FROZEN POLICY MIXTURES: HOW DEPLOYMENT LIMITS FEEDBACK IN IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how deployment affects imitation learning from clean expert labels in finite policy classes containing the expert. For two stochastic policies with a uniform prior and Bayesian updates, frozen per-step mixing has worst-case cumulative squared trajectory Hellinger loss of iterated-logarithm order in episode length. Matching bounds identify a feedback mechanism: failure ends informative visits, limiting the next update. Sampling once per episode or updating after each label removes this horizon dependence. This law is specific to trajectory discrepancy; no cumulative return-gap bound uniform in rounds and stochastic classes holds even for one-step episodes. On our two-state construction, however, terminal-success rewards yield the same separation in cumulative return gaps: iterated-logarithmic for frozen mixing and constant for both alternatives. We also characterize batch updates and likelihood tempering. For deterministic classes, matching return-gap bounds resolve the proposed class-size improvement for WARM-STAGGER; plurality voting achieves the proposed rates in expectation under its single-state query model. Controlled experiments validate the trajectory and return calculations, update choices, larger stochastic classes, and deterministic voting.

1 INTRODUCTION

An imitation learner can draw one candidate policy for the whole episode or draw again before each action. Resampling changes the states it reaches and hence the expert labels available for subsequent updates. How much can this deployment choice change trajectory discrepancy and return shortfall for the same candidate policies?

Learning from the states visited by the learner is the premise of DAgger (Ross et al., 2011). Per-episode and per-step mixtures are distinguished explicitly in Li & Zhang (2022; 2025), and Li & Zhang (2025) deploy exponential-weighted per-step mixtures in STAGGER and WARM-STAGGER. Pointwise likelihood-weighted mixtures also arise in distillation from noisy experts (Sriraman et al., 2026). We study mixtures whose weights stay fixed throughout an episode and are revised after its labels arrive. A wrong policy remains available at every step, including after labels have supplied evidence against it.

This delay can make cumulative trajectory Hellinger loss grow with episode length even for two policies. In Figure 1, the expert takes safe actions in an active state, while the other policy can cause failure. Failure leads to an absorbing state where both policies agree, ending informative feedback. Frozen weights prevent the labels already collected from reducing that episode's risk. Failure changes the trajectory distribution and limits the next update.

We analyze a uniform initial distribution over a class containing the expert. After each episode, a policy's weight is multiplied by the likelihood of the independent expert labels observed at the visited states. For two stochastic policies, the worst-case cumulative squared Hellinger loss between the learner's and expert's full trajectory laws has *iterated-logarithm* dependence on episode length. This is the number of repeated logarithms needed to bring the length below one. The two-state construction attains the order, and a matching upper bound holds in every environment under this feedback model. Retaining one policy per episode or updating after each label removes this dependence. A uniform cumulative return-gap law of the same form is impossible; Section 3 gives the finite-round return implication.

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

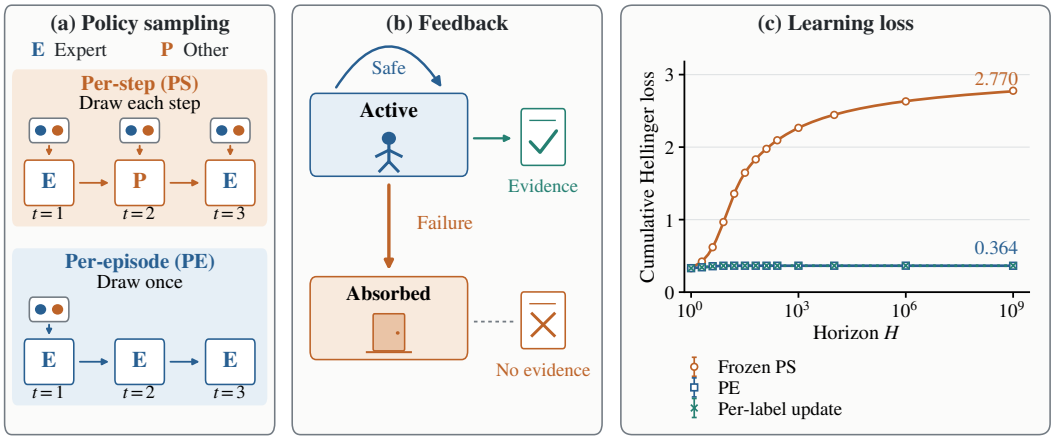

Figure 1: Deployment affects feedback and trajectory loss. (a) Example policy draws. (b) Failure ends informative visits. (c) Cumulative Hellinger loss. Curves: exact losses; markers: simulations with pointwise 95% intervals (Section 8).

The finite-horizon effect is visible in controlled simulations. Across 13 episode lengths, we compare the three rules using 100,000 independent learning runs per setting, with the same policies, prior, and rule for obtaining labels. Each learner collects labels on its own trajectories. Frozen mixing has exact cumulative loss 2.27 at $H = 10^3$ and 2.77 at $H = 10^9$, whereas both alternatives have loss 0.364 to three decimals. The difference is a factor of six to eight despite the slow growth: even at $H = 2^{1024}$, the exact frozen loss is only 3.33. Further experiments vary update timing and strength, test stochastic classes with more candidates, and compare deterministic deployments under both feedback protocols.

Our contributions are threefold.

1. Trajectory discrepancy under frozen mixing. For two stochastic policies, worst-case cumulative trajectory Hellinger loss has iterated-logarithm order, uniformly over the number of learning rounds. A two-state construction and a potential for the posterior odds give matching lower and upper bounds (Theorem 1).

2. The effects of deployment and update timing. We characterize how deployment and update timing change Hellinger loss. Per-episode sampling and per-label updating remove the horizon dependence. Batch updates make the interval between updates the relevant scale; likelihood tempering adds loss from weaker updates (Theorems 2 and 3, and Theorem 4).

3. Deterministic laws and the WARM-STAGGER conjecture. For deterministic classes, we give matching return-gap laws. The cost of resampling depends on both class size and horizon after offline demonstrations, and the same dependence persists for the averaged iterate with full-trajectory interactive labels (Theorem 5; Section 6). These results also resolve the class-size conjecture for WARM-STAGGER's per-step rule. A deterministic instance contradicts the proposed improvement. Under the same single-state query model, plurality deployment achieves the proposed rates in expectation (Theorem 7; Proposition 8).

2 PROBLEM SETUP

An episode consists of H steps in a Markov decision process (MDP). Its state space $\mathcal{S}$ is finite or countable and contains the time index; stationary policies therefore include time-dependent behavior. The remaining components are a finite action set $\mathcal{A}$, an initial distribution ρ, and transition kernels. We normalize rewards by requiring the total return of every trajectory to lie in $[0, 1]$. For a policy π, $J(\pi)$ denotes its expected total return.

Let Π be a finite class of stationary stochastic policies containing the expert $\pi^\star$, with uniform prior w_1. Deterministic policies are included as point masses. Write B for the class size and $M = B - 1$

for the number of non-expert policies. We assume at least two candidates; the singleton case has zero loss. All rollouts use the same initial law and transition kernels.

Under per-step mixing (PS), the learner independently draws a candidate from the current weights before each action and samples that candidate's action at the visited state. In round t, the deployed policy is $\mu_t(\cdot \mid s) = \sum_\pi w_t(\pi)\pi(\cdot \mid s)$. Under per-episode sampling (PE), it draws once from the same weights and follows that candidate throughout. Weights stay fixed within an episode unless an update rule explicitly says otherwise.

The stochastic results measure squared Hellinger distance $\mathsf{h}^2(P, Q) = 1 - \int \sqrt{dP\, dQ}$ between state-action trajectory laws and receive independent expert labels at visited states. The deterministic results measure expected return shortfall after offline demonstrations or interaction. Full-trajectory annotation labels every visited state; the single-state protocol queries a uniformly chosen time in each episode. Each result specifies which protocol it uses.

Class-size dependence is expressed through

$$h := \min\{H, M\}, \qquad A_H(M) := h\Big(1 + \log \frac{M}{h}\Big) = \begin{cases} H\big(1 + \log \frac{B-1}{H}\big), & H < B - 1, \\ B - 1, & H \geq B - 1, \end{cases}$$

All logarithms are natural. The bounds $A_H(M) \leq \min\{M,\, H(1 + \log M)\}$ and $A_H(M) \geq \min\{M, H\}$ locate this quantity between the shorter of the horizon and class size and its logarithmic enlargement.

3 How frozen weights limit feedback

Can a fixed class continue to incur trajectory Hellinger loss as episodes grow longer? For stochastic candidates, a label can lower a wrong policy's weight without eliminating it. The relevant state of learning is therefore a likelihood-weighted distribution rather than a set of surviving policies. For example, the NAIL algorithm of Sriraman et al. (2026) raises trajectory label likelihoods to the power $1/H$; using the likelihoods without this exponent gives the untempered Bayesian posterior. The corresponding deployment rule is the pointwise mixture used in on-policy distillation.

The remaining weight matters because a wrong policy can take the learner away from informative states. Freezing that weight prevents labels collected earlier in the episode from changing later actions. We show that this feedback mechanism produces an *iterated logarithm* dependence on the horizon with just two policies.

Under the Bayesian protocol, each episode supplies an independent expert label $a'_h \sim \pi^\star(\cdot \mid s_h)$ at each of the H visited states, then updates $w_{t+1}(\pi) \propto w_t(\pi) \prod_h \pi(a'_h \mid s_h)$ after the episode. Write $\mathrm{CUM}(n, H) = \sup \sum_{t=1}^n \mathbb{E}\, \mathsf{h}^2(P^{\pi^\star}, P^{\mu_t})$ for the worst-case cumulative loss. Here the supremum permits MDPs with a finite action set, a countable state space with the time index adjoined, arbitrary transitions, and returns in $[0, 1]$, and over two-policy classes $\{\pi^\star, \pi^1\}$ of stationary stochastic policies. The protocol in this definition fixes three choices: an untempered whole-trajectory Bayesian posterior, a pointwise mixture, and labels at all visited states. Classical Bayesian sequence prediction bounds cumulative predictive divergence using the prior mass of the truth and updates after each observed symbol (Hutter, 2003). Here the frozen mixture determines the trajectory and can end its own informative visits; that sequential prediction guarantee does not directly control this trajectory loss.

For clarity, the iterated logarithm is defined using natural logarithms: $\log^\star H = \min\{j \geq 0 : \log^{\circ j} H < 1\}$.

Theorem 1 (Iterated-logarithm law for two policies)**.** *For every integer $H \geq 1$,*

$$\tfrac{1}{8} \log^\star H \;\leq\; \sup_{n \geq 1} \mathrm{CUM}(n, H) \;\leq\; 56\,(1 + \log^\star H),$$

and the instance of Proposition 9 below has infinite-round cumulative loss $S_H \geq \max\{1 - \frac{\sqrt{3}}{2},\, \frac{7}{8}(\log^\star H - 2)_+\}$. The upper bound holds for every two-policy stochastic class and every MDP.

The stochastic laws here and in Section 4 measure trajectory discrepancy. For returns in $[0,1]$, Theorem 1 gives the finite-round bound

$$\sum_{t=1}^{n} \mathbb{E}\,|J(\pi^\star) - J(\mu_t)| \le \sqrt{112\,n(1+\log^\star H)}. \tag{1}$$

A cumulative return-gap bound uniform over both rounds and two-policy stochastic classes is impossible even at $H = 1$ (Appendix B.7). On the two-state construction, awarding one for an episode without failure instead gives an iterated-logarithm cumulative return gap for frozen mixing and constant cumulative gaps for per-episode sampling and per-label updating (Corollary 14). This return separation is specific to the construction and its terminal reward.

The example randomizes the expert between safe actions. The mechanism persists with a deterministic expert. An expert that always chooses one safe action, paired with a wrong policy that plays † with probability $\frac{1}{2}$, gives the same likelihood ratio and recursion. Thus the wrong policy's randomization allows it to retain weight after repeated informative labels. For the matching upper bound, a supermartingale controls the square root of the posterior odds. A potential allocates one term to each exponential-tower level, paying for the loss incurred before the odds pass that scale.

Appendix B gives the two-state construction and the complete tower-potential argument.

4 Controlling trajectory loss through deployment and updates

The construction uses both repeated policy draws and weights that cannot respond to labels within the episode. We can change either choice. Sampling once preserves the policy choice across steps; updating after each label lets the evidence affect the next action. For two stochastic policies, each change yields a cumulative Hellinger loss bound independent of the horizon under the same expert-label feedback rule.

Theorem 2 (Per-episode sampling and within-episode updating)**.** *In the two-policy setting of Theorem 1, let $\mathcal{C}_{\mathrm{PE}}(H)$ be the worst-case infinite-round cumulative loss when one policy $\pi \sim w_t$ is drawn at the start of each episode and followed throughout (deployed law $P^{(t)} = (1-w_t)P^{\pi^\star} + w_t P^{\pi^\dagger}$), and $\mathcal{C}_{\mathrm{WU}}(H)$ that of pointwise mixing when the posterior is updated after every label (the expert label at the state visited at step h is revealed after the action at step h is taken; the action at step h uses the posterior given the labels of steps $1,\dots,h-1$ of the current episode and all earlier episodes). Then for every $H \ge 1$,*

$$1 - \tfrac{1}{\sqrt{2}} \;\le\; \mathcal{C}_{\mathrm{PE}}(H) \;\le\; \frac{11}{1-e^{-1}} < 18, \qquad 1 - \tfrac{1}{\sqrt{2}} \;\le\; \mathcal{C}_{\mathrm{WU}}(H) \;\le\; 5\log 2 < 3.47.$$

4.1 Likelihood tempering and trajectory loss

Does reducing the strength of the update control the loss? Tempering raises a likelihood to a power below one. Each informative label then moves the odds less, delaying the reduction in the wrong policy's weight. For example, NAIL of Sriraman et al. (2026) uses the trajectory likelihood exponent $1/H$. Here we use that exponent within the clean Bayesian protocol. The contaminated labels, query model, and loss in that work are different.

Consider the tempered update $r_{t+1} = r_t \Lambda_t^\alpha$, $\alpha \in (0,1]$, with deployment, feedback, and loss unchanged. Here the odds compare the alternative policy to the expert, and the episode likelihood ratio is the product of the alternative-to-expert likelihood ratios of its labels. Let $\mathcal{C}_\alpha(H)$ denote the corresponding worst case over two-policy classes and MDPs. The cumulative loss has additive dependence on $1/\alpha$ and $\log^\star H$. In particular, tempering by $1/H$ makes its worst-case order linear in H.

Theorem 3 (Tempered law)**.** *For every integer $H \ge 1$ and $\alpha \in (0,1]$, with $k = \log^\star H$,*

$$\tfrac{1}{12}\Big(\frac{1}{\alpha} + k\Big) \;\le\; \mathcal{C}_\alpha(H) \;<\; 56\Big(\frac{1}{\alpha} + k\Big),$$

and the instance of Proposition 9 has $S_{\alpha,H} \ge \max\{\frac{1}{4\alpha}, \frac{k}{8}\}$. In particular $\frac{H}{4} \le \mathcal{C}_{1/H}(H) < 112H$, and for that instance $S_{1/H,H}/H \to \frac{1}{2}$; more precisely $\frac{H}{2} - \frac{4}{\log 2} \le S_{1/H,H} \le \frac{H}{2} + 2\log H + 2\log 2 + 1$ for $H \ge 2$. For any sequence $\alpha_H \in (0,1]$, $\mathcal{C}_{\alpha_H}(H) = \Theta(\max\{\alpha_H^{-1}, \log^\star H\})$: the transition between the two regimes is at $\alpha_H \asymp 1/\log^\star H$.

4.2 How long may the weights remain fixed?

Freezing the weights for an entire episode gives $\Theta(\log^\star H)$ worst-case cumulative loss; updating after each label gives $\Theta(1)$. Between these endpoints, the learner may refresh every L steps. The tower height then depends on this interval, uniformly over the length of the episode.

Theorem 4 (Batch-length law)**.** *Let $1 \le L \le H$, partition each episode into batches of lengths $\ell_j = \min\{L, H - jL\}$, $0 \le j < \lceil H/L \rceil$, and refresh the Bayesian mixture after each batch (labels of a batch are used from the next batch on). Then, uniformly in $H \ge L$, the worst-case cumulative loss $\mathcal{C}_L(H)$ over two-policy classes and MDPs satisfies*

$$\tfrac{1}{8}\log^\star L \;\le\; \mathcal{C}_L(H) \;<\; \min\bigl\{56(1+\log^\star L),\; 5L\log 2\bigr\}$$

(the second upper bound is the informative one only for small L), and the self-truncating instance has $S_{L,H} \ge \max\{1 - \frac{\sqrt{3}}{2}, \frac{7}{8}(\log^\star L - 2)_+\}$.

The proofs and the quantitative protocol comparison are in Appendix B.

5 How much does per-step mixing cost?

We now specialize to deterministic candidates and return gaps. The same labels that reduce stochastic likelihoods eliminate inconsistent deterministic policies.

A *deterministic policy* chooses one action at each state: $\pi : \mathcal{S} \to \mathcal{A}$. In this section the expert $\pi^\star$ is deterministic. The learner knows a finite class $\mathcal{B}$ of deterministic policies, and we assume $\pi^\star \in \mathcal{B}$. Realizability thus means that the expert is one of the known candidates. Write $B = |\mathcal{B}|$ for the class size and $M = B - 1$ for the number of non-expert policies, indexed $\pi_1, \dots, \pi_M$. We consider classes with at least two policies; a singleton class incurs zero loss.

Before interaction, the learner observes n independent expert trajectories $\tau^{(1)}, \dots, \tau^{(n)}$ with the same distribution. These trajectories record the states visited by $\pi^\star$ and the expert action at each state. The candidates agreeing with every recorded action form the *version space*,

$$\mathcal{V} = \{\pi \in \mathcal{B} :\; \pi(s) = \pi^\star(s) \text{ for every state } s \text{ visited in the data}\},$$

which contains $\pi^\star$ by construction. Write $q = |\mathcal{V}|$.

Under per-step mixing (PS), the learner draws a fresh uniformly random $\pi \in \mathcal{V}$ at each step, independently of its earlier draws, and takes action $\pi(s)$ at the current state s. Its stochastic policy is therefore $\mu_{\mathcal{V}}(a \mid s) = |\{\pi \in \mathcal{V} : \pi(s) = a\}|\,/q$. This is the deployment rule of Warm-Stagger (Li & Zhang, 2025, Def. 1) and the deterministic instance of the "pointwise mixture" in noisy-expert distillation (Sriraman et al., 2026).

Under per-episode sampling (PE), the learner draws a uniformly random $\pi \in \mathcal{V}$ once, at the start of the episode, and uses that policy at every step. This is the Gibbs rule of probably approximately correct (PAC) learning, applied with trajectories as examples.

For either deployment rule $\mathrm{R} \in \{\mathrm{PS}, \mathrm{PE}\}$, we measure the expected return lost relative to the expert, $\mathrm{GAP}_{\mathrm{R}} = \mathbb{E}_{\text{data}}\bigl[J(\pi^\star) - J(\pi_{\mathrm{R}}(\mathcal{V}))\bigr]$. The *worst-case gap* $\mathcal{R}_{\mathrm{R}}(B, H, n)$ takes the supremum over all MDPs, policy classes, and experts satisfying the assumptions above. The quantity $A_H(M)$ from Section 2 controls the worst-case gap.

