-probability regret bounds, realized reward regret bounds, and the necessity of zero initialization remain open.