Theorem 5 (Worst-case offline gaps)**.** *Let $B \ge 2$, $H \ge 1$ and $n \ge 1$. Then*

$$\tfrac{1}{864}\min\Bigl\{1, \tfrac{A_H(M)}{n}\Bigr\} \;\le\; \mathcal{R}_{\mathrm{PS}}(B, H, n) \;\le\; \min\Bigl\{1, \tfrac{A_H(M)}{n}\Bigr\}, \tag{2}$$

$$\tfrac{1}{16}\min\Bigl\{1, \tfrac{\log B}{n}\Bigr\} \;\le\; \mathcal{R}_{\mathrm{PE}}(B, H, n) \;\le\; \min\Bigl\{1, \tfrac{2\log B}{n}\Bigr\}. \tag{3}$$

For $n = 0$ the exact values are $\mathcal{R}_{\mathrm{PE}}(B, H, 0) = 1 - 1/B$ and $\mathcal{R}_{\mathrm{PS}}(B, H, 0) = 1 - B^{-H}$ (attained when $|\mathcal{A}| \ge B$; for smaller action sets they remain upper bounds). Consequently, uniformly in $B \ge 2$, $H \ge 1$, $n \ge 0$,

$$\mathcal{R}_{\mathrm{PS}}(B, H, n) = \Theta\Bigl(\min\Bigl\{1, \tfrac{A_H(B-1)}{n+1}\Bigr\}\Bigr), \qquad \mathcal{R}_{\mathrm{PE}}(B, H, n) = \Theta\Bigl(\min\Bigl\{1, \tfrac{\log B}{n+1}\Bigr\}\Bigr).$$

The upper bounds follow from the surviving policies' disagreement probabilities; Appendix D constructs instances that match them up to constants. Under per-episode sampling, the class contributes $\log B$. Resampling changes this term to $A_H(B-1) = \min\{H, B-1\}(1 + \log \frac{B-1}{\min\{H,B-1\}})$. The resulting factor can reach H when $B \gg H$, or $(B-1)/\log B$ when $B \le H$. The difference already appears with the data fixed: the two rules encounter the same expert trajectory, but have different opportunities to select a policy that disagrees along it.

Appendix C gives the complete upper-bound proof.

6 What interaction removes

After each episode, the learner removes every candidate that disagrees with an expert label at a visited state, then deploys uniformly over the remaining version space using PS or PE. A potential charges the episode's deviation probability to these removals.

The potential bound can be combined with the offline guarantee by returning a uniformly chosen pre-update interactive iterate. The total number of episodes then determines the rate, while the class-size and horizon dependence still records which deployment rule was used.

Theorem 6 (Hybrid laws for the averaged iterate)**.** *Let $n \ge 0$ offline episodes be followed by $N \ge 1$ interactive episodes, $T = n + N$, and let the output be the rule deployed from a uniformly chosen one of the N pre-update interactive version spaces. Then*

$$\tfrac{1}{864}\min\Big\{1, \tfrac{A_H(M)}{T}\Big\} \le \mathcal{R}^{\mathrm{avg}}_{\mathrm{PS}} \le \min\Big\{1, \tfrac{2A_H(M)}{T}\Big\}, \qquad \tfrac{1}{16}\min\Big\{1, \tfrac{\log B}{T}\Big\} \le \mathcal{R}^{\mathrm{avg}}_{\mathrm{PE}} \le \min\Big\{1, \tfrac{3\log B}{T}\Big\}.$$

The final iterate needs a separate argument. Appendix E supplies the potential proof, the full hybrid proof, and the remaining gap between its final-iterate bounds.

7 Consequences for Warm-Stagger

The deployment distinction settles a proposed improvement to Warm-Stagger. The issue is whether the offline term can depend only logarithmically on the number of base policies. Li & Zhang (2025) study hybrid imitation learning with N_{off} offline expert trajectories and N_{int} interactive rounds. Their annotation oracle labels one sampled state in each round.

For deterministic bases, its exponential weights are uniform on the candidates that remain consistent with the offline data. The first deployed policy is therefore our per-step mixture. The rare-context construction gives the following consequence; Appendix F records the original guarantee and the exact algorithmic reduction. In the following display, $n = N_{\mathrm{off}} = m^2$ is the number of offline trajectories and K counts the rare contexts absent from those trajectories.

Theorem 7 (The conjecture is false for Warm-Stagger)**.** *There is no universal constant C such that* Warm-Stagger*, run as specified, satisfies $J(\pi^E) - J(\widehat{\pi}) \le C\,R\log(B/\delta)/N_{\mathrm{off}}$ (for any fixed $\delta \in (0,1)$, or in expectation) on every instance satisfying the hypotheses of their Theorem 6. Concretely, on the rare-context instance of Section D.4 with $H = m$, $B = m+1$, $N_{\mathrm{off}} = m^2$, $N_{\mathrm{int}} = 1$, $R = \mu = 1$, the output is $\pi^1 = \mu_{\mathcal{V}}$ and*

$$J(\pi^E) - J(\widehat{\pi}) = \frac{K}{n}\Big[1 - \Big(\frac{K}{K+1}\Big)^m\Big], \qquad m\big(J(\pi^E) - J(\widehat{\pi})\big) \xrightarrow[m\to\infty]{\mathbb{P}} e^{-1}\big(1 - e^{-e}\big) = 0.3436\ldots,$$

so the gap is $\Theta(1/m) = \Theta(H/N_{\mathrm{off}})$ while the conjectured bound is $O(\log(m/\delta)/m^2)$. More generally, for deterministic base classes the worst-case expected *offline term of* Warm-Stagger*'s first iterate is $\Theta(\min\{1, A_H(B-1)/N_{\mathrm{off}}\})$ (Theorem 5), which lies between the conjectured $\log B/N_{\mathrm{off}}$ and the proved $H\log B/N_{\mathrm{off}}$ up to constants: the factor lost is $\min\{H, B-1\}(1 + \log\frac{B-1}{\min\{H,B-1\}})/\log B$.*

The separation persists beyond $N_{\mathrm{int}} = 1$. Each interactive round labels one state and removes at most one non-expert policy on this instance, so the asymptotics persist for $N_{\mathrm{int}} = o(m)$. The feedback models must be kept distinct: Warm-Stagger labels one state per round and has a high-probability guarantee. Theorem 6 labels the full trajectory and bounds expected gaps. Under full-trajectory annotation, the averaged-iterate rate of per-step mixing is $\Theta(\min\{1, A_H(B-1)/(N_{\mathrm{off}} + N_{\mathrm{int}})\})$.

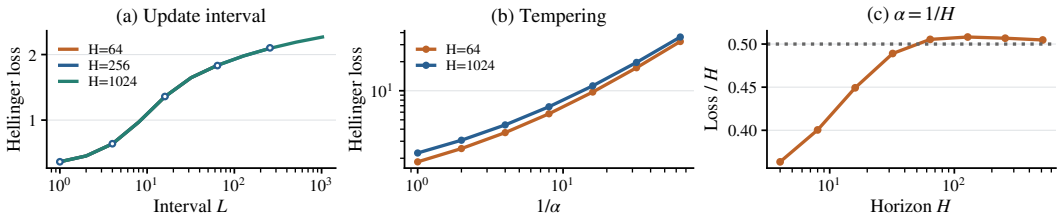

Figure 2: Update sensitivity: (a) frozen interval, (b) likelihood exponent, (c) inverse-horizon tempering. Curves: exact losses; dotted: analytic limit; open markers: 20,000-run estimates at horizon 256 with pointwise 95% intervals.

The conjectured $\log B$ dependence holds for per-episode sampling by Theorem 5. Under single-state feedback, deterministic plurality voting achieves the corresponding *expected* rates while keeping WARM-STAGGER's single-state query model and its state-wise online term. To compare directly to class size, we state the bound with $\log B$ in place of the conditional $\log B_{\mathrm{bc}}$; the guarantee is in expectation rather than with high probability.

Proposition 8 (Plurality vote restores the conjectured guarantee)**.** *Let the bases be deterministic and let the learner deploy, in round n, the plurality policy $v_n(s) = \arg\max_a |\{\pi \in \mathcal{V}_n : \pi(s) = a\}|$ (ties broken by a fixed rule), keeping* WARM-STAGGER*'s state query and version-space update, with $N_{\mathrm{int}} \geq 1$. Assume returns in $[0, R]$ and μ-recoverability in the reward-shortfall sense $V_h^{\pi^E}(s) - Q_h^{\pi^E}(s, a) \leq \mu$ for all h, s, a (the convention needed for the online argument of Li & Zhang (2025); their printed definition has the opposite sign, which we read as a typo). Then, in expectation over data and rollouts, the first-step uniform mixture $\widehat{\pi}$ of $v_1, \ldots, v_{N_{\mathrm{int}}}$ satisfies (omit the first term if $N_{\mathrm{off}} = 0$)*

$$\mathbb{E}\big[J(\pi^E) - J(\widehat{\pi})\big] \;\leq\; \min\Big\{\frac{4R\log B}{N_{\mathrm{off}}},\; \frac{\mu H \log_2 B}{N_{\mathrm{int}}}\Big\}.$$

These deployments are classical. The TRAGGER algorithm of Li & Zhang (2025) uses first-step mixtures and trajectory labels for a $\log B$ guarantee; NAILGUN (Sriraman et al., 2026) uses weighted plurality for noisy experts. Our comparison isolates the price of resampling within an episode: its offline dependence can grow from $\log B$ to $A_H(B-1)$.

Both proofs appear in Appendix F.

8 EXPERIMENTS

We compare exact recursions with independent learning histories on explicit MDPs. Appendix H gives the instances, grids, estimators, and code. Return and Hellinger measurements are distinguished below.

8.1 TRAJECTORY LOSS AND TERMINAL-SUCCESS RETURNS

On the two-state instance, we compare frozen per-step mixing (PS), per-episode sampling (PE), and per-label updating (WU) over 13 horizons, with 100,000 independent histories per setting. WU receives each label after the current action. All 39 Hellinger estimates are within 0.008 of their exact references (Figure 1; Table 3). A separate experiment awards one for an episode without failure. From one thousand to one billion steps, exact cumulative return gaps rise from 2.331 to 2.779 for PS, while PE and WU remain at 0.632 to three decimals. Independent 100,000-run estimates differ from the return references by less than 0.0037 (Figure 7, Appendix H). These reward-specific results illustrate Corollary 14.

8.2 UPDATE TIMING AND LIKELIHOOD STRENGTH

Figure 2 covers 49 update settings, with eight checks using 20,000 independent histories each. At horizon 1,024, refreshing every 1, 16, 64, or 256 steps gives losses 0.364, 1.358, 1.836, and 2.095;

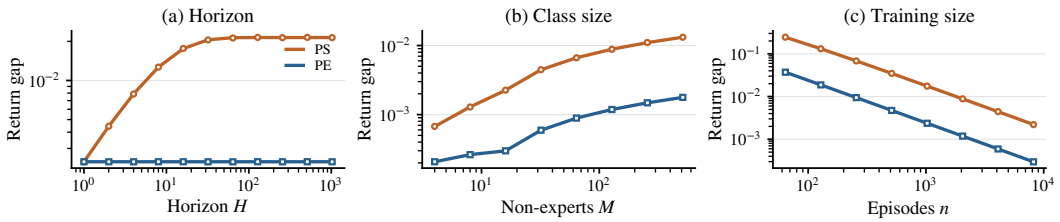

Figure 3: Bernoulli deployment gaps versus (a) horizon, (b) class size, and (c) training size. Disagreement probability is fixed in (a), recalibrated in (b,c). Curves: exact gaps; markers: simulations with pointwise 95% intervals.

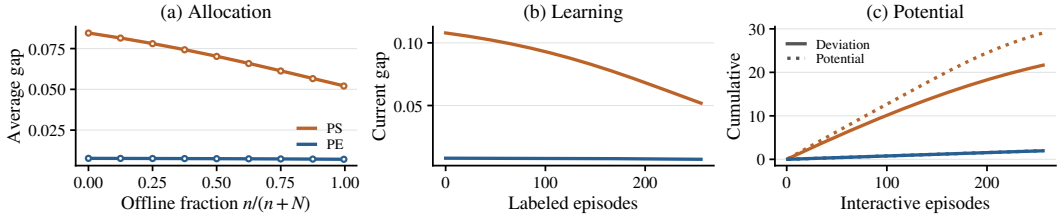

Figure 4: Full-trajectory learning: (a) allocation of 256 episodes, (b) current-iterate gaps, (c) cumulative deviation and potential decrease. Curves: exact values; markers: simulations with pointwise 95% intervals.

freezing for the whole episode gives 2.268. Shared intervals give nearly identical losses across horizons, as suggested by Theorem 4.

At horizon 1,024, lowering the likelihood exponent from one to one sixty-fourth raises loss from 2.268 to 36.009. Inverse-horizon tempering gives loss-to-horizon ratios 0.5082 and 0.5047 at horizons 128 and 512, approaching the limit in Theorem 3.

8.3 Deterministic resampling amplifies surviving errors

Figure 3 uses 20,000 Bernoulli training histories per setting. On the fixed instance in (a), PE stays at 0.00237 while PS rises to 0.02140 as the horizon grows from one to 1,024. The fixed disagreement law causes saturation. Panels (b,c) recalibrate the disagreement probability at each class or sample size; each PS/PE pair shares its instance. For 128 non-experts and horizon 16, increasing demonstrations from 64 to 8,192 reduces PS gaps from 0.2449 to 0.002216 and PE gaps from 0.03714 to 0.0002964. Each setting is a different member of the lower-bound family.

8.4 Interaction and the allocation of demonstrations

On a fixed Bernoulli instance, we divide 256 episodes between offline data and full-trajectory interaction (Figure 4). A shared set of 20,000 independent histories checks nine allocations. All-interactive average gaps are 0.08457 for PS and 0.007673 for PE; with 255 offline episodes they are 0.05215 and 0.007102. The final survivor law is unchanged, but later-starting averages include better-informed iterates. The common worst-case order in Theorem 6 allows this finite-instance difference. At every interactive round, expected potential decrease dominates cumulative deviation, as Theorem 20 requires.

8.5 Rare contexts and the location of a query

The rare-context instance tests the Warm-Stagger refutation (Figure 5). At 128 contexts, exact PS and PE gaps are 0.002670 and 0.00005975, a factor of 44.68. Their normalized values, 0.3417 and 0.9789, approach the respective limits 0.3436 and one. Appendix H gives the occupancy law and feedback comparison.

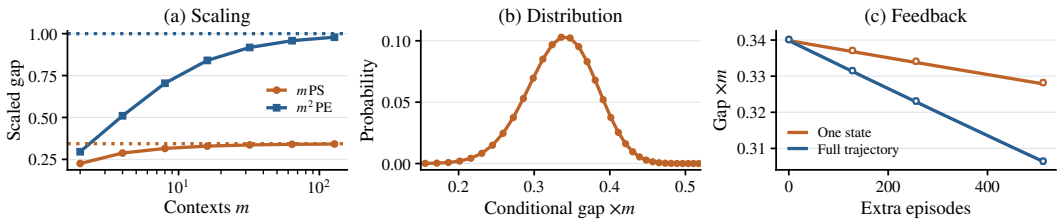

Figure 5: Rare-context deployment: (a) normalized gaps, (b) conditional-gap distribution, (c) single-state versus full-trajectory feedback. Curves: exact values; dotted: limits; markers in (c): simulations with pointwise 95% intervals.

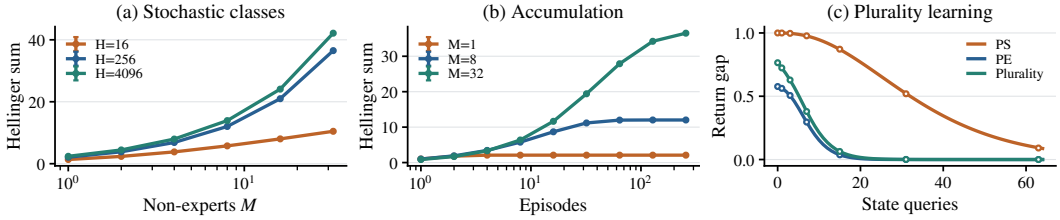

Figure 6: (a,b) Stochastic context copies over up to 256 rounds. (c) Plurality with single-state queries. Markers: 20,000-run estimates with pointwise 95% intervals; curves in (c): exact gaps.

Starting from 23.54 surviving non-experts at 64 contexts, 512 further episodes leave 22.52 with single-state queries and 20.77 with full-trajectory labels. A query after failure is uninformative, while a full trajectory always includes the initial state. This comparison fixes episode count; the two protocols use different label budgets (Table 6).

8.6 Stochastic classes and nontrivial plurality learning

We extend the stochastic construction to one informative context per non-expert, as in Theorem 13. Figure 6(a,b) uses 20,000 independent histories at each of 18 settings. At horizon 256, increasing the number of non-experts from one to 32 raises the estimated 256-round Hellinger sum from 2.093 to 36.510. Each context updates only its own posterior odds ratio. These finite-round measurements do not resolve the uniform worst-case order.

For plurality, we use 64 Bernoulli non-experts with disagreement probability 0.6 and one offline demonstration. Every rule then receives one uniformly timed state query per episode. Plurality has initial gap 0.765, so the vote is not trivially correct. At horizon 16, the exact average gaps over 256 rounds are 0.14019 for PS, 0.01857 for PE, and 0.02447 for plurality (panel (c)). Plurality improves on PS here, while PE learns faster. Across four horizons, 20,000 independent histories per rule validate the survivor recursion. Appendix H reports the query-mistake counts and full results.

The experiments use explicit synthetic MDPs; the proofs establish the worst-case rates.

9 Conclusion

Frozen per-step mixing can end its own informative visits. For two stochastic policies under clean-label Bayesian updates, this yields a sharp iterated-logarithm law for cumulative squared trajectory Hellinger loss. A uniform cumulative return-gap law does not follow: uniform return bounds over stochastic classes must depend on the number of rounds. The two-state construction nevertheless gives an actual return separation under terminal-success rewards, while the deterministic results establish matching return-gap laws. Deployment, feedback, and the loss metric must therefore be specified together. Sharp rates for larger stochastic classes and final interactive iterates remain open, as does the necessity of the horizon and recoverability dependence under single-state feedback (Appendix G).

AI USE STATEMENT

We used generative AI tools primarily for experimental code development and manuscript polishing, with additional assistance in theoretical analysis, proof drafting, and experimental design and interpretation. The authors are responsible for the final content.

ETHICS STATEMENT

This work studies theoretical imitation-learning models and controlled synthetic experiments. It involves no human participants, personal data, or deployment in real-world decision systems. The results are limited to the stated policy classes, feedback protocols, and loss functions; they do not establish safety or reliability for deployed imitation-learning systems.

REPRODUCIBILITY STATEMENT

Appendices B through F give the proofs and their assumptions. Appendix H specifies the experimental settings, estimators, seeds, and numerical checks, including the terminal-return study and the stochastic-class and voting experiments in Section H.8 and the following subsections. The accompanying anonymous code includes simulation scripts, saved results, dependency versions, and figure and table renderers; its README gives reproduction commands. The experiments require no external datasets or pretrained models.

REFERENCES

Michael K. Cohen, Marcus Hutter, and Neel Nanda. Fully general online imitation learning. *Journal of Machine Learning Research*, 23(334):1–30, 2022.

Dylan J. Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2407.15007.

Dylan J. Foster, Yanjun Han, Jian Qian, and Alexander Rakhlin. Online estimation via offline estimation: An information-theoretic framework. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024b.

Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264, 1953.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Marcus Hutter. Optimality of universal Bayesian sequence prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003.

Yichen Li and Chicheng Zhang. On efficient online imitation learning via classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.12868.

Yichen Li and Chicheng Zhang. Interactive and hybrid imitation learning: Provably beating behavior cloning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2412.07057v3.

Nived Rajaraman, Lin F. Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *AISTATS*, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ved Sriraman, Peihan Liu, Daniel Hsu, and Adam Block. Behavior cloning is not all you need: The optimality of on-policy distillation for noisy expert feedback, 2026. arXiv:2606.30923.

A PROTOCOL DISTINCTIONS AND RELATED WORK

This appendix places the guarantees side by side and records the distinctions needed when comparing them to earlier work. A distribution over base policies can be deployed by drawing once per episode or anew at every step. Both choices arise in interactive imitation learning in the DAgger family. For a finite class containing a realizable expert, the *version space* $\mathcal{V}$ contains the policies consistent with all expert labels seen so far. Natural online learners, including exponential weights with log loss and Bayesian aggregation, assign uniform or posterior weight to $\mathcal{V}$.

The first deployment rule is a first-step mixture, also called the Gibbs rule in PAC learning. The second is an each-step mixture, equivalently the stochastic policy $\mu_{\mathcal{V}}(a \mid s) = |\{\pi \in \mathcal{V} : \pi(s) = a\}| \,/\, |\mathcal{V}|$. Li and Zhang run exponential weights over the each-step mixing class in STAGGER and WARM-STAGGER. Bayesian aggregation in on-policy distillation from noisy experts also deploys a pointwise mixture. The analysis of Li & Zhang (2025) pays a factor $\log \widetilde{B}$, where $\widetilde{B} \le B^H$ is the size of a finite step-wise completion of the each-step mixing class. The conjectured improvement replaces that factor by $\log B$.

The deterministic analysis gives exact worst-case orders for version-space deployment (Sections 5 through 7). For stochastic classes, we study pointwise mixing of a whole-trajectory Bayesian posterior, optionally tempered, with expert labels at every visited state (Section 3). Write $M = B - 1$ for the number of non-expert policies and $A_H(M) = \min\{H, M\}\big(1 + \log \frac{M}{\min\{H,M\}}\big)$. Tables 1 and 2 separate return-gap guarantees from cumulative trajectory-loss guarantees. These choices are part of the statements and cannot be interchanged.

Deployment	Data and output	Expected return gap	Result
Per-step mixture	Offline n trajectories	$\Theta(\min\{1, A_H(M)/(n+1)\})$	Thm. 5
One policy per episode	Offline n trajectories	$\Theta(\min\{1, \log B/(n+1)\})$	Thm. 5
Either	$+N$ full-trajectory rounds; averaged iterate	Same laws with $n + N$	Thm. 6
WARM-STAGGER	Offline data; one state query per round	$\Theta(H/N_{\mathrm{off}})$ on the counterexample	Thm. 7
Plurality vote	Same single-state protocol	$\min\{\frac{4R\log B}{N_{\mathrm{off}}}, \frac{\mu H \log_2 B}{N_{\mathrm{int}}}\}$	Prop. 8

Table 1: Deterministic classes: deployment, feedback, and expected return gaps.

Deployment	Weight updates	Cumulative trajectory h^2	Result
Per-step mixture	After each episode	$\Theta(\log^\star H)$	Thm. 1
Per-step mixture	Likelihood$^\alpha$; after episode	$\Theta(1/\alpha + \log^\star H)$	Thm. 3
One policy per episode	After each episode	$\Theta(1)$	Thm. 2
Per-step mixture	After each label	$\Theta(1)$	Thm. 2
Per-step mixture	Every L steps	$\Theta(\log^\star L)$	Thm. 4
Per-step mixture; $M+1$ policies	After each episode	$\Theta(A_H(M))$ if $M \ge H^2$; $\Theta(M \log^\star H)$ under the theorem's tower-height condition	Thm. 13

Table 2: Stochastic classes: expected cumulative trajectory Hellinger losses with full-trajectory labels and a uniform prior. Two policies unless stated otherwise.

The deterministic potential counts eliminations. For stochastic policies, a positive likelihood can preserve a candidate after every label, so this count need not decrease. In the two-state MDP, a two-policy class has cumulative Hellinger loss tending to n as $H \to \infty$. A catastrophic action ends the mixture's informative visits, and its posterior odds decrease only through the tower

recursion $B_{t+1} \approx B_t + 2^{B_t}$. The worst-case upper and lower bounds have the same order: $\sum_{t \le n} \mathbb{E}\, \mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \le 56(1 + \log^\star H)$ for every n, every two-policy stochastic class, and every MDP, while the instance has infinite-round cumulative loss at least $\frac{1}{8} \log^\star H$. Its guaranteed loss increases by $\frac{7}{8}$ at each level of the tower $e, e^e, e^{e^e}, \ldots$ The upper bound follows from a Bhattacharyya supermartingale for the square root of the posterior odds and the *tower potential* $F_H(r) = \sqrt{r} + \sum_{j \le m} \min\{1, \sqrt{r e^{x_j}}\}$, $x_{j+1} = e^{x_j}$. Each of its $m + 1 \le 1 + \log^\star H$ units pays for one round of constant loss.

With the same prior and expert-label rule, sampling one policy per episode or updating the mixture after every label gives cumulative loss $\Theta(1)$ for every H. Refreshing every L steps gives worst-case cumulative loss $\Theta(\log^\star L)$ uniformly in $H \ge L$. The duration for which weights stay fixed is therefore the scale controlling the tower height. Tempering the likelihood by α, as NAIL does with $\alpha = 1/H$, gives $\Theta(1/\alpha + \log^\star H)$ through the rescaled potential $F_H(r/\alpha^2)$. Consequently, $1/H$-tempering costs $\Theta(H)$, with the transition between regimes at $\alpha \asymp 1/\log^\star H$. For $M + 1$ stochastic policies (Theorem 13), summing tower potentials and using an entropy potential gives $\mathcal{C}_M(H) \le \min\{56M(1 + \log^\star H), 11H \log(M + 1)\}$. The lower-bound constructions imply $\mathcal{C}_M(H) \ge \max\{A_H(M)/16, \frac{7}{8} M(\log^\star H - \log^\star(16(M+1)) - 1)_+\}$. The uniform order between the resulting regimes, including at $M = H$, remains open.

After n offline expert trajectories, the worst-case expected suboptimality is $\Theta(\min\{1, A_H(M)/(n+1)\})$ for per-step mixing and $\Theta(\min\{1, \log B/(n+1)\})$ for per-episode sampling, with explicit constants. At $n = 0$, the exact values are $1 - B^{-H}$ and $1 - 1/B$. Coupling the learner to the expert up to the first deviation isolates the source of the deterministic gap: per-episode sampling deviates with probability $\frac{1}{|\mathcal{V}|} \sum_{\pi \in \mathcal{V}} \mathbf{1}\{\tau \in D_\pi\}$, which contributes an *indicator* for each surviving policy, whereas per-step mixing deviates with probability at most $\frac{1}{|\mathcal{V}|} \sum_{\pi \in \mathcal{V}} |\{t : \pi(s_t) \ne \pi^\star(s_t)\}|$, which contributes a *count* capped at H. The clipped-harmonic potential $\Phi_H(q) = \sum_{j \le q} \min\{1, H/j\}$ also telescopes under interactive full-trajectory annotation. It gives matching hybrid laws for the averaged iterate with $T = n + N_{\text{int}}$ episodes, and two-sided bounds for the final iterate.

The rare-context counterexample uses $H = m$, $B = m + 1$, and $N_{\text{off}} = m^2$, and satisfies every hypothesis of their Theorem 6. The algorithm has gap $\Theta(1/m) = \Theta(H/N_{\text{off}})$, with limit constant $e^{-1}(1 - e^{-e})$, while the conjectured bound is $O(\log(m/\delta)/m^2)$. For deterministic base classes, the worst-case expected offline term of WARM-STAGGER's first iterate is $\Theta(\min\{1, A_H(B-1)/N_{\text{off}}\})$. Their Theorem 6 additionally allows stochastic bases. Its single-state queries are different from the full-trajectory Bayesian protocol in Section 3; the stochastic version of WARM-STAGGER is consequently outside these results. Deploying a plurality vote attains the conjectured guarantee with $\log B$ in both terms, in expectation.

The per-episode upper bound follows by treating trajectories as examples and integrating the finite-class realizable consistency bound $\mathbb{P}(\text{err} > \varepsilon) \le \min\{1, M(1 - \varepsilon)^n\}$ over ε (Haussler et al., 1994; Shalev-Shwartz & Ben-David, 2014). This classical trajectory-level guarantee supplies the baseline against which the per-step cost is measured. Li and Zhang already distinguish first-step and each-step mixtures. Our analysis supplies the per-step law $A_H(M)$ and its matching lower bound, the rollout-specific potentials and hybrid laws, the refutation, and the iterated-logarithm law and tower potential for stochastic classes. The $\log^\star H$ dependence arises because fixed mixture weights determine which informative states are visited before failure.

Classical Bayesian sequence prediction (Hutter, 2003) bounds the cumulative expected predictive KL divergence by $\log(1/\text{prior mass of the truth})$. Predictions are evaluated on the observed sequence, with updates after each symbol. Cohen et al. (2022) study a conservative Bayesian imitator with selective expert queries in environments that need not reset. Here the weights stay fixed throughout an episode, and we measure Hellinger distance between the complete trajectory laws of the expert and the mixture. The learner can therefore prevent itself from reaching the states that would have provided further evidence. The entropy potential of Theorem 13 yields $O(H \log(M + 1))$; for two policies its initial budget is $\log(1/\text{prior}) = \log 2$. The tower potential captures the additional structure needed for the iterated-logarithm bound.

A separate line of work studies horizon-free behavior cloning under log loss (Foster et al., 2024a), the fundamental limits of imitation learning and the value of interaction (Rajaraman et al., 2020; Ross & Bagnell, 2010). For stochastic experts, the pointwise likelihood-weighted mixture used

in the noisy-expert distillation of Sriraman et al. (2026) is the analogue of per-step mixing (Section 4.1). Missing-mass estimation (Good, 1953) describes the mass of unobserved contexts in the deterministic constructions; Rajaraman et al. (2020) used it to analyse behavior cloning.

A.1 Quantitative summary

For reference, we collect the parameter dependence and constants of the results in one place. For Warm-Stagger, Li & Zhang (2025) conjectured that the $\log \widetilde{B}$ in their guarantee, where $B \le \widetilde{B} \le B^H$, can be replaced by $\log B$. The stochastic and deterministic analyses measure different losses, as detailed below.

For two stochastic policies containing the expert, with a uniform prior and untempered Bayesian updates, the worst-case cumulative trajectory Hellinger loss of frozen per-step mixing is $\Theta(\log^\star H)$ in the horizon H. The upper bound is $56(1+\log^\star H)$ for every class and environment, uniformly in the number of rounds. A two-state construction gives a lower bound of $\frac{1}{8}\log^\star H$. Failure prevents further informative visits in the episode. As a result, the accumulated visit count follows the approximate recursion $B_{t+1} \approx B_t + 2^{B_t}$. The upper bound combines a supermartingale for the square root of the posterior odds with a potential whose terms follow an exponential tower.

Under the same expert-label rule, sampling one policy per episode or updating after every label has worst-case cumulative loss $\Theta(1)$. Tempering the likelihood by α, as Nail does with $\alpha = 1/H$, gives $\Theta(1/\alpha + \log^\star H)$. This is linear in H at $\alpha = 1/H$. For $M+1$ stochastic policies, the two-sided bounds are tight when $M \ge H^2$, where the order agrees with that for deterministic classes, and when H is a much taller tower than M, giving $\Theta(M \log^\star H)$.

For deterministic classes and n offline expert trajectories, per-step mixing has worst-case expected suboptimality $\Theta(\min\{1, A_H(B-1)/(n+1)\})$, with $A_H(M) = \min\{H, M\}(1 + \log \frac{M}{\min\{H,M\}})$. Per-episode sampling has order $\Theta(\min\{1, \log B/(n+1)\})$. When offline demonstrations are followed by full-trajectory interactive labels, the averaged iterate obeys the corresponding matching laws. The explicit instance refuting the Warm-Stagger conjecture has gap $\Theta(H/N_{\text{off}})$. Plurality voting attains the conjectured rates in expectation under that algorithm's single-state query model.

B Stochastic constructions and proofs

We first construct the process that attains the two-policy lower bound, then prove the uniform upper bound and its extensions. Each proof uses the feedback protocol stated with its main result.

For deterministic policies with clean labels, the likelihood update eliminates inconsistent candidates and leaves equal weights on the rest. Pointwise Bayesian mixing then reduces exactly to the per-step version-space rule. Theorem 6, proved in Appendix E, applies to the Bernoulli instance with $n = 0$ offline episodes and gives cumulative loss $\Omega(\min\{T, A_H(M)\})$ over T rounds. On this construction, trajectory Hellinger loss lies between half the failure probability and the failure probability, so the return-gap lower bound transfers to this loss. For $M \ge e^H$ and $T \ge H \log M$, this becomes $\Omega(H \log M)$. Thus the horizon factor in the $H \log |\Pi|$ fast term for pointwise-mixture aggregation is necessary even within this deterministic subclass. The matching deterministic upper bound uses a potential Φ_H. The two-policy stochastic construction precludes an $O(A_H(M))$ upper bound for stochastic classes.

B.1 An action can end informative feedback

For two deterministic policies, $A_H(1) = 1$. For two stochastic policies, cumulative loss can instead grow without bound in H, at rate $\Theta(\log^\star H)$. The scale of the lower bound is seen through the exponential tower $T_0 = 1$, $T_{j+1} = e^{T_j}$, whose first relevant levels are $T_2 \approx 15$, $T_3 \approx 3.8 \cdot 10^6$. The instance has cumulative loss at least $\frac{7}{8}(j-1)$ whenever $H \ge T_j$, $j \ge 2$. This is a lower envelope and does not require S_H to be monotone in H.

Proposition 9 (No $O(A_H)$ bound for stochastic classes)**.** *There is a class of two stationary stochastic policies $\{\pi^\star, \pi^1\}$ on an MDP with two physical states (the time index adjoined) such that $\sum_{t=1}^{n} \mathbb{E}\,\mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \to n$ as $H \to \infty$ for every fixed n. Since $A_H(1) = 1$, no bound $C\,A_H(|\Pi| - 1)$ with an absolute constant C can hold.*

Proof. Consider two physical states: a is initial and b is absorbing. The actions are $\{0, 1, \dagger\}$. At a, actions $0, 1$ keep the process at a, while $\dagger$ moves it to b. Let $\pi^\star(\cdot \mid a) = \pi^\star(\cdot \mid b) = (\frac{1}{2}, \frac{1}{2}, 0)$, $\pi^1(\cdot \mid a) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, $\pi^1(\cdot \mid b) = (\frac{1}{2}, \frac{1}{2}, 0)$. Each of the H visited states receives an expert label (including the visit to a at which $\dagger$ is selected; the remaining steps are spent at b). Expert labels lie in $\{0, 1\}$. Each label at a has likelihood ratio $\frac{1}{2}$ under π^1 versus $\pi^\star$, whereas a label at b has ratio 1. Labels obtained after absorption therefore carry no further information, although the learner continues to act until the episode ends.

Let K_t be the number of visits to a in episode t and $B_t = \sum_{r<t} K_r$; then the posterior odds are $w_t(\pi^1)/w_t(\pi^\star) = 2^{-B_t}$, so $w_t(\pi^1) = 1/(1 + 2^{B_t})$ and the mixture plays $\dagger$ at each visit to a with probability $p(B_t) = \frac{1}{2} w_t(\pi^1) = \frac{1}{2(1+2^{B_t})}$. Hence $K_t = \min\{G_t, H\}$ with $G_t \mid B_t \sim \mathrm{Geom}(p(B_t))$ on $\{1, 2, \dots\}$.

The trajectory comparison is especially simple because the expert never leaves a. On those trajectories, every mixture action probability is the expert's probability times $1 - p(B_t)$. Therefore $dP^{\mu_t} = (1 - p(B_t))^H dP^{\pi^\star}$ on the expert's support, and $\mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) = g_H(B_t) := 1 - (1 - p(B_t))^{H/2}$.

We couple all horizons by generating the geometric variables from the same uniforms $U_1, U_2, \dots$. The untruncated process $B^\infty_{t+1} = B^\infty_t + G^\infty_t$ is almost surely finite at every fixed t. For $H \geq \max_{t<n} G^\infty_t$, the truncated process agrees with it up to episode n. Almost surely, for every $t \leq n$, the probability $p(B_t) = p(B^\infty_t) > 0$ is therefore eventually independent of H, and $g_H(B_t) \to 1$. Dominated convergence gives $\sum_{t \leq n} \mathbb{E} g_H(B_t) \to n$. □

Lemma 10 (Lower bound)**.** *In the instance of Proposition 9,* $S_H := \sum_{t \geq 1} \mathbb{E}\, g_H(B_t)$ *satisfies* $S_H \geq \max\{1 - \frac{\sqrt{3}}{2},\ \frac{7}{8}(\log^\star H - 2)_+\}$*, and also* $S_H \geq \frac{1}{8} \log^\star H$*.*

Proof. Let $T_0 = 1$, $T_{j+1} = e^{T_j}$, and consider the untruncated process $B^\infty_1 = 0$, $B^\infty_{t+1} = B^\infty_t + G_t$, $G_t \mid B^\infty_t = b \sim \mathrm{Geom}(p(b))$. Coupling through common uniforms in the inverse geometric distribution gives $B^H_t \leq B^\infty_t$ for every t, because $p(b)$ is decreasing in b. It suffices first to control the untruncated count by the tower: $\mathbb{P}\{B^\infty_t \leq T_t\ \forall t \geq 2\} \geq \frac{31}{32}$. At the first level, $\mathbb{P}(B^\infty_2 > 15) = (\frac{3}{4})^{15} < \frac{1}{64}$ and $15 < T_2$. Next, if $b \leq T := T_t$ with $t \geq 2$ then $p(b) \geq 2^{-T-2}$, and since $T \geq 15$, $\lfloor e^T - b \rfloor \geq e^T - T - 1 \geq \frac{1}{2} e^T$, so the geometric tail gives $\mathbb{P}(b + G_t > T_{t+1} \mid B^\infty_t = b) \leq (1 - 2^{-T-2})^{e^T/2} \leq \exp(-e^{(1-\log 2)T}/8) \leq \exp(-e^{T/4}/8)$. As $T_t \geq 15 + 8(t-2)$ and $e^{15/4} > 40$, this probability is at most $e^{-5(t-1)}$. Summing the first-exit probabilities gives $\mathbb{P}\{\exists t \geq 2 : B^\infty_t > T_t\} < \frac{1}{64} + \sum_{j \geq 1} e^{-5j} < \frac{1}{32}$.

On the event that the count remains below the tower, the odds stay large enough to incur constant loss at successive tower levels. Let $k = \log^\star H \geq 3$, so that $H \geq T_{k-1} \geq T_2 > 15$. In the first round $g_H(0) = 1 - (\frac{3}{4})^{H/2} \geq 1 - (\frac{3}{4})^8 > \frac{7}{8}$. For $2 \leq t \leq k-2$, on the event above, $\frac{H}{2} p(B^H_t) \geq H 2^{-T_t - 3} \geq T_{t+1} 2^{-T_t - 3} = e^{(1-\log 2)T_t}/8 > 5$, so $g_H(B^H_t) \geq 1 - e^{-5}$ and $\mathbb{E} g_H(B^H_t) \geq \frac{31}{32}(1 - e^{-5}) > \frac{7}{8}$. Summing the $k-2$ rounds $t = 1, \dots, k-2$ gives $S_H \geq \frac{7}{8}(k-2)$. For every $H \geq 1$, $S_H \geq g_H(0) \geq 1 - (\frac{3}{4})^{1/2} = 1 - \frac{\sqrt{3}}{2}$. The uniform form $S_H \geq \frac{k}{8}$ follows by also checking the smaller tower heights: for $k = 1$, $1 - \frac{\sqrt{3}}{2} > \frac{1}{8}$; for $k = 2$, $H \geq 3$ and $g_H(0) \geq 1 - (\frac{3}{4})^{3/2} > \frac{1}{4}$; for $k \geq 3$, $\frac{7}{8}(k-2) \geq \frac{k}{8}$. □

B.2 A POTENTIAL FOR INFORMATION COLLECTED ON THE ROLLOUT

The upper bound must use labels where the learner actually travels. Information available only on the expert's trajectory cannot justify its posterior update. Fix a two-policy class $\{\pi^\star, \pi^1\}$. Write $r_t = w_t(\pi^1)/w_t(\pi^\star)$ for the posterior odds, with $r_1 = 1$ and $w_t := w_t(\pi^1) = r_t/(1 + r_t)$. At a state s, the policies' disagreement is measured by $d(s) = \mathsf{h}^2(\pi^\star(\cdot \mid s), \pi^1(\cdot \mid s)) \in [0, 1]$. A round multiplies the odds by its label likelihood ratio: $r_{t+1} = r_t \Lambda_t$, where $\Lambda_t = \prod_{h=1}^H \pi^1(a'_h \mid s_h)/\pi^\star(a'_h \mid s_h)$.

We first relate trajectory loss to information collected at visited states. A related sequential Hellinger comparison is given by Foster et al. (2024b, Lem. F.2); related forms appear in Foster et al. (2024a) and the augmented-Hellinger comparison of Sriraman et al. (2026). Here we prove a comparison

clipped at $\min\{1, wD\}$ using a stopped likelihood-ratio process and a predictable budget. This comparison supplies the information bound. The tower potential then converts it into a loss bound with iterated-logarithm dependence.

Lemma 11 (Trajectory loss from information on the learner's states)**.** *Let $0 \le w < 1$, $\mu_w = (1-w)\pi^\star + w\pi^1$ (pointwise), $P = P^{\pi^\star}$, $Q = P^{\mu_w}$ (same initial distribution and transitions), and for a trajectory $(s_1, \dots, s_H) \sim Q$ let $D = \sum_{h=1}^{H} d(s_h)$. Then $\mathsf{h}^2(P, Q) \le 22\,\mathbb{E}_Q[\min\{1, wD\}]$.*

Proof. For $w = 0$, both trajectory laws agree. Assume $0 < w < 1$, so $P \ll Q$. Let $e_h = \mathsf{h}^2(\pi^\star(\cdot \mid s_h), \mu_w(\cdot \mid s_h))$; concavity of the square root gives $e_h \le w\, d(s_h)$. Let $\mathcal{F}_h$ be generated by the first h actions and resulting states ($\mathcal{F}_0$ by the initial state) and $Z_h = \sqrt{dP|_{\mathcal{F}_h}/dQ|_{\mathcal{F}_h}}$, $Z_0 = 1$. Since the transition kernels coincide, $\mathbb{E}_Q[Z_h \mid \mathcal{F}_{h-1}] = Z_{h-1}(1 - e_h)$, so for any stopping time $\tau \le H$, $\ell_\tau := 1 - \mathbb{E}_Q Z_\tau = \mathbb{E}_Q \sum_{h \le \tau} Z_{h-1} e_h$ and $\mathbb{E}_Q(Z_\tau - 1)^2 = \mathbb{E}_Q Z_\tau^2 - 2\mathbb{E}_Q Z_\tau + 1 = 2\ell_\tau$ (as $\mathbb{E}_Q Z_\tau^2 = \mathbb{E}_Q[dP/dQ|_{\mathcal{F}_\tau}] = 1$).

To control the likelihood-ratio factor, stop when it leaves a fixed interval or the accumulated information reaches unit scale. Let $A_h = w \sum_{i \le h} d(s_i)$ and $\tau = \min(\{h : Z_h \notin (\frac{1}{2}, 2) \text{ or } A_h \ge 1\} \cup \{H\})$. Before stopping, $Z_{h-1} < 2$ and $A_{h-1} < 1$, each increment of A is at most $w \le 1$, so pathwise $A_\tau \le \min\{A_H, 2\} \le 2\min\{A_H, 1\}$; with $a := \mathbb{E}_Q \min\{A_H, 1\}$, this gives $\ell_\tau \le 2\mathbb{E}_Q \sum_{h \le \tau} e_h \le 2\mathbb{E}_Q A_\tau \le 4a$.

The remaining loss can be charged to the probability of stopping. Partition $\{\tau < H\}$ into $R = \{\tau < H, Z_\tau \notin (\frac{1}{2}, 2)\}$ and $C = \{\tau < H, Z_\tau \in (\frac{1}{2}, 2)\}$; on C, $A_\tau \ge 1$ hence $A_H \ge 1$, so $Q(C) \le a$. Since $z \le 2(z-1)^2$ for $z \in [0, \frac{1}{2}] \cup [2, \infty)$, $\mathbb{E}_Q[Z_\tau \mathbf{1}\{R\}] \le 2\mathbb{E}_Q(Z_\tau - 1)^2 = 4\ell_\tau$, and $\mathbb{E}_Q[Z_\tau \mathbf{1}\{C\}] \le 2Q(C) \le 2a$. By nonnegativity of Z, $\mathbb{E}_Q Z_H \ge \mathbb{E}_Q[Z_\tau \mathbf{1}\{\tau = H\}] = \mathbb{E}_Q Z_\tau - \mathbb{E}_Q[Z_\tau \mathbf{1}\{\tau < H\}]$, and $\mathsf{h}^2(P, Q) = 1 - \mathbb{E}_Q Z_H \le \ell_\tau + \mathbb{E}_Q[Z_\tau \mathbf{1}\{\tau < H\}] \le 5\ell_\tau + 2a \le 22a$. □

Stopping only when the square-root likelihood ratio leaves the same interval also gives an unclipped comparison: the stopped loss is at most twice the expected sum of one-step losses, and the stopping remainder is at most four times the stopped loss. Thus the full trajectory loss is at most ten times that sum, as used in the within-episode argument below.

The proof of Lemma 11 needs only $e_h \le a_h \le 1$, where a_h is measurable before action h. Thus any such predictable budget gives $\mathsf{h}^2(P, Q) \le 22\,\mathbb{E}_Q \min\{1, \sum_h a_h\}$. Section B.3 uses this form for more than two policies.

We next construct the potential that pays for this clipped information budget. Its terms are capped square roots at successive scales of the odds. Spacing those scales along an exponential tower makes the initial potential grow only with the number of tower levels.

Lemma 12 (Tower potential)**.** *Let $x_0 = 0$, $x_{j+1} = e^{x_j}$, $m = \min\{j \ge 0 : x_j \ge \log H\}$, and*

$$F_H(r) = \sqrt{r} + \sum_{j=1}^{m} \min\bigl\{1, \sqrt{r e^{x_j}}\bigr\}, \qquad r \ge 0.$$

Then $m \le \log^\star H$ and $F_H(1) = m + 1$. If $0 \le D \le H$ and $\Lambda \ge 0$ is a random variable with $\mathbb{E}\Lambda \le 1$ and $\mathbb{E}\sqrt{\Lambda} \le e^{-D}$, then for every $r \ge 0$, with $w = r/(1+r)$,

$$F_H(r) - \mathbb{E}F_H(r\Lambda) \;\ge\; c_0 \min\{1, wD\}, \qquad c_0 = 1 - e^{-1/2}.$$

Proof. First, $F_H(1) = 1 + m$ because $e^{x_j} \ge 1$. With $T_0 = 1$, $T_{j+1} = e^{T_j}$, we have $x_j = T_{j-1}$ for $j \ge 1$. The definition $\log^\star H = k$ means $T_{k-1} \le H < T_k$, so $m = \min\{j : T_j \ge H\} \le k$.

Every summand of F_H is increasing and concave. Jensen's inequality and $\mathbb{E}\Lambda \le 1$ therefore give a nonnegative expected decrease for each one. It is therefore enough to select a summand whose scale matches the current odds and information. For $f_a(r) = \min\{1, \sqrt{r/a}\}$, if $r \le a$, then $f_a(r) - \mathbb{E}f_a(r\Lambda) \ge \sqrt{r/a}(1 - \mathbb{E}\sqrt{\Lambda}) \ge \sqrt{r/a}(1 - e^{-D})$. If $r \ge a$, then $f_a(r) = 1$ and $\mathbb{E}f_a(r\Lambda) \le \sqrt{r/a}\,\mathbb{E}\sqrt{\Lambda}$, so the decrease is at least $1 - \sqrt{r/a}\,e^{-D}$. The uncapped summand decreases by at least $\sqrt{r}(1 - e^{-D})$. Note $1 - e^{-D} \ge (1 - e^{-1})\min\{1, D\}$ and $1 - e^{-1} > c_0$.

Case $r \ge 1$. $\sqrt{r}(1 - e^{-D}) \ge (1 - e^{-1})\min\{1, D\} \ge c_0 \min\{1, wD\}$ since $w \le 1$.

Case $0 < r < 1$, $x := -\log r$. If $x \geq x_m$, consider the last summand ($a = e^{-x_m}$; the uncapped one if $m = 0$), initially $q = \sqrt{re^{x_m}}$; as $e^{x_m} \geq H$ and $r \leq e^{-x_m}$, $e^{x_m}/r \geq e^{2x_m} \geq H^2$, so $q \geq rH$, and the decrease is at least $q(1-e^{-D}) \geq rH(1-e^{-1})\min\{1, D\} \geq (1-e^{-1})rD$ (using $D \leq H$), which is $\geq c_0 \min\{1, wD\}$ because $r \geq w$. Otherwise pick $i < m$ with $x_i \leq x \leq x_{i+1}$ and consider the i-th summand (uncapped if $i = 0$), initially $q_i = \sqrt{re^{x_i}}$; then $q_i \geq \sqrt{r} \geq r$ and $q_i \geq re^{x_i} = rx_{i+1}$ (as $re^{x_i} \leq 1$). If $D \leq x_{i+1}$ the decrease is at least $q_i(1-e^{-D}) \geq (1-e^{-1})rD$ (for $D \leq 1$ use $q_i \geq r$; for $D \geq 1$ use $q_i \geq rx_{i+1} \geq rD$), again $\geq c_0 \min\{1, wD\}$. If $D \geq x_{i+1}$, the $(i+1)$-st summand has $a = e^{-x_{i+1}} \leq r$, is initially 1, and decreases by at least $1 - \sqrt{r/a}\, e^{-D} = 1 - \exp(\frac{x_{i+1}-x}{2} - D) \geq 1 - e^{-D/2} \geq 1 - e^{-1/2} = c_0$, since $x \geq x_i \geq 0$ gives $\frac{x_{i+1}-x}{2} \leq \frac{x_{i+1}}{2} \leq \frac{D}{2}$ and $D \geq x_{i+1} \geq x_1 = 1$. □

Proof of the upper bound in Theorem 1. We now combine the two lemmas. First fix the learner's visited states so that the label likelihood can be analyzed conditionally. Condition on the past and the entire learner trajectory $(s_1, \ldots, s_H)$ in round t. Because the deployed policy is frozen during the episode, the labels remain conditionally independent draws $a'_h \sim \pi^\star(\cdot \mid s_h)$. Hence $\mathbb{E}[\Lambda_t \mid \text{past}, \text{traj}] = \prod_h \sum_{a:\pi^\star(a|s_h)>0} \pi^1(a \mid s_h) \leq 1$ and $\mathbb{E}[\sqrt{\Lambda_t} \mid \text{past}, \text{traj}] = \prod_h \sum_a \sqrt{\pi^\star(a \mid s_h)\pi^1(a \mid s_h)} = \prod_h (1 - d(s_h)) \leq e^{-D_t}$ with $D_t = \sum_h d(s_h) \leq H$.

Apply Lemma 12 conditionally with $r = r_t$, $\Lambda = \Lambda_t$, then average over the trajectory $\sim P^{\mu_t}$. This gives $F_H(r_t) - \mathbb{E}[F_H(r_{t+1}) \mid \text{past}] \geq c_0\, \mathbb{E}_{P^{\mu_t}} \min\{1, w_t D_t\}$, and Lemma 11 gives $\mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \leq 22\, \mathbb{E}_{P^{\mu_t}} \min\{1, w_t D_t\}$. Therefore $\mathbb{E}\, \mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \leq \frac{22}{c_0}\big(\mathbb{E}F_H(r_t) - \mathbb{E}F_H(r_{t+1})\big)$; all expectations are finite because $\mathbb{E}r_t \leq 1$ and $0 \leq F_H(r) \leq \sqrt{r} + m$. Summing these potential differences over $t \leq n$ telescopes. Using $F_H \geq 0$, we obtain $\sum_{t \leq n} \mathbb{E}\, \mathsf{h}^2 \leq \frac{22}{c_0} F_H(1) = \frac{22}{c_0}(m+1) < 56(1 + \log^\star H)$, as $22/(1 - e^{-1/2}) < 56$. □

B.3 Classes of M stochastic policies

For larger classes, two budgets become relevant: a tower potential for each non-expert policy and an entropy budget for the expert's posterior mass. Keep the same protocol and take $\Pi = \{\pi^\star, \pi^1, \ldots, \pi^M\}$. Define $\mathcal{C}_M(H) = \sup \sup_n \sum_{t \leq n} \mathbb{E}\, \mathsf{h}^2(P^{\pi^\star}, P^{\mu_t})$, where the supremum ranges over MDPs and stochastic policy classes. Write $k = \log^\star H$, $s = \log^\star(16(M+1))$, and recall $A_H(M) = \min\{H, M\}(1 + \log \frac{M}{\min\{H,M\}})$.

Theorem 13 (Multipolicy bounds)**.** *For all integers $M, H \geq 1$,*

$$\max\Big\{\frac{A_H(M)}{16},\ \frac{k}{8},\ \frac{7}{8}M(k-s-1)_+\Big\} \ \leq\ \mathcal{C}_M(H) \ \leq\ \min\big\{56M(1+k),\ 11H\log(M+1)\big\}.$$

Consequently $\mathcal{C}_M(H) \geq \frac{1}{32}(A_H(M) + \log^\star H)$; if $M \geq H^2$ then $\frac{1}{16}A_H(M) \leq \mathcal{C}_M(H) \leq 22A_H(M)$; if $\log^\star H \geq 2(\log^\star(16(M+1))+1)$ then $\frac{7}{16}M\log^\star H \leq \mathcal{C}_M(H) \leq 112M\log^\star H$; and along $M = 2^H$, $\mathcal{C}_M(H) = \Theta(H^2) = \Theta(A_H(M))$, so none of the three candidates $M(1+\log^\star H)$, $A_H(M) + M\log^\star H$, $A_H(M)(1+\log^\star H)$ is the uniform order.

Each budget is tight in a different parameter regime. For $M \geq H^2$, they have the same order as for deterministic classes, with no additional $\log^\star$ factor. When the tower height of H exceeds twice that of M in the sense stated in the theorem, each wrong policy can contribute its own $\log^\star H$ term. The uniform order between these regimes remains open. For instance, at $M = H$ the bounds are $H/16 \leq \mathcal{C}_H(H) \leq 56H(1 + \log^\star H)$.

Proof of the upper bounds. Let $r_t(i) = w_t(\pi^i)/w_t(\pi^\star)$, $R_t = \sum_i r_t(i)$, $w_t(i) = r_t(i)/(1+R_t)$, and $r_{t+1}(i) = r_t(i)\Lambda_t(i)$ with $\Lambda_t(i) = \prod_h \pi^i(a'_h \mid s_h)/\pi^\star(a'_h \mid s_h)$. Conditionally on the past and the learner trajectory, $\mathbb{E}\Lambda_t(i) \leq 1$ and $\mathbb{E}\sqrt{\Lambda_t(i)} = \prod_h (1 - d_i(s_h)) \leq e^{-D_t(i)}$ with $D_t(i) = \sum_h d_i(s_h)$, for each i separately. These conditional identities apply separately for each i, without an independence assumption.

Sum of towers. To apply the two-policy argument coordinatewise, sum the tower potentials: $\Phi_t = \sum_i F_H(r_t(i))$. Lemma 12 gives $\Phi_t - \mathbb{E}[\Phi_{t+1} \mid \text{past}] \geq c_0 \mathbb{E}_{P^{\mu_t}} \sum_i \min\{1, w'_t(i) D_t(i)\}$ with $w'_t(i) = r_t(i)/(1 + r_t(i)) \geq w_t(i)$; since $\sum_i \min\{1, a_i\} \geq \min\{1, \sum_i a_i\}$ for $a_i \geq 0$, and since

by concavity $e_h \le \sum_i w_t(i) d_i(s_h) =: a_h \le 1$ is a predictable budget, the predictable-budget form of Lemma 11 gives $\mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \le 22\mathbb{E}_{P^{\mu_t}} \min\{1, \sum_i w_t(i) D_t(i)\} \le \frac{22}{c_0}(\Phi_t - \mathbb{E}[\Phi_{t+1} \mid \text{past}])$. Telescoping from $\Phi_1 = M(m+1) \le M(1+k)$ gives the first bound.

Entropy. For the second bound, track the posterior mass of the expert instead of individual odds. Let $V_t = \log(1 + R_t) = -\log w_t(\pi^\star)$. Fix the learner's states $s = (s_1, \dots, s_H)$ and let $L_{\star,s} = \bigotimes_h \pi^\star(\cdot \mid s_h)$, $L_{i,s} = \bigotimes_h \pi^i(\cdot \mid s_h)$, $\bar{L}_s = w_t(\pi^\star) L_{\star,s} + \sum_i w_t(i) L_{i,s}$ (laws of the label vector). The posterior identity gives $V_t - \mathbb{E}[V_{t+1} \mid \text{past}, s] = \mathbb{E}_{L_{\star,s}} \log(L_{\star,s}/\bar{L}_s) = \mathrm{KL}(L_{\star,s} \| \bar{L}_s) =: K_t(s)$. The h-th marginal of $\bar{L}_s$ is $\mu_t(\cdot \mid s_h)$, so by data processing and $\mathrm{KL} \ge 2\mathsf{h}^2$, $K_t(s) \ge \mathrm{KL}(\pi^\star(\cdot \mid s_h) \| \mu_t(\cdot \mid s_h)) \ge 2e_h$ for every h, whence $\sum_h e_h \le \frac{H}{2} K_t(s)$. Lemma 11 with $a_h = e_h$ gives $\mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \le 22\mathbb{E}_{P^{\mu_t}} \sum_h e_h \le 11H(V_t - \mathbb{E}[V_{t+1} \mid \text{past}])$, and telescoping from $V_1 = \log(M+1)$ with $V_t \ge 0$ gives the second bound. □

Proof of the lower bounds. Deterministic lower bound $A_H(M)/16$. Draw a context J uniformly from $[M]$ in each episode. The expert plays 0 everywhere. Policy i plays the catastrophic action 1 exactly in context i, leading to an absorbing state where all policies agree. The first visit to an unresolved context eliminates its policy: the visited state receives a label even if the learner plays 1 there. With q unresolved contexts, the conditional loss in an unresolved context is $g(q) = 1 - (\frac{q}{q+1})^{H/2}$, the per-round loss is $qg(q)/M$, and the mean number of rounds at level q is M/q, so the infinite-round cumulative loss is exactly $D_M(H) = \sum_{q=1}^{M} g(q)$. Since $g(q) \ge 1 - e^{-H/(2(q+1))} \ge \frac{1}{2}\min\{1, \frac{H}{2(q+1)}\} \ge \frac{1}{8}\min\{1, H/q\}$ and $\sum_{q=1}^{M} \min\{1, H/q\} \ge \frac{1}{2}\int_0^M \min\{1, H/x\} dx = \frac{1}{2} A_H(M)$, we get $D_M(H) \ge A_H(M)/16$.

Two-policy embedding $k/8$. Add to the two-policy instance an ordinary context of probability $1 - \varepsilon$ in which expert and gadget policy agree, and $M - 1$ distractor policies assigning probability zero to every expert label at the first ordinary-context state (made pairwise distinct, and distinct from the gadget policy, on unreachable states). On the event that the first episode is ordinary (probability $1 - \varepsilon$), it eliminates the distractors and leaves expert and gadget policy with equal weight; conditionally on this event, ordinary episodes have likelihood ratio 1 and zero loss, and the gadget episodes reproduce the two-policy process in their own visit index, so $\mathcal{C}_M(H) \ge (1 - \varepsilon) S_H \ge (1 - \varepsilon) k/8$ by Lemma 10. Letting $\varepsilon \downarrow 0$ gives the bound.

M copies $\frac{7}{8} M(k - s - 1)_+$. Draw J uniformly from $[M]$; in context j use the gadget of Proposition 9 with $\pi^\star = (\frac{1}{2}, \frac{1}{2}, 0)$, $\pi^j = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$ at alive states, and every π^i, $i \ne j$, agreeing with the expert; all agree in the absorbing state. Let B_j be the cumulative number of alive-state labels obtained in context j; then $r(j) = 2^{-B_j}$ and the catastrophe probability per alive step in context j is $p_j = \frac{2^{-B_j}}{2(1 + \sum_i 2^{-B_i})} \ge \frac{2^{-B_j}}{2(M+1)}$, with conditional loss $1 - (1 - p_j)^{H/2}$.

A separate geometric comparison for each context bounds how quickly its evidence can grow. Index the episodes of context j by ℓ. By inverse-geometric coupling, the value of B_j before the ℓ-th visit is at most $\bar{B}_\ell$, where $\bar{B}_1 = 0$, $\bar{B}_{\ell+1} = \bar{B}_\ell + \bar{G}_\ell$, $\bar{G}_\ell \mid \bar{B}_\ell = b \sim \mathrm{Geom}(2^{-b}/(2u))$, $u = M + 1$. Put $t_\ell = T_{s+\ell-1}$, so $t_1 > 16u \ge 32$. If $b \le t = t_\ell$, the geometric tail gives $\mathbb{P}(b + \bar{G}_\ell > e^t \mid b) \le \exp(-e^{(1-\log 2)t}/(4u)) \le \exp(-4e^{t/4}/t)$; for $t \ge 32$ the function $f(t) = 4e^{t/4}/t$ is increasing with $f(32) > 5$ and $f(t+8) \ge 2f(t)$, and $t_{\ell+1} \ge t_\ell + 8$, so the failure probability at level ℓ is at most $e^{-5\ell}$ and $\mathbb{P}\{\bar{B}_\ell \le t_\ell \ \forall \ell\} > \frac{127}{128}$.

The tower event leaves enough posterior weight for each context to contribute constant loss on its early visits. Let $n = (k - s - 1)_+$; for $1 \le \ell \le n$, $H \ge T_{k-1} \ge T_{s+\ell} = e^{t_\ell}$, and on the good event $\frac{H}{2} p_j \ge H 2^{-t_\ell}/(4u) \ge e^{(1-\log 2)t_\ell}/(4u) > 5$, so each of the first n visits to context j has expected conditional loss at least $\frac{127}{128}(1 - e^{-5}) > \frac{7}{8}$. The conditional expectation of the contextual loss equals the trajectory Hellinger loss of the round, every context is visited infinitely often, and Tonelli's theorem sums the losses over contexts and visit indices: $\mathcal{C}_M(H) \ge \frac{7}{8} Mn$.

Consequences. Combining the first two lower bounds uses $\max\{A/16, k/8\} \ge (A + k)/32$. If $M \ge H^2$ then $\log M \le 2\log(M/H)$, so $H \log(M+1) \le 2H(1 + \log \frac{M}{H}) = 2A_H(M)$. If $k \ge 2(s+1)$ then $k - s - 1 \ge k/2$ and $1 + k \le 2k$. For $M = 2^H$, $H \ge 4$: $M \ge H^2$ and $A_H(2^H) = H(1 + H\log 2 - \log H) = \Theta(H^2)$. □

B.4 PROOF OF THE TEMPERED LAW

Proof of the upper bound. The upper bound follows by rescaling the tower potential to reflect the smaller change in odds. By Jensen's inequality, using the concavity of $x \mapsto x^\alpha$, conditionally on the past and the learner trajectory, $\mathbb{E}\Lambda_t^\alpha \le (\mathbb{E}\Lambda_t)^\alpha \le 1$ and $\mathbb{E}\Lambda_t^{\alpha/2} \le (\mathbb{E}\sqrt{\Lambda_t})^\alpha \le e^{-\alpha D_t}$. Define the *scaled* tower potential $\Psi_{\alpha,H}(r) = F_H(r/\alpha^2) = \frac{\sqrt{r}}{\alpha} + \sum_{j=1}^m \min\{1, \sqrt{r e^{x_j}}/\alpha\}$, so that $\Psi_{\alpha,H}(1) = \alpha^{-1} + m \le \alpha^{-1} + k$. It is enough to show that for $L \ge 0$ with $\mathbb{E}L \le 1$, $\mathbb{E}\sqrt{L} \le e^{-\alpha D}$ and $0 \le D \le H$, $\Psi_{\alpha,H}(r) - \mathbb{E}\Psi_{\alpha,H}(rL) \ge c_0 \min\{1, \frac{r}{1+r}D\}$. As before, each summand has a nonnegative expected decrease; the choice of summand depends on the odds.

For $r \le \alpha^2$, put $u = r/\alpha^2$. Lemma 12 (with $\alpha D \le H$ in place of D) gives a decrease of at least $c_0 \min\{1, \frac{u}{1+u}\alpha D\}$, and $\frac{\alpha u}{1+u} = \frac{\alpha r}{\alpha^2+r} \ge \frac{r}{1+r}$ because $\alpha(1+r) - (\alpha^2+r) = (1-\alpha)(\alpha - r) \ge 0$. For $r \ge \alpha^2$, the uncapped summand alone decreases by at least $\frac{\sqrt{r}}{\alpha}(1 - e^{-\alpha D})$: for $\alpha D \le 1$ this is $\ge \beta\sqrt{r}\, D \ge \beta \frac{r}{1+r} D$ with $\beta = 1 - e^{-1} > c_0$, and for $\alpha D \ge 1$ it is $\ge \beta\sqrt{r}/\alpha \ge \beta$.

Apply this decrease bound conditionally with $L = \Lambda_t^\alpha$ and average over the trajectory. Lemma 11 then gives $\mathsf{h}^2(P^{\pi^\star}, P^{\mu_t}) \le \frac{22}{c_0}(\Psi(r_t) - \mathbb{E}[\Psi(r_{t+1}) \mid \text{past}])$, and telescoping gives $\sum_{t \le n} \mathbb{E}\,\mathsf{h}^2 \le \frac{22}{c_0}(\alpha^{-1} + m) < 56(\alpha^{-1} + k)$. □

Proof of the lower bounds. In the instance of Proposition 9 the tempered odds are $r_t = 2^{-\alpha B_t}$, so $K_t \mid B_t = b \sim \min\{\mathrm{Geom}(p_\alpha(b)), H\}$ with $p_\alpha(b) = \frac{1}{2(1+2^{\alpha b})}$ and conditional loss $g_{\alpha,H}(b) = 1 - (1 - p_\alpha(b))^{H/2}$.

The $1/(4\alpha)$ bound. For the first lower bound, integrate the catastrophe probability. The resulting potential cannot decrease faster than a constant multiple of the incurred loss. Extend p_α to real $b \ge 0$ and let $V(b) = \int_b^\infty p_\alpha(x)dx = \log(1 + 2^{-\alpha b})/(2\alpha \log 2)$, so $V(0) = 1/(2\alpha)$. Since p_α is decreasing, $\mathbb{E}[V(b) - V(b+K) \mid b] \le p_\alpha(b)\mathbb{E}[K \mid b] = 1 - (1 - p_\alpha(b))^H = g_{\alpha,H}(b)(1 + (1 - p_\alpha(b))^{H/2}) \le 2g_{\alpha,H}(b)$. As $B_t \ge t-1$, $V(B_t) \to 0$ boundedly, and telescoping gives $\frac{1}{2\alpha} \le 2S_{\alpha,H}$.

The tower lower bound. Couple the tempered and untempered processes through common inverse-distribution uniforms. Then $B_t^\alpha \le B_t^1$ inductively, since $\alpha B_t^\alpha \le B_t^1$ implies $p_\alpha(B_t^\alpha) \ge p_1(B_t^1)$ and hence $K_t^\alpha \le K_t^1$. Thus $g_{\alpha,H}(B_t^\alpha) \ge g_{1,H}(B_t^1)$ pathwise, and $S_{\alpha,H} \ge S_{1,H} \ge k/8$ by Lemma 10. Hence $S_{\alpha,H} \ge \max\{\frac{1}{4\alpha}, \frac{k}{8}\} \ge \frac{1}{12}(\alpha^{-1} + k)$.

The asymptotic at $\alpha = 1/H$. To obtain the limiting constant, we refine the integral potential and bound the error of approximating its decrease by a first-order term. Put $a = \log 2/H$, $h = H/2$, $p(b) = \frac{1}{2(1+e^{ab})}$, $q = 1 - p$, $g(b) = 1 - q(b)^h$, $f(b) = p(b)/(1 + q(b)^h)$ and $I(b) = \int_b^\infty f$. For $K = \min\{\mathrm{Geom}(p(b)), H\}$, $f(b)\mathbb{E}[K \mid b] = g(b)$, and f decreases, so $\mathbb{E}[I(b) - I(b+K) \mid b] \le g(b)$ and telescoping gives $S_{1/H,H} \ge I(0)$. Substituting p as the variable ($p'(b) = -ap(1-2p)$), $I(0) = \frac{1}{a}\int_0^{1/4} \frac{dp}{(1-2p)(1+(1-p)^h)} = \frac{H}{2} - J$ with $0 \le J = \frac{1}{a}\int_0^{1/4} \frac{(1-p)^h dp}{(1-2p)(1+(1-p)^h)} \le \frac{2}{a}\int_0^{1/4} e^{-hp}dp \le \frac{2}{ah} = \frac{4}{\log 2}$. For the upper bound, we control the error in replacing a potential difference by its first-order term. The inequality $hp(1-p)^{h-1} \le 1$ gives $0 \le \frac{d}{dp}\frac{p}{1+(1-p)^h} \le 2$, hence $0 \le -f'(b) \le 2ap(b)$ and $0 \le kf(b) - [I(b) - I(b+k)] \le ap(b)k^2$ for $k \ge 0$; with $\mathbb{E}K^2 \le \frac{2}{p}\mathbb{E}K$ for the truncated geometric variable, $g(b) \le \mathbb{E}[I(b) - I(b+K) \mid b] + 2a\mathbb{E}[K \mid b]$. Let $b_0 = \log H/a$ and $\tau = \inf\{t : B_t \ge b_0\}$; as $1 \le K_t \le H$, $B_\tau < b_0 + H$, and summing over $t < \tau$, $\mathbb{E}\sum_{t<\tau} g(B_t) \le I(0) + 2a\mathbb{E}B_\tau \le \frac{H}{2} + 2\log H + 2\log 2$.

After this stopping time, the posterior odds themselves bound the remaining loss. For $t \ge \tau$, with $r = e^{-ab}$ and $r' = re^{-aK}$: concavity on $[0, H]$ gives $1 - e^{-aK} \ge \frac{K}{H}(1 - e^{-aH}) = \frac{K}{2H}$, so $\mathbb{E}[r - r' \mid b] \ge \frac{r}{2H}\mathbb{E}[K \mid b]$, while $g(b) \le p(b)\mathbb{E}[K \mid b] \le \frac{r}{2}\mathbb{E}[K \mid b]$; thus $g(b) \le H\,\mathbb{E}[r - r' \mid b]$ and telescoping from τ (where $r_\tau \le 1/H$) gives $\mathbb{E}\sum_{t \ge \tau} g(B_t) \le 1$. □

The count recursion also permits a finite-horizon comparison of the exponents. Directed-rounding evaluation, as in Proposition 9 but with p_α and truncation error $\le Hz^{N+1}/(2(1-z))$, $z = 2^{-\alpha}$, gives $S_{\alpha,10^3} = 2.2653,\ 3.0712,\ 6.8237,\ 35.998$ and $S_{\alpha,10^6} = 2.6340,\ 3.4863,\ 7.3998,\ 37.117$ for $\alpha = 1, \frac{1}{2}, \frac{1}{8}, \frac{1}{64}$, against $\alpha^{-1} + \log^\star H = 4, 5, 11, 67$; and for $\alpha = 1/H$ the exact recursion

gives $S_{1/H,10^2} = 50.80$ and $S_{1/H,10^3} = 503.09$, inside the analytic envelopes $[44.2, 61.6]$ and $[494.2, 516.2]$ of Theorem 3.

B.5 Proof for per-episode and per-label deployment

Proof. Lower bound. Let $\pi^\star$ be uniform on actions $\{0, 1\}$ and π^1 uniform on $\{2, 3\}$ at the initial state. All actions lead to a common absorbing state where the policies agree. The two trajectory laws have disjoint supports. Under either protocol, the first deployed law is their equal mixture, with $\mathsf{h}^2 = 1 - 1/\sqrt{2}$. The first expert label then eliminates π^1.

Per-episode sampling. Let $\Delta = \mathsf{h}^2(P^{\pi^\star}, P^{\pi^1})$ and, for $i \in \{\star, 1\}$, $I_i = \mathbb{E}_{P^{\pi^i}}[1 - \prod_h (1 - d(s_h))]$. Apply the predictable-budget form of Lemma 11 with reference law P^{π^i}, the other policy's trajectory law as target, and $a_h = d(s_h)$. The conditional Hellinger distance of the step kernels is exactly $d(s_h)$ because the transitions coincide. Absolute continuity of the target law with respect to the reference is not required: the proof of Lemma 11 uses only that L_h, the likelihood-ratio process, is a nonnegative supermartingale with $\mathbb{E}_Q Z_\tau^2 \le 1$. We obtain $\Delta \le 22\, \mathbb{E}_{P^{\pi^i}} \min\{1, D\}$. The pointwise inequality $1 - \prod_h (1 - d(s_h)) \ge 1 - e^{-D} \ge (1 - e^{-1}) \min\{1, D\}$ then yields $I_i \ge \frac{1-e^{-1}}{22}\Delta$ for both i.

Conditionally on the deployed trajectory, $\mathbb{E}\sqrt{\Lambda_t} = \prod_h (1 - d(s_h))$, and the trajectory follows $\pi^\star$ with probability $1 - w_t$ and π^1 with probability w_t, so with $r = r_t$, $\sqrt{r} - \mathbb{E}[\sqrt{r_{t+1}} \mid \text{past}] = \sqrt{r}\,((1 - w_t) I_\star + w_t I_1) \ge \frac{1-e^{-1}}{22}\sqrt{r}\,\Delta$. By convexity of h^2 and $1 + r \ge 2\sqrt{r}$, the loss is $\mathsf{h}^2(P^{\pi^\star}, P^{(t)}) \le w_t \Delta \le \frac{\sqrt{r}}{2}\Delta$. Hence the loss is at most $\frac{11}{1-e^{-1}}(\sqrt{r_t} - \mathbb{E}[\sqrt{r_{t+1}} \mid \text{past}])$, and telescoping from $r_1 = 1$ gives the bound (the odds are a nonnegative supermartingale, so all expectations are finite).

Within-episode updating. Let $V(r) = \log(1 + r)$ and let $r_{t,h}$ be the odds before label h of episode t; the mixture at state s is $\mu_s = (p_s + r q_s)/(1 + r)$ with $p_s = \pi^\star(\cdot \mid s)$, $q_s = \pi^1(\cdot \mid s)$. For a label $A' \sim p_s$ the update is $r^+ = r q_s(A')/p_s(A')$, so $V(r) - V(r^+) = \log(p_s(A')/\mu_s(A'))$ and $V(r) - \mathbb{E}[V(r^+)] = \mathrm{KL}(p_s \| \mu_s) \ge 2\mathsf{h}^2(p_s, \mu_s) =: 2e_{t,h}$. Summing within the episode gives $2\mathbb{E}[\sum_h e_{t,h} \mid \text{past}] \le V(r_{t,1}) - \mathbb{E}[V(r_{t,H+1}) \mid \text{past}]$.

For within-episode updating, earlier labels influence subsequent actions. We therefore compare laws augmented by those labels before marginalizing to the action-state trajectories. Under the expert law, actions and independent labels follow p_s. Under the learner law, actions follow μ_s and labels follow p_s. Both laws run the same posterior recursion on the labels. At every augmented history, the conditional Hellinger distance between the step kernels is exactly $e_{t,h}$, because the label and transition kernels coincide. Marginalizing and applying the sequential comparison $\mathsf{h}^2(P, Q) \le 10\, \mathbb{E}_Q \sum_h e_h$ (the first half of the proof of Lemma 11, valid for singular kernels since L_h is a nonnegative supermartingale) gives $\mathsf{h}^2(P^{\pi^\star}, P^{(t)}) \le 10\, \mathbb{E}[\sum_h e_{t,h} \mid \text{past}] \le 5(V(r_{t,1}) - \mathbb{E}[V(r_{t,H+1}) \mid \text{past}])$. Since $r_{t,H+1} = r_{t+1,1}$, $r_{1,1} = 1$ and $V \ge 0$, telescoping gives $5 \log 2$. □

On the two-state instance, these alternatives coincide at the level of both the count process and the trajectory loss. Under both, the number K of alive-state labels in an episode with prior count b has $\mathbb{P}(K = k \mid b) = w_b 2^{-k}$ for $k < H$, and conditional loss $1 - \sqrt{(1 + 2^{-b-H})/(1 + 2^{-b})}$. For within-episode updating, the survival factors $\prod_{j<k}(1 - p_{b+j})$ telescope to $(1 + 2^{-b-k})/(1 + 2^{-b})$. This gives $u_b = 1/(2^b + 2)$ for $1 \le b < H$ and infinite-round loss $S_H = 0.3643$ to four decimals for both $H = 10^3$ and 10^6.

B.6 Proof for batch updates

Proof. Upper bound. To obtain the upper bound, charge each batch to a tower potential at the batch scale. Number all batches of all episodes consecutively, and let r_b, w_b be the odds and weight frozen during batch b, $D_b = \sum_{h \in b} d(s_h) \le L$. Conditionally on the history before the batch and its trajectory, the batch's labels are independent expert draws (the batch's actions do not depend on them), so $\mathbb{E}\Lambda_b \le 1$ and $\mathbb{E}\sqrt{\Lambda_b} = \prod_{h \in b}(1 - d(s_h)) \le e^{-D_b}$, and Lemma 12 at scale L (its proof uses only $D \le L$ in place of $D \le H$, with the tower stopped at $m_L = \min\{j : x_j \ge \log L\} \le \log^\star L$) gives $F_L(r_b) - \mathbb{E}[F_L(r_{b+1})] \ge c_0 \mathbb{E} \min\{1, w_b D_b\}$.

For the trajectory comparison, augment both episode laws by all labels, as in Theorem 2. At step h, the conditional Hellinger distance is $e_h \le w_{b(h)} d(s_h)$, a predictable budget, so $\mathsf{h}^2(P^{\pi^\star}, P^{(t)}) \le 22\,\mathbb{E}\min\{1, \sum_{b\in t} w_b D_b\} \le 22\,\mathbb{E}\sum_{b\in t}\min\{1, w_b D_b\}$. Summing over all batches through episode n and telescoping F_L from $F_L(1) = m_L + 1$ gives $\frac{22}{c_0}(m_L+1) < 56(1+\log^\star L)$. The bound $5L\log 2$ follows from the entropy potential $V(r) = \log(1+r)$ exactly as in Theorem 2, with the batch label-vector KL K_b satisfying $\sum_{h\in b} e_h \le \frac{L}{2} K_b$ by data processing.

Lower bound. In the self-truncating instance with pre-episode count b, while alive a completed batch adds its length to the count and the catastrophe probability per step is frozen at $p(b+jL)$ during batch j; on the expert's support the learner's density is the constant $\prod_j q(b+jL)^{\ell_j}$ with $q = 1-p$, so the episode loss is exactly $g_{L,H}(b) = 1 - \prod_j q(b+jL)^{\ell_j/2} \ge 1 - q(b)^{L/2}$.

For the lower bound, it suffices that the early failures occur before the first batch ends. On that event, batch updating cannot yet change the count process. Let $k = \log^\star L \ge 3$, so $L \ge T_{k-1}$. On the event $E = \{B_t^\infty \le T_t\ \forall t \ge 2\}$ of Lemma 10 (probability $\ge \frac{31}{32}$), for $i \le k-3$ the first-batch geometric draw satisfies $G_i \le T_{i+1} \le T_{k-2} < L$, so each of these episodes has its catastrophe within the first batch, its remaining labels are uninformative, and its count increment equals G_i; hence the batched count agrees with the frozen process for $2 \le t \le k-2$ on E and $\frac{L}{2}p(B_t) \ge L2^{-T_t-3} > 5$, giving $\mathbb{E}g_{L,H}(B_t) \ge \frac{31}{32}(1-e^{-5}) > \frac{7}{8}$; the first episode has loss $1 - (\frac{3}{4})^{L/2} > \frac{7}{8}$ as $L \ge 16$. Thus $S_{L,H} \ge \frac{7}{8}(k-2)$, and the remaining cases are as in Lemma 10. □

The protocol comparison can now be summarized with the feedback rule held fixed. Every label is obtained at a visited state after its action; deployment, update timing, and likelihood exponent are the remaining choices. All entries are $\sup_n$ of the expected cumulative loss:

execution	worst-case cumulative loss (two policies)	
pointwise mixture, weights frozen per episode	$\Theta(\log^\star H)$	Thm. 1
same, likelihood tempered by α	$\Theta(1/\alpha + \log^\star H)$	Thm. 3
one policy sampled per episode	$\Theta(1)$	Thm. 2
pointwise mixture, updated after every label	$\Theta(1)$	Thm. 2
pointwise mixture, refreshed every L steps	$\Theta(\log^\star L)$	Thm. 4

For untempered pointwise mixing, holding the weights fixed for L steps allows observations to accumulate without changing subsequent actions in that batch. The corresponding worst-case bound has tower height $\log^\star L$.

B.7 Metric scope and exact numerical evaluation

The loss metric is essential. Cumulative Hellinger loss has the stated bound, but cumulative expected return gaps admit no uniform bound over two-policy classes, even for $H = 1$. To see this, fix $0 < \delta \le \frac{1}{8}$ and use two actions, with reward 1 for action 1. Let $\pi^\star(1) = \frac{1}{2} + \delta$, $\pi^1(1) = \frac{1}{2}$. The round-t gap is δw_t. With $m = t-1$ labels, $\mathbb{E}w_t \ge \frac{1}{2}(1 - \mathrm{TV}(P_m, Q_m)) \ge \frac{1}{2}(1 - \sqrt{2m\delta^2}) \ge \frac{3}{8}$ for $t \le N = \lfloor 1/(32\delta^2)\rfloor$ (using $\mathbb{E}_P[q/(p+q)] \ge \frac{1}{2}\sum\min\{p,q\}$ and Pinsker with $\mathrm{KL}(P_m\|Q_m) \le 4m\delta^2$), so $\sum_{t\le N}\mathbb{E}[J(\pi^\star) - J(\mu_t)] \ge \frac{3}{8}\delta N \ge \frac{3}{512\delta} \to \infty$ as $\delta \downarrow 0$. An expected-return bound that depends on the number of rounds is still available. By $\mathrm{TV} \le \sqrt{2\mathsf{h}^2}$ and Cauchy-Schwarz, Theorem 1 gives $\sum_{t\le n}\mathbb{E}\,|J(\pi^\star) - J(\mu_t)| \le \sqrt{112\,n(1+\log^\star H)}$.

Corollary 14 (Return separation on the two-state instance)**.** *In the MDP of Proposition 9, give reward one precisely when the entire episode avoids action †, and zero otherwise. Use the same uniform prior and clean-label Bayesian updates. For deployment $d \in \{\mathrm{PS}, \mathrm{PE}, \mathrm{WU}\}$, let $g_t^d = \mathsf{h}^2(P^{\pi^\star}, P_t^d)$ and $\Delta_t^d = J(\pi^\star) - J_t^d$, conditional on the history before round t. Then*

$$g_t^d \le \Delta_t^d = 2g_t^d - (g_t^d)^2 \le 2g_t^d.$$

Consequently, $\sum_{t\ge1}\mathbb{E}\Delta_t^{\mathrm{PS}} = \Theta(1+\log^\star H)$, whereas both per-episode sampling and per-label updating have cumulative expected return gap $\Theta(1)$, uniformly over $H \ge 1$. These statements concern this construction and reward, not all stochastic classes.

Proof. The reward is Markov after adjoining time: set all rewards before step H to zero, and set $r_H(s,u) = \mathbf{1}\{s = a,\ u \in \{0,1\}\}$, where a is the active state. This includes failure on the last

action. The expert has return one, and a learner's return equals its probability of avoiding failure. Rewards do not enter the posterior updates.

Condition on the pre-episode odds r. On the expert's trajectory support, the learner's density is a constant c_d times the expert's density. Frozen per-step mixing gives

$$c_{\mathrm{PS}} = \left(1 - \frac{r}{2(1+r)}\right)^H, \qquad c_{\mathrm{PE}} = \frac{1 + r2^{-H}}{1+r}.$$

For per-label updating, each label on a surviving path halves the odds, independently of its identity. Marginalizing over those labels therefore gives

$$c_{\mathrm{WU}} = \prod_{j=0}^{H-1} \left(1 - \frac{r2^{-j}}{2(1+r2^{-j})}\right) = \prod_{j=0}^{H-1} \frac{1 + r2^{-(j+1)}}{1 + r2^{-j}} = \frac{1 + r2^{-H}}{1+r}.$$

For every deployment, the trajectory Hellinger loss is $1 - \sqrt{c_d}$ and the return gap is $1 - c_d$. Substitution gives the identity and inequalities. Taking expectations and summing over rounds transfers the upper bounds in Theorems 1 and 2, and the instance lower bound in Lemma 10. The positive constant lower bound for each alternative already follows from its first-round return gap $(1 - 2^{-H})/2 \geq 1/4$. □

The recurrence in Proposition 9 can be evaluated directly, providing a reference for simulations of the learning process. Each count value b of B_t is visited at most once, so the infinite-round loss is $S_H = \sum_{b \geq 0} u_b \, g_H(b)$. The visit probabilities satisfy $u_0 = 1$ and $u_b = \sum_{a<b} u_a q_a(b-a)$, where $q_a(j) = p(a)(1-p(a))^{j-1}$ for $j < H$ and $(1-p(a))^{H-1}$ for $j = H$. The kernel is zero for larger increments. Truncating at $b \leq M$ costs at most $H2^{-M-1}$. Directed-rounding evaluation gives $S_H = 2.2653,\ 2.5411,\ 2.6340,\ 2.7698,\ 3.0433,\ 3.3319$ for $H = 10^3,\ 2^{16},\ 10^6,\ 10^9,\ 10^{30},\ 2^{1024}$ (rounded to four decimals; each value is certified to $\pm 10^{-15}$ before rounding), and $S_{2^{65536}} \geq 3.1861$, so $S_{2^{65536}} - S_{2^{16}} > 0.64$. The finite-horizon values change slowly. Their asymptotic order is established by the bounds, not by extrapolating these numbers: Lemma 10 gives the increasing envelope $S_{\lceil T_j \rceil} \geq \frac{7}{8}(j-1)$ for $j \geq 2$. The growth is $\Theta(\log^\star H)$; in particular, $S_H / \log\log H \to 0$.

C Proofs of the deterministic upper bounds

C.1 Disagreement and survival notation

Survival does not imply agreement on a fresh trajectory. For each non-expert policy π_i, record the event that it disagrees with the expert somewhere and the probability of that event:

$$D_i = \{\tau : \exists t,\ \pi_i(s_t) \neq \pi^\star(s_t)\} \quad \text{(trajectories on which } \pi_i \text{ disagrees somewhere)}, \qquad p_i = \mathbb{P}_{\tau \sim \pi^\star}(D_i),$$

and its survival indicator by $I_i = \mathbf{1}\{\pi_i \in \mathcal{V}\}$. Independence across training trajectories gives $\mathbb{E} I_i = (1-p_i)^n$. All policies are tested on the same data, so the indicators $(I_i)_i$ need not be independent. The analysis below permits this dependence.

The two deployment rules can first be compared through a single event: the learner ever chooses an action different from the expert. Run the learner and expert from the same initial state with the same transition randomness until the learner first disagrees with $\pi^\star$. Without a disagreement, their trajectories and returns coincide. With a disagreement, their return difference is at most 1 by normalization. Thus either rule satisfies

$$J(\pi^\star) - J(\pi_{\mathrm{R}}(\mathcal{V})) \ \leq \ d_{\mathrm{R}}(\mathcal{V}) := \mathbb{P}\big(\text{the learner deviates from } \pi^\star \text{ somewhere} \mid \mathcal{V}\big), \tag{4}$$

The deterministic constructions make a deviation lose the full return and hence attain equality. Until the first deviation, the learner follows the states $s_1, \ldots, s_H$ of the coupled expert trajectory $\tau \sim \pi^\star$. A surviving policy can disagree with $\pi^\star$ only at a state absent from the training data.

C.2 Deviation along a fixed expert trajectory

Lemma 15 (Conditional deviation probabilities)**.** *Fix the version space $\mathcal{V}$ with $q = |\mathcal{V}|$ and let $\tau = (s_1, \ldots, s_H) \sim \pi^\star$ be the coupled expert trajectory. Let $a_t = a_t(\mathcal{V}, \tau) = |\{\pi \in \mathcal{V} : \pi(s_t) \neq \pi^\star(s_t)\}|$.*

(a) $d_{\mathrm{PE}}(\mathcal{V}) = \frac{1}{q}\sum_{i=1}^{M} I_i p_i$.

(b) $d_{\mathrm{PS}}(\mathcal{V}) = \mathbb{E}_\tau\big[1 - \prod_{t=1}^{H}(1 - a_t/q)\big]$, *and* $d_{\mathrm{PS}}(\mathcal{V}) \le \min\{1, H/q\}\sum_{i=1}^{M} I_i p_i$.

Proof. (a) PE chooses one uniformly random $\pi \in \mathcal{V}$ and uses it throughout the episode. On the coupled expert trajectory, it departs from $\pi^\star$ precisely when τ satisfies $\tau \in D_\pi$. The expert choice $\pi = \pi^\star$ contributes no deviation. Averaging over π and then τ therefore yields $\frac{1}{q}\sum_{\pi\in\mathcal{V}\setminus\{\pi^\star\}} p_\pi = \frac{1}{q}\sum_i I_i p_i$.

(b) Conditional on τ, the PS draws at the successive expert states are independent. At s_t, a draw disagrees with $\pi^\star$ with probability a_t/q; the first such disagreement is exactly the learner's first deviation. Thus the deviation probability is $1 - \prod_t(1 - a_t/q)$. To account for repeated opportunities, use the union bound $1 - \prod_t(1 - a_t/q) \le \sum_t a_t/q$ and count how often each survivor can disagree: $\sum_{t=1}^{H} a_t = \sum_{\pi\in\mathcal{V}} |\{t : \pi(s_t) \ne \pi^\star(s_t)\}| \le H\sum_{\pi\in\mathcal{V}} \mathbf{1}\{\tau \in D_\pi\}$. Averaging with $\mathbb{E}_\tau$ gives $d_{\mathrm{PS}}(\mathcal{V}) \le \frac{H}{q}\sum_i I_i p_i$. There is also a bound independent of the number of draws: $1 - \prod_t(1 - a_t/q) \le \mathbf{1}\{\exists t : a_t > 0\} \le \sum_{\pi\in\mathcal{V}} \mathbf{1}\{\tau \in D_\pi\}$. Indeed, any survivor that disagrees at s_t has $\tau \in D_\pi$. This gives $d_{\mathrm{PS}}(\mathcal{V}) \le \sum_i I_i p_i$; combining the bounds yields $\min\{1, H/q\}\sum_i I_i p_i$. □

C.3 From survivor counts to the offline rate

Both rules lead to $w_r(q)\sum_i I_i p_i$, with $w_r(q) := \min\{1, r/q\}$ for $r \ge 1$: PE uses $r = 1$, and PS uses $r = H$. We can therefore average over the training data in a single calculation. Count the surviving policies whose disagreement probabilities exceed a given threshold, then integrate over thresholds. The calculation uses each policy's survival probability individually; it allows arbitrary dependence among the survival indicators.

Lemma 16 (Survival-count integral). *For $r \ge 1$, $M \ge 1$, $n \ge 1$,*

$$\mathbb{E}\Big[w_r(q)\sum_{i=1}^{M} I_i p_i\Big] \le U_r(M,n) := \int_0^1 \min\{r,\, M(1-t)^n\}\,dt = \begin{cases} \dfrac{M}{n+1}, & M \le r, \\ r\Big[1 - \dfrac{n}{n+1}\Big(\dfrac{r}{M}\Big)^{1/n}\Big], & M > r. \end{cases}$$

Moreover $U_H(M,n) \le A_H(M)/n$ and $U_1(M,n) \le (1+\log M)/n \le 2\log(M+1)/n$.

Proof. Let $C(t) = \sum_i I_i \mathbf{1}\{p_i > t\}$ for $t \in [0,1]$. The identity $p_i = \int_0^1 \mathbf{1}\{p_i > t\}\,dt$ writes a policy's disagreement probability as an integral of threshold indicators. Summing over the survivors gives $w_r(q)\sum_i I_i p_i = \int_0^1 w_r(q)C(t)\,dt$. Because $C(t) \le \sum_i I_i \le q$, the weighted count obeys $w_r(q)C(t) \le \min\{C(t),\, rC(t)/q\} \le \min\{r, C(t)\}$.

The expected count is controlled by the survival probabilities: $\mathbb{E}C(t) = \sum_{i:p_i>t}(1-p_i)^n \le M(1-t)^n$. Concavity of $x \mapsto \min\{r, x\}$, followed by Jensen's inequality and Tonelli's theorem, yields $\mathbb{E}\int_0^1 \min\{r, C(t)\}dt \le \int_0^1 \min\{r, \mathbb{E}C(t)\}dt \le \int_0^1 \min\{r, M(1-t)^n\}dt = U_r(M,n)$.

It remains to evaluate this integral. If $M \le r$, its integrand is $M(1-t)^n$ on the whole interval, so its value is $M/(n+1)$. If $M > r$, the integrand changes form at $t_0 = 1 - (r/M)^{1/n}$. Integrating on either side gives $U_r = rt_0 + \int_{t_0}^1 M(1-t)^n dt = rt_0 + \frac{M}{n+1}(1-t_0)^{n+1} = r\big(1 - (r/M)^{1/n}\big) + \frac{r}{n+1}(r/M)^{1/n} = r\big[1 - \frac{n}{n+1}(r/M)^{1/n}\big]$.

To read off the dependence on horizon and class size, apply $(1-t)^n \le e^{-nt}$ and extend the domain to $[0,\infty)$: $U_r \le \int_0^\infty \min\{r, Me^{-nt}\}dt$. For $M \le r$, the resulting integral equals M/n. For $M > r$, the breakpoint is $t_1 = \frac{1}{n}\log(M/r)$, and integration gives $rt_1 + \int_{t_1}^\infty Me^{-nt}dt = \frac{r}{n}\log\frac{M}{r} + \frac{r}{n} = \frac{r}{n}\big(1 + \log\frac{M}{r}\big)$. Taking $r = H$ gives $A_H(M)/n$ in both cases because $h = \min\{H, M\}$. Taking $r = 1$ gives $U_1 \le (1+\log M)/n$, also for $M = 1$. The remaining comparison follows from $2\log(M+1) - \log M = \log\big(M + 2 + 1/M\big) \ge \log 4 > 1$, which implies $1 + \log M \le 2\log(M+1)$. □

Proof of the upper bounds in Theorem 5. Combine the first-deviation reduction (4) with Lemma 15 to obtain $\mathrm{GAP}_{\mathrm{PE}} \le \mathbb{E}[w_1(q)\sum_i I_i p_i]$ and $\mathrm{GAP}_{\mathrm{PS}} \le \mathbb{E}[w_H(q)\sum_i I_i p_i]$. Lemma 16 then gives $\min\{1, U_1\} \le \min\{1, 2\log B/n\}$ and $\min\{1, U_H\} \le \min\{1, A_H(M)/n\}$. The normalization of returns supplies the cap at 1.

With $n = 0$, all policies remain: $\mathcal{V} = \mathcal{B}$. PE selects $\pi^\star$ with probability $1/B$, so its gap is at most $1 - 1/B$. PS selects $\pi^\star$ at each of the H steps with probability B^{-H}, which bounds its gap by $1 - B^{-H}$. Both bounds are attained by an MDP that returns 1 exactly when all actions match the expert, with each of the M remaining policies selecting its own distinct non-expert action at every state. This construction uses $|\mathcal{A}| \ge B$. □

Remark 17 (The potential Φ_H)**.** The interactive proof will track how much the version space shrinks after each episode. For this purpose, $A_H(M)$ has a useful sum representation: up to a factor 2, it equals the clipped harmonic potential $\Phi_H(M) = \sum_{j=1}^{M} \min\{1, H/j\}$: if $H \ge M$ both equal M; otherwise $\Phi_H(M) = H + H\sum_{j=H+1}^{M} \frac{1}{j} \le H(1 + \log\frac{M}{H}) = A_H(M) \le \Phi_H(M) + H\log 2 \le 2\Phi_H(M)$. The potential form lets us sum the decreases across interactive episodes in Section 6.

D Constructions attaining the deterministic laws

Can the repeated opportunities in the upper bound occur on a realizable instance? We assign each non-expert policy its own random disagreement event. Training eliminates that policy only when the event is observed. At test time, a surviving event can expose the learner to the same wrong policy at every step, which attains the per-step amplification.

D.1 An MDP with independent disagreement regions

Fix $M \ge 1$, $H \ge 1$ and $p \in (0, 1)$. At the start of every episode, the learner observes a freshly drawn vector $X = (X_1, \ldots, X_M)$ of independent $\mathrm{Bernoulli}(p)$ bits. The *alive* states are (t, X) for $t = 1, \ldots, H$; an absorbing *failure* state also carries the time index. There are two actions. Action 0 at (t, X) advances to $(t+1, X)$ when $t < H$. Action 1 sends the learner to failure. The only reward is 1, received for action 0 at the final alive state (H, X), so every trajectory returns either 0 or 1.

The expert always chooses 0. Non-expert policy π_i chooses X_i at each alive state (t, X) and 0 in failure. The resulting class contains $B = M + 1$ distinct deterministic stationary policies. The only reward is lost as soon as the learner deviates, so (4) holds with equality.

Policy π_i disagrees with $\pi^\star$ exactly on episodes with $X_i = 1$. It survives the n training episodes exactly when all of them have $X_i = 0$. Independence of the coordinates makes the survival indicators independent Bernoulli(a) variables with $a = (1-p)^n$. The count of surviving non-expert policies is $K = \sum_i I_i \sim \mathrm{Bin}(M, a)$. On a fresh episode with bits X, let $L = \sum_i I_i X_i$ count the survivors that disagree. Conditional on K, we have $L \sim \mathrm{Bin}(K, p)$. PE fails precisely when the policy selected at the start of the episode disagrees, giving

$$\mathrm{GAP}_{\mathrm{PE}} = p\,\mathbb{E}\Big[\frac{K}{K+1}\Big]. \tag{5}$$

For PS, the disagreement probability remains fixed at $L/(K+1)$ while the learner remains alive. Repeated draws therefore give

$$\mathrm{GAP}_{\mathrm{PS}} = \mathbb{E}\Big[1 - \Big(1 - \frac{L}{K+1}\Big)^H\Big]. \tag{6}$$

Lemma 18 (Elementary inequalities)**.** *(a) For $x \ge 0$, $1 - e^{-x} \ge \frac{1}{2}\min\{1, x\}$. (b) For $0 \le x \le 1$ and $H \ge 1$, $1 - (1-x)^H \ge \frac{Hx}{1+Hx}$. (c) For $x \ge 1$, $\log(2x) \ge (\log 2)(1 + \log x)$. (d) $\log 2 \ge 2/3$.*

Proof. (a) The two ranges give the bound directly: for $x \ge 1$, $1 - e^{-x} \ge 1 - e^{-1} \ge 1/2$; for $x \le 1$, $1 - e^{-x} \ge x - x^2/2 \ge x/2$. (b) The desired form follows by rearranging $(1-x)^H \le e^{-Hx} \le 1/(1+Hx)$. (c) At $x = 1$, both sides equal $\log 2$. The left derivative $1/x$ is at least the right derivative $(\log 2)/x$, which preserves the ordering beyond that point. (d) Jensen's inequality gives $\log 2 = \int_1^2 dx/x \ge 2/3$ because $1/x$ is convex: the integral is at least $1/\bar{x}$ for $\bar{x} = 3/2$. □

D.2 PER-EPISODE SAMPLING

The PE lower bound needs few survivors with appreciable disagreement probabilities. We choose the event probability so that the expected number of non-expert survivors is below one, but the chance of retaining at least one stays bounded away from zero.

Proof of the PE lower bound in Theorem 5. Set $p = 1 - (2M)^{-1/n}$. Then $a = 1/(2M)$ and $\mathbb{E}K = 1/2$, while $\mathbb{P}(K \geq 1) = 1 - (1 - \frac{1}{2M})^M \geq 1 - e^{-1/2} \geq \frac{1}{4}$. On the event that some non-expert policy survives, its total sampling weight is at least half: $K/(K+1) \geq \frac{1}{2}\mathbf{1}\{K \geq 1\}$. Substituting into (5) yields $\mathrm{GAP}_{\mathrm{PE}} \geq \frac{p}{2}\mathbb{P}(K \geq 1) \geq \frac{p}{8}$. Lemma 18(a) gives $p = 1 - e^{-\log(2M)/n} \geq \frac{1}{2}\min\{1, \log(2M)/n\}$. Since $\log(2M) \geq \log(M+1) = \log B$, we obtain $\mathrm{GAP}_{\mathrm{PE}} \geq \frac{1}{16}\min\{1, \log B/n\}$. □

The denominator $|\mathcal{V}|$ in Lemma 15(a) therefore cannot reduce the worst-case rate to $1/n$. The expected number of non-expert survivors is only one half. Their contribution remains large because a surviving policy disagrees on a new episode with probability $\approx \log(2M)/n$.

D.3 PER-STEP MIXING

For PS, we must also account for the number of draws. The next lemma converts the binomial number of disagreeing survivors into a deviation probability. The bound saturates when repeated draws already cause a constant failure probability.

Lemma 19 (Binomial amplification)**.** *For $K \geq 1$, $p \in (0,1)$ and $L \sim \mathrm{Bin}(K,p)$,*

$$g_H(K,p) := \mathbb{E}\Big[1 - \Big(1 - \frac{L}{K+1}\Big)^H\Big] \;\geq\; \frac{HKp}{K+1+H+HKp} \;\geq\; \frac{1}{6}\min\{1,\, p\min(H,K)\}.$$

Proof. Let $Z = HL/(K+1)$. Lemma 18(b) reduces the desired probability to $g_H \geq \mathbb{E}[Z/(1+Z)]$. The Cauchy-Schwarz inequality gives $(\mathbb{E}Z)^2 = (\mathbb{E}[\sqrt{Z/(1+Z)}\sqrt{Z(1+Z)}])^2 \leq \mathbb{E}[Z/(1+Z)]\,\mathbb{E}[Z+Z^2]$, so $\mathbb{E}[Z/(1+Z)] \geq (\mathbb{E}Z)^2/\mathbb{E}[Z+Z^2]$. The binomial moments are $\mathbb{E}Z = HKp/(K+1)$ and $\mathbb{E}Z^2 = H^2(Kp(1-p) + K^2p^2)/(K+1)^2 \leq H^2(Kp + K^2p^2)/(K+1)^2$, so

$$\mathbb{E}\Big[\frac{Z}{1+Z}\Big] \geq \frac{H^2K^2p^2/(K+1)^2}{HKp/(K+1) + H^2(Kp+K^2p^2)/(K+1)^2} = \frac{HKp}{(K+1)+H+HKp}.$$

To expose the saturation scale, suppose first that $p\min(H,K) \geq 1$. Then $HKp \geq \max(H,K) \geq \frac{1}{3}(K+1+H)$, and the ratio is at least $\frac{HKp}{3HKp+HKp} = \frac{1}{4} \geq \frac{1}{6}$. If instead $p\min(H,K) < 1$, then $HKp < \max(H,K) \leq K+1+H$, giving the bound $\frac{HKp}{2(K+1+H)} \geq \frac{HKp}{6\max(H,K)} = \frac{p\min(H,K)}{6}$. Here we used $K+1+H \leq 3\max(H,K)$, which holds for $K \geq 1$, and $HK/\max(H,K) = \min(H,K)$. □

Proof of the PS lower bound in Theorem 5. Choose $h = \min\{H, M\}$ and $p = 1 - (h/(2M))^{1/n}$. This sets $a = h/(2M)$ and $\lambda := \mathbb{E}K = h/2$. To show that many policies survive with constant probability, use the second moment $\mathbb{E}K^2 = \lambda^2 + Ma(1-a) \leq \lambda^2 + \lambda$, so the Paley-Zygmund inequality gives $\mathbb{P}(K \geq \lambda/2) \geq \frac{(\lambda/2)^2}{\mathbb{E}K^2} \geq \frac{\lambda^2}{4(\lambda^2+\lambda)} = \frac{\lambda}{4(\lambda+1)} \geq \frac{1}{12}$ because $\lambda \geq 1/2$. On the event $K \geq h/4$, we have $\min(H,K) \geq h/4$ since $h \leq H$. Applying (6) and Lemma 19, $\mathrm{GAP}_{\mathrm{PS}} \geq \frac{1}{12}\cdot\frac{1}{6}\min\{1, ph/4\} \geq \frac{1}{288}\min\{1, hp\}$. By Lemma 18(a), $hp = h(1 - e^{-\log(2M/h)/n}) \geq \frac{h}{2}\min\{1, \log(2M/h)/n\} \geq \frac{1}{2}\min\{1, h\log(2M/h)/n\}$, and by Lemma 18(c),(d), $h\log(2M/h) \geq (\log 2)\,h(1 + \log(M/h)) \geq \frac{2}{3}A_H(M)$. Hence $\mathrm{GAP}_{\mathrm{PS}} \geq \frac{1}{288}\cdot\frac{1}{2}\cdot\frac{2}{3}\min\{1, A_H(M)/n\} = \frac{1}{864}\min\{1, A_H(M)/n\}$. □

D.4 THE RARE-CONTEXT INSTANCE

The following construction replaces independent disagreement bits by mutually exclusive rare contexts. It also underlies the refutation of the WARM-STAGGER conjecture in Section 7. Set $m \geq 2$,

$n \geq \max\{m, 2\}$, $H = m$, $B = m + 1$. Draw a context $j \in \{0, 1, \ldots, m\}$ at the start of the episode. Each context $j \geq 1$ has probability $1/n$, and context 0 receives the remaining mass. The context stays in the state, with actions and rewards as in Section D.1. Policy π_j chooses action 1 at every alive state in context j and 0 elsewhere. Thus observing context j under the expert eliminates exactly π_j. Let K count the contexts $j \geq 1$ absent from the data. The two gaps are

$$\mathrm{GAP}_{\mathrm{PE}} = \frac{1}{n}\,\mathbb{E}\Big[\frac{K}{K+1}\Big], \qquad \mathrm{GAP}_{\mathrm{PS}} = \frac{1}{n}\,\mathbb{E}\Big[K\Big(1 - \Big(\frac{K}{K+1}\Big)^m\Big)\Big], \tag{7}$$

because each unobserved context j appears with probability $1/n$. In that context, PE fails exactly when its selected policy is π_j; PS fails when any of its m independent draws selects π_j. Using $\mathbb{E}K = m(1-1/n)^n \geq m/4$ for $n \geq 2$ and $1-(K/(K+1))^m \geq 1-e^{-m/(K+1)} \geq 1-e^{-1/2} \geq \frac{1}{4}$ for $1 \leq K \leq m$ gives

$$\frac{1}{8n} \leq \mathrm{GAP}_{\mathrm{PE}} \leq \frac{1}{n}, \qquad \frac{m}{16n} \leq \mathrm{GAP}_{\mathrm{PS}} \leq \frac{m}{n}.$$

For $n = m^2$, the gaps become $\mathrm{GAP}_{\mathrm{PS}} = \Theta(1/m)$ and $\mathrm{GAP}_{\mathrm{PE}} = \Theta(1/m^2)$, a factor-$H$ separation. Here the PE gap is smaller than its worst-case order $\Theta(\log B/n)$. For a finite-size comparison, evaluating (7) using a recursion for the law of K gives $(n/m)\mathrm{GAP}_{\mathrm{PS}} = 0.288, 0.315, 0.329$ and $n\,\mathrm{GAP}_{\mathrm{PE}} = 0.51, 0.70, 0.84$ at $m = 4, 8, 16$. Their respective limits are $e^{-1}(1-e^{-e}) = 0.3436\ldots$ and 1.

E Interaction and final-iterate proofs

E.1 Accounting for removed policies

Does collecting labels on the learner's own trajectory remove the deployment cost? In each interactive episode, the learner deploys its current version space $\mathcal{V}_t$ using the chosen rule. The expert supplies $\pi^\star(s)$ at every visited state, and the learner removes inconsistent policies after the episode: $\mathcal{V}_{t+1} = \{\pi \in \mathcal{V}_t : \pi(s) = \pi^\star(s) \text{ on all visited } s\}$. Write $d_t = d_{\mathrm{R}}(\mathcal{V}_t)$ for the conditional deviation probability in episode t (Lemma 15); by (4), it bounds that episode's return gap.

A deviation now provides evidence as well as incurring loss: its label eliminates every candidate that disagrees there. We measure this progress with a potential of the version-space size. If its expected decrease dominates the episode's deviation probability, summing over episodes bounds the entire learning process.

Theorem 20 (Interactive potential bound)**.** *Let $\Psi_r(q) = \sum_{j=2}^{q} \min\{1, r/j\}$, $\Psi_r(1) = 0$. For N interactive episodes starting from any version space $\mathcal{V}_0$ (e.g. the offline one),*

$$\mathbb{E}\sum_{t=0}^{N-1} d_t \;\leq\; \mathbb{E}\,\Psi_r(|\mathcal{V}_0|) \qquad (r = H \textit{ under PS},\ r = 1 \textit{ under PE}),$$

hence $\mathbb{E}\sum_{t<N} d_t \leq A_H(M)$ under PS and $\leq \log B$ under PE.

Proof. Fix the version space $\mathcal{V}$ of size q and a coupled expert trajectory τ. Record which candidates have disagreed along each prefix: $S_t = \{\pi \in \mathcal{V} : \pi \text{ disagrees with } \pi^\star \text{ at one of } s_1, \ldots, s_t\}$, $k_t = |S_t|$ with $k_0 = 0$. Let T be the first-deviation step, setting $T = H$ if no deviation occurs. The learner visits $s_1, \ldots, s_T$, so annotation removes all policies in S_T and may remove others as well. Hence $|\mathcal{V}'| \leq q - k_T$. Monotonicity of Ψ_r then implies $\Psi_r(q) - \Psi_r(|\mathcal{V}'|) \geq \Psi_r(q) - \Psi_r(q-k_T) = \sum_{j=q-k_T+1}^{q} \min\{1, r/j\} \geq k_T \min\{1, r/q\}$.

PS. Sum the probabilities of a first disagreement to get $\mathbb{E}_{\text{draws}} \sum_{t\leq T} a_t/q = \sum_t \frac{a_t}{q} \prod_{s<t}(1-\frac{a_s}{q}) = 1 - \prod_t (1-\frac{a_t}{q})$. The summand indexed by t is the probability of the first disagreement occurring at step t. Their sum is therefore the deviation probability conditional on τ. Denote it by $\delta(\tau)$. Every survivor counted in a_t for $t \leq T$ belongs to S_T and appears at most $T \leq H$ times. Therefore $\sum_{t\leq T} a_t \leq H k_T$.

When $q \geq H$, we have $\min\{1, H/q\} = H/q$. The potential drop is at least $k_T H/q \geq \sum_{t\leq T} a_t/q$, whose expectation over the draws is $\delta(\tau)$. When $q < H$, we have $\min\{1, H/j\} = 1$ for every $j \leq q$.

The drop is then at least k_T and hence ≥ 1 on a deviation, since the disagreeing policy belongs to S_T. Its expectation is again at least $\delta(\tau)$. Averaging over τ yields $\mathbb{E}[\Psi_H(q) - \Psi_H(|\mathcal{V}'|) \mid \mathcal{V}] \geq d_{\mathrm{PS}}(\mathcal{V})$.

PE. Reaching s_t before a deviation requires the selected policy to avoid the earlier disagreements, an event with probability $(q - k_{t-1})/q$ because it must lie outside S_{t-1}. On this event $T \geq t$, so annotation removes every policy in S_t. We can charge the policies in $S_t \setminus S_{t-1}$ to disjoint portions of the potential decrease. Whenever the removed set contains S_t, that part is at least $\sum_{j=q-k_t+1}^{q-k_{t-1}} \frac{1}{j} \geq \frac{k_t - k_{t-1}}{q - k_{t-1}}$. It follows that $\mathbb{E}[\Psi_1(q) - \Psi_1(|\mathcal{V}'|) \mid \mathcal{V}, \tau] \geq \sum_t \frac{q-k_{t-1}}{q} \cdot \frac{k_t - k_{t-1}}{q - k_{t-1}} = \frac{k_H}{q}$, which equals the PE deviation probability conditional on τ: the selected policy deviates exactly when it belongs to S_H. For either rule, summing the expected decreases over episodes gives $\mathbb{E}\Psi_r(|\mathcal{V}_0|) - \mathbb{E}\Psi_r(|\mathcal{V}_N|) \leq \mathbb{E}\Psi_r(|\mathcal{V}_0|)$. The stated class-size bounds follow from $\Psi_H(B) \leq \Phi_H(M) \leq A_H(M)$ and $\Psi_1(B) = \sum_{j=2}^{B} 1/j \leq \log B$. □

E.2 Proof of the hybrid law

Proof. Upper bounds. The offline bound continues to hold after further consistent pruning. To see this, apply Lemma 16. Indeed, for any random $\mathcal{W} \subseteq \mathcal{V}_0$ that contains $\pi^\star$, $w_r(|\mathcal{W}|)\#\{\pi_i \in \mathcal{W} : p_i > t\} \leq \min\{r, \#\{\pi_i \in \mathcal{V}_0 : p_i > t\}\}$. The threshold-count argument in that proof therefore applies with $\mathcal{W}$ replacing $\mathcal{V}_0$, and every interactive iterate satisfies $\mathbb{E}d_t \leq U_r(M, n)$. Omit this bound when $n = 0$. Separately, averaging Theorem 20 gives $\frac{1}{N}\mathbb{E}\sum_{t<N} d_t \leq \Psi_r(B)/N$. Taking the smaller of the offline and interactive bounds gives $\min\{1, A_H(M)/n, A_H(M)/N\} \leq \min\{1, 2A_H(M)/T\}$ under PS and $\min\{1, 2\log B/n, \log B/N\} \leq \min\{1, 3\log B/T\}$ under PE, using $\min\{x/n, x/N\} \leq 2x/(n+N)$ and $\min\{2x/n, x/N\} \leq 3x/(n+N)$.

Lower bounds. Take the MDP of Section D.1. Choose p as in each offline lower bound, replacing n by T. Annotation at the initial state $(1, X)$ reveals $\pi^\star(1, X) = 0$ and removes exactly the survivors i for which $X_i = 1$. The learner's actions cannot change this update. After s episodes of either kind, the survivor count thus has law $K_s \sim \mathrm{Bin}(M, (1-p)^s)$. The survivors after T episodes form a subset of those after $s \leq T$ episodes, giving $K_s \geq K_T$ on every realization. Here K_T has the offline law with T samples, with $\mathbb{E}K_T = 1/2$ for the PE choice and $h/2$ for the PS choice.

This coupling reduces the remaining calculation to the offline lower bounds. For PE, $K/(K+1)$ increases with K, so $\mathbb{E}[K_s/(K_s+1)] \geq \mathbb{E}[K_T/(K_T+1)] \geq \frac{1}{2}\mathbb{P}(K_T \geq 1) \geq \frac{1}{8}$. For PS, the event $K_T \geq h/4$ has probability $\geq 1/12$ by the Paley-Zygmund argument for K_T. On that event $K_s \geq h/4$, so Lemma 19 applies with $\min(H, K_s) \geq h/4$. Each iterate has the same lower bound with the same constants, as does their average. □

E.3 The final version space

The averaged-iterate guarantee does not directly characterize the final version space. For that output, the offline bound survives pruning, and a second argument counts all candidates that can be eliminated. The resulting bounds are as follows.

Proposition 21 (Final version space). *For deployment from the final version space $\mathcal{V}_T$ after $T = n + N \geq 1$ episodes,*

$$\tfrac{1}{864}\min\Big\{1, \tfrac{A_H(M)}{T}\Big\} \leq \mathcal{R}_{\mathrm{PS}}^{\mathrm{last}} \leq \min\Big\{1, \tfrac{A_H(M)}{n}, \tfrac{M}{T+1}\Big\}, \qquad \tfrac{1}{16}\min\Big\{1, \tfrac{\log B}{T}\Big\} \leq \mathcal{R}_{\mathrm{PE}}^{\mathrm{last}} \leq \min\Big\{1, \tfrac{2\log B}{n}, \tfrac{M}{T+1}\Big\},$$

where the $1/n$ terms are omitted when $n = 0$.

Proof. The $1/n$ terms follow from the offline bounds, which remain valid under further consistent pruning. The lower bounds follow as in Theorem 6: on the Bernoulli MDP, the final version space has the same law as an offline version space after T samples.

For the $M/(T+1)$ term, let $u(\mathcal{V}) = \mathbb{P}_{\tau \sim \pi^\star}(\exists \pi \in \mathcal{V} : \tau \in D_\pi)$ be the chance that some surviving policy disagrees on a fresh expert trajectory. Both rules have deviation probability at most $u(\mathcal{V})$. Whenever this event occurs, an episode of either kind eliminates at least one survivor. Offline, the whole expert trajectory is annotated. Interactively, the learner reaches the first expert state at which any survivor disagrees, and that state is annotated. Consequently, $\mathbb{E}[|\mathcal{V}_t| - |\mathcal{V}_{t+1}| \mid \mathcal{V}_t] \geq u(\mathcal{V}_t)$. Since pruning makes $u(\mathcal{V}_t)$ non-increasing in t, we may append one hypothetical episode and sum to obtain $(T+1)\mathbb{E}u(\mathcal{V}_T) \leq \sum_{t=0}^{T} \mathbb{E}u(\mathcal{V}_t) \leq M$. □

The upper and lower rates agree when $n \gtrsim N$. When interaction dominates, the upper and lower bounds can still differ by factors of at most $M/A_H(M)$ for PS and $M/\log B$ for PE. The exact last-iterate law in that regime remains open (Section G).

F Algorithm details and proofs for Warm-Stagger

F.1 The original guarantee and deployment rule

Warm-Stagger (their Algorithm 2) initializes an online learner with the data-consistent class $\mathcal{B}_{\mathrm{bc}} = \{\pi \in \mathcal{B} : \pi(s_h) = a_h \ \forall (s,a)_{1:H} \in D_{\mathrm{off}}\}$. In round $n = 1, \dots, N_{\mathrm{int}}$, it obtains and executes a policy π^n, then samples $s^n \sim d^{\pi^n}$ by choosing a uniformly random time step from the rollout. It queries the expert at s^n and updates the learner using log loss $\ell^n(\pi) = \log(1/\pi(a^{\star,n} \mid s^n))$. The output is a first-step uniform mixture of $\pi^1, \dots, \pi^{N_{\mathrm{int}}}$.

The per-round deployment is an exponential-weighted distribution over *each-step mixing policies*. For $u \in \Delta(\mathcal{B})$, the deployed policy is $\bar{\pi}_u(a \mid s) = \sum_\pi u(\pi)\pi(a \mid s)$, "executed by drawing $\pi \sim u$ freshly-at-random at each step" (their Definition 1). The weights start uniformly (their Proposition 41). Their Theorem 6 assumes μ-recoverability, returns in $[0, R]$, and deterministic realizability, and gives, with probability $1 - \delta$,

$$J(\pi^E) - J(\widehat{\pi}) \le O\Big(\min\Big\{\frac{R\log(\widetilde{B}/\delta)}{N_{\mathrm{off}}},\ \frac{\mu H \log(B_{\mathrm{bc}}/\delta)}{N_{\mathrm{int}}}\Big\}\Big), \qquad B \le \widetilde{B} \le B^H,$$

The larger class size $\widetilde{B}$ counts policies in the *step-wise completion* of the base class (their Definition 5). A completed policy concatenates base policies over time, giving $B \le \widetilde{B} \le B^H$. An each-step mixture has the same trajectory law as a first-step mixture over this larger class. After Theorem 6 in their Section 4.1, the authors write: "We conjecture that the $\log \widetilde{B}$ dependence may be sharpened to $\log B$, and leave this as an interesting open question."

The reduction to our deterministic setting is exact. One inconsistent action incurs log loss $+\infty$ and removes that base from exponential weights. Consistent bases keep equal weights. Thus the distribution is uniform on the current version space, and π^1 is precisely the per-step mixture $\mu_{\mathcal{V}}$ from Section 5, with $\mathcal{V} = \mathcal{B}_{\mathrm{bc}}$. Its worst-case expected gap is consequently governed by the offline laws.

F.2 Proof of the refutation

Proof. We verify the assumptions on the rare-context instance before computing its gap. Transitions are deterministic given the context, the deterministic base class contains the expert, and returns lie in $\{0, 1\}$, so $R = 1$. It also has $\mu = 1$: one wrong action loses at most 1 relative to the expert continuation. With $N_{\mathrm{int}} = 1$, the output is the first policy, the uniform per-step mixture over $\mathcal{B}_{\mathrm{bc}} = \{\pi^E\} \cup \{\pi_j : j \text{ unobserved}\}$.

Its conditional gap is given by (7), where K counts the unobserved rare contexts. We have $K/m \to e^{-1}$ in probability: the mean is $\mathbb{E}K = m(1-1/n)^n$, and $(1-2/n)^n \le (1-1/n)^{2n}$ makes the non-observation indicators negatively correlated, giving $\mathrm{Var}(K) \le m$. Substituting this concentration into the conditional gap gives the stated limit. Since the gap concentrates at $\Theta(1/m)$, it exceeds any $C\log(m/\delta)/m^2$ for sufficiently large m, with probability $\ge 1 - \delta$ and in expectation. The general expected rate follows from Theorem 5; its lower-bound instance also satisfies the assumptions with $R = \mu = 1$. □

F.3 Proof of the plurality guarantee

Proof. Offline term. Consider the coupled expert trajectory τ. If v_n first differs from π^E at s_t, its chosen action has at least as many votes as the expert action. Consequently, at most half of the surviving policies can vote for the expert action. At least half of $\mathcal{V}_n$ then disagrees with π^E at s_t, which gives $\mathbf{1}\{v_n \text{ deviates on } \tau\} \le \frac{2}{|\mathcal{V}_n|}\sum_{\pi\in\mathcal{V}_n}\mathbf{1}\{\tau \in D_\pi\}$. Since $\mathcal{V}_n \subseteq \mathcal{V}_0$ results from consistent pruning, the offline bound in the proof of Theorem 6 still applies. Averaging gives $\mathbb{E}[\text{deviation}] \le 2U_1(M, N_{\mathrm{off}}) \le 4\log B/N_{\mathrm{off}}$, and the gap is at most R times the deviation probability.

Online term. Let $\varepsilon_n = \mathbb{P}_{s\sim d^{v_n}}(v_n(s) \neq \pi^E(s))$ be the probability that plurality voting is wrong at the queried state. The performance-difference identity and μ-recoverability imply $J(\pi^E) - J(v_n) = H\,\mathbb{E}_{s\sim d^{v_n}}[V^{\pi^E}(s) - Q^{\pi^E}(s, v_n(s))] \leq \mu H \varepsilon_n$. Whenever the queried action is wrong, the update eliminates at least half of $\mathcal{V}_n$. Starting from $|\mathcal{V}_0| \leq B$, there can be at most $\log_2 B$ such rounds. Given the full history before round n, the queried mistake has probability ε_n. Thus $\mathbb{E}\sum_n \varepsilon_n = \mathbb{E}[\#\{n : \text{queried mistake}\}] \leq \log_2 B$. Averaging over $n \leq N_{\text{int}}$ yields $\mathbb{E}[J(\pi^E) - J(\widehat{\pi})] \leq \mu H \log_2 B / N_{\text{int}}$. □

G Scope and remaining questions

Sections 5 through 7 compare the two deployment rules for finite classes of deterministic base policies that contain the expert, with returns in $[0, 1]$ (or $[0, R]$ after scaling). Section 3 studies stochastic classes under Bayesian pointwise mixing, including tempered likelihoods and updates after batches. The upper bounds permit arbitrary MDPs and dependence among disagreement regions. The lower bounds are witnessed by an explicit MDP. The rate statements take the worst case over this class of instances. On a particular instance, the per-step gap is bounded above by $\mathbb{E}\min\{1, \sum_t a_t/q\}$ and can be much smaller, so the worst-case order alone does not determine the cost of a deployment rule in a given environment.

When most data come from interaction, the exact final-iterate law remains unresolved. The bounds leave a factor $M/A_H(M)$ for per-step mixing or $M/\log B$ for per-episode sampling.

For larger stochastic classes, the uniform order of $\mathcal{C}_M(H)$ between the two regimes of Theorem 13 is also open. At $M = H$, for example, the bounds are $H/16$ and $56H(1 + \log^\star H)$. The two potentials use different information about the process: the sum-of-towers potential does not account for the fact that at most H states are visited per round, while the entropy potential does not account for the tower.

For alternative deployments, the M-policy extensions of Theorem 2 are only partly understood. The within-episode bound $5\log(M + 1)$ extends verbatim using the potential $-\log w(\pi^\star)$. For per-episode sampling with M policies, applying the two-policy argument to each odds ratio against the expert gives an $O(M)$ bound: each sampled non-expert policy provides information on its own disagreement. The optimal dependence on class size remains open.

A further question is whether the hybrid laws extend to high probability with the $\log(1/\delta)$ dependence of Li & Zhang (2025).

Under WARM-STAGGER's single-state query model with recoverability parameter μ, plurality vote achieves $\min\{\log B/N_{\text{off}}, \mu H \log B/N_{\text{int}}\}$ in expectation. The minimax rate over all algorithms is unknown; in particular, it remains to determine whether the best algorithm must pay the μH factor.

Holding the policies and label rule fixed gives a direct empirical comparison: frozen per-step mixing accumulates more loss than either drawing one policy per episode or updating after each label. Across 39 settings with 100,000 independent runs each, simulated losses remain within 0.008 of the exact calculations. At one billion steps, the observed excess over per-episode sampling is 2.414, with a pointwise 95% confidence interval from 2.410 to 2.418. The comparison measures how deployment and update timing affect learning on this two-state instance with the candidate policies and prior unchanged. The theoretical bounds establish the iterated-logarithm worst-case order; the experiment shows the resulting gap at finite horizons.

H Experimental methods and validation

H.1 Baseline deployment comparison

We isolate deployment on the two-state construction in Proposition 9 by keeping the environment and policies fixed. The construction also supplies an exact recursion against which to check the simulated learning histories. We use the construction's expert and alternative policy, a uniform prior, and independent expert labels at every visited state. The three deployment rules are frozen per-step mixing (PS), one policy drawn per episode (PE), and per-step mixing updated after each

label (WU). Under WU, the label arrives after the current action, as required by Theorem 2. Each rule therefore learns from the states it reaches itself.

We use 13 horizons from one to one billion steps. For each horizon and method, we simulate 100,000 independent learning runs, arranged in 20 batches of 5,000 with master seed 20260925. We measure cumulative squared Hellinger loss between the expert's trajectory law and the deployed trajectory laws, using $\mathsf{h}^2(P,Q) = 1 - \int \sqrt{dP\, dQ}$. For each episode, we calculate the conditional loss analytically and simulate only the state visits that determine the posterior update. The Monte Carlo uncertainty thus comes from the learning histories; trajectory distance requires no further estimation. Figure 1 reports pointwise normal 95% confidence intervals, with standard errors computed across independent runs.

Long episodes can be simulated without enumerating uninformative states: the distribution of the active-state duration is available exactly. For PS, we draw the time spent in the active state from the construction's exact truncated geometric distribution. For PE, we first draw a policy and then its active-state duration. We simulate WU separately, action by action, and update its odds after every informative label. A run stops when its cumulative active-state visits exceed $N = \lceil \log_2 H \rceil + 48$. The omitted expected cumulative loss is at most $H2^{-N-1}$ for PS and 2^{-N} for PE and WU, each below 4×10^{-15}. We compute the exact reference values from the renewal recursion using 60-decimal interval arithmetic and the same tail bounds. Every resulting enclosure is narrower than 10^{-12}.

At a horizon of one, all three methods agree. In longer episodes, frozen PS continues to expose the learner to the wrong policy after labels have supplied evidence against it. A failure then ends the informative part of the episode, leaving fewer labels with which to update those weights. Its cumulative loss rises from 2.265 at one thousand steps to 2.770 at one billion steps. Both alternatives have loss 0.364 to three decimals at these horizons (Table 3). The simulations agree with the exact calculation: across all 39 settings, the largest absolute difference between a Monte Carlo mean and its exact reference is below 0.008.

We also compare the independent Monte Carlo means directly, propagating their sampling uncertainty to the estimated excess loss. At one thousand steps, the observed excess loss of PS over PE is 1.901, with a pointwise 95% confidence interval from 1.897 to 1.905. At one billion steps, the excess is 2.414, with an interval from 2.410 to 2.418. The corresponding excesses over WU are 1.902 (1.898 to 1.906) and 2.413 (2.409 to 2.417). These intervals use the standard error of the difference between independent Monte Carlo means. They describe sampling uncertainty on this instance and are not simultaneous confidence statements across horizons. Both alternatives have lower observed loss by much more than the simulation's sampling uncertainty.

This baseline experiment measures trajectory Hellinger loss. It does not measure reward shortfall or WARM-STAGGER's single-state query protocol. The iterated-logarithm order and horizon-independent guarantees follow from the theorems; the curves illustrate finite-horizon losses.

Episode length	Frozen PS	Per-episode	Per-label update	PS / per-episode
1	0.330757	0.330757	0.330757	1.00
1,000	2.265257	0.364348	0.364348	6.22
1,000,000	2.633959	0.364348	0.364348	7.23
1,000,000,000	2.769808	0.364348	0.364348	7.60

Table 3: Exact infinite-round cumulative trajectory Hellinger loss on the two-state instance, rounded to six decimals. Ratios use the unrounded values. Per-episode and per-label deployment have exactly equal loss on this instance; the comparison does not establish equality in other environments.

H.2 TERMINAL-SUCCESS RETURNS

The return experiment keeps the two-state MDP, policies, uniform prior, expert-label rule, and horizon grid unchanged. Its only nonzero reward is one at the final step if the entire episode avoids failure, including that step's action. The expert's return is one. Corollary 14 gives the connection to trajectory Hellinger loss, separately for each round and learning history. We evaluate return gaps directly; applying the nonlinear identity to an already summed Hellinger loss would be incorrect.

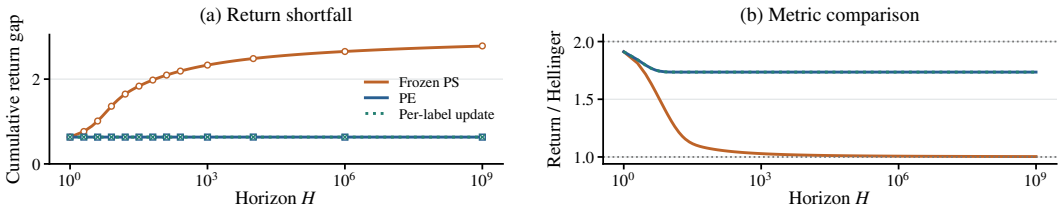

Figure 7: Terminal-success returns: (a) cumulative gaps, (b) ratio to cumulative Hellinger loss. Curves: exact; markers in (a): 100,000-run estimates with pointwise 95% intervals.

The script `code/mixing_return.py` uses master seed 20260928 and 100,000 independent learning histories for each of the 39 settings, in batches of 5,000. It samples the informative-state count under PS, the initial policy and failure time under PE, and individual actions with posterior updates under WU. At each round it integrates terminal reward conditional on the current posterior. Thus Monte Carlo uncertainty comes from learning histories, rather than extra test rollouts. The three methods use independent random streams. Their pointwise normal 95% intervals are not simultaneous confidence statements. At horizon one the count increases deterministically, so the conditional-return estimator has zero variance.

For reference values, the count-visit recursion in Appendix B.7 is evaluated with the terminal failure probability replacing Hellinger loss. At cutoff $C = \lceil \log_2 H \rceil + 48$, omitted expected return gaps are bounded by $H2^{-C-1}$ for PS and 2^{-C} for PE and WU. Every tail is below 4×10^{-15}. References use 60-decimal directed interval arithmetic; all resulting enclosures have width below 10^{-12}. Forty-two preliminary checks enumerate small-horizon action paths and four-round learning histories, validating both the transition kernels and return calculation. Across the 39 settings, the largest absolute difference between a Monte Carlo mean and its reference is 0.003687; the largest standardized difference among settings with nonzero variance is 2.083. These are pointwise checks, without seed selection.

The file `code/results/mixing_return.json` and its CSV companion record the references, estimates, intervals, and provenance. Run `code/plot_mixing_return.py` after the experiment to reproduce Figure 7. These results establish a performance separation on the specified reward and construction; the absence of a uniform cumulative return-gap bound over stochastic classes is unchanged.

H.3 Expanded study: settings and estimators

The extension script is `code/mixing_extensions.py`; its master seed is 20260926. Independent NumPy streams identify each setting, with 20,000 histories per simulated setting. We save all numerical curves and summaries in `code/results/mixing_extensions.json`, together with the generating source hash. CSV files expose the update, deterministic, and rare-context grids. The figures and tables are generated directly from these calculations.

For deterministic experiments we sample the training history and integrate the risk of a fresh deployment analytically. This removes additional randomness from test rollouts. Normal intervals use the empirical variance across independent histories. PS and PE share each sampled version space, and the allocation study reuses histories across its nine outputs. The displayed intervals are marginal, pointwise statements; we do not claim simultaneous coverage or independence of these paired estimates. Exact curves are evaluated in double precision. Their analytical truncation bounds control omitted terms, and do not certify floating-point rounding error.

H.4 Batch and tempered count recursions

The number of informative labels before an episode is a sufficient statistic for the two-state construction. For count b, exponent α, and frozen interval L, the failure probability at step $j+1$ on the surviving path is

$$p_j(b) = \frac{1}{2(1 + 2^{\alpha(b + L\lfloor j/L \rfloor)})}, \qquad 0 \le j < H.$$

Study	Parameter grid / fixed quantities	Settings
Batch updates	Horizons 64, 256, 1,024; power-of-two intervals up to each horizon	27
Likelihood exponent	Horizons 64 and 1,024; seven exponents from one to 1/64	14
Inverse-horizon exponent	Eight power-of-two horizons from 4 to 512	8
Deterministic laws	Eleven horizons, eight class sizes, eight sample sizes	27
Data allocation	Nine splits of a fixed 256-episode budget	9
Rare contexts	Seven power-of-two context counts from 2 to 128	7
Feedback comparison	Single-state or full-trajectory labels; 64 contexts	2

Table 4: Added experiment grids. Exact calculations cover every entry. Monte Carlo checks use five batch settings and three additional tempering checks, all deterministic and rare-context settings, one shared set of histories for the allocation study, and independent histories for each feedback protocol: 900,000 additional histories in total.

The probability of collecting exactly k informative labels is the product of survival probabilities through step $k-1$, times the failure probability at step k when $k < H$. At $k = H$ the remaining survival mass is used, regardless of whether the last action fails. Every collected label, including the one at the state where the unsafe action is chosen, contributes to the next episode's odds. The conditional trajectory loss is

$$g_{L,H,\alpha}(b) = 1 - \prod_{j=0}^{H-1} (1 - p_j(b))^{1/2}.$$

A forward renewal recursion sums this loss against the probability of visiting each pre-episode count. Counts increase by at least one per episode, so a count can be visited at most once. Writing $z = 2^{-\alpha}$, the uncomputed tail after count N is bounded by $Hz^{N+1}/(2(1-z))$. We choose the cutoff so that this is below 10^{-13}. All batch lengths in the plotted grid divide the horizon.

For simulation, each frozen batch draws its active-state duration from a geometric law and clips it to the batch length. Completed surviving batches update the odds; a failure ends informative visits for the remainder of the episode. This is a separate sampling implementation from the forward recursion. Its eight checks use horizon 256 with intervals 1, 4, 16, 64, and 256 at unit exponent, and frozen episodes with exponents one half, one eighth, and one thirty-second.

Interval	Exponent	Exact recursion	Monte Carlo	Pointwise 95% interval
1	1	0.364348	0.364003	[0.362755, 0.365251]
4	1	0.637162	0.639295	[0.637070, 0.641520]
16	1	1.357745	1.359850	[1.354420, 1.365281]
64	1	1.835759	1.831837	[1.823609, 1.840065]
256	1	2.095147	2.102631	[2.093630, 2.111631]
256	1/2	2.868464	2.865262	[2.853281, 2.877244]
256	1/8	6.479770	6.486776	[6.463271, 6.510282]
256	1/32	18.954656	18.936991	[18.888754, 18.985228]

Table 5: Independent checks of the update recursions at horizon 256, with 20,000 histories per row. Exact values use a count tail below one trillionth; this bound does not include floating-point rounding. Intervals describe simulation uncertainty.

H.5 Bernoulli finite sums and interaction

Conditional on a training history with K surviving non-expert policies, the number of fresh disagreements is binomial with parameters K and p. We sum over that count to evaluate PS failure; PE failure is the disagreement probability times the fraction of surviving candidates that are non-

experts. A second binomial sum over K, with survival probability $(1-p)^n$, gives the offline mean. This evaluates Equations 5 and 6 directly.

For the horizon sweep, we fix $M = 128$, $n = 1024$, and $p = 1 - (16/(2M))^{1/n}$. Only the horizon varies, from one to 1,024 in powers of two. For the class-size sweep, $H = 16$ and $n = 2048$, with M from 4 to 512 in powers of two. For the sample-size sweep, $M = 128$ and $H = 16$, with n from 64 to 8,192. In these last two sweeps we set $p = 1 - (\min\{H, M\}/(2M))^{1/n}$ separately at each setting. Both deployments always use the same resulting instance.

The allocation experiment fixes $M = 64$, $H = 16$, and $T = 256$, with $p = 1 - (16/128)^{1/256}$. Offline counts are 0, 32, 64, 96, 128, 160, 192, 224, and 255. Each full-trajectory episode reveals the initial disagreement vector, so each survivor remains independently with probability $1-p$, regardless of deployment. We simulate this transition through all 256 episodes. Averaging the conditional risks before the interactive updates implements the prescribed output rule. The potential calculation subtracts the expected final potential from its initial value using the same exact binomial survivor law. After 256 fully interactive episodes, cumulative deviations are 21.65018 for PS and 1.96420 for PE, below potential decreases of 29.05739 and 1.97175. Current-iterate gaps fall from 0.10789 to 0.05186 for PS and from 0.007965 to 0.007094 for PE. The saved results include every intermediate round.

H.6 Rare-context occupancy and single-state queries

Let K be the number of unobserved rare contexts after the offline data. At each offline observation the occupancy chain moves from K to $K-1$ with probability K/n, and otherwise stays put. We iterate this finite chain from the full set of contexts and evaluate the two conditional gaps in Equation 7. The independent simulation instead draws the complete context histogram from a multinomial law and counts the unobserved coordinates.

During a PS rollout in an unresolved context, the failure hazard is $1/(K+1)$. A uniform time query is informative precisely if it occurs no later than the first failure. Thus the probability of removing a policy in one interactive episode is

$$\frac{K}{n}\frac{1}{m}\sum_{h=1}^{m}\left(\frac{K}{K+1}\right)^{h-1}$$

under a single query, compared with K/n under full-trajectory labels. Both chains start from the same offline distribution. Independent simulations explicitly draw the unresolved context event, query time, and geometric failure time. We record pointwise confidence intervals at 0, 128, 256, and 512 extra episodes. Plurality is not evaluated on this rare-context construction: breaking ties in favor of the safe action makes it perfect. The separate Bernoulli voting experiment below gives it a nonzero initial gap.

Feedback	Extra episodes	Survivors	Exact PS gap	Estimated PS gap
One state	0	23.54	0.005310	0.005315 ± 0.000010
One state	128	23.28	0.005262	0.005268 ± 0.000010
One state	256	23.02	0.005215	0.005221 ± 0.000010
One state	512	22.52	0.005123	0.005128 ± 0.000010
Full trajectory	0	23.54	0.005310	0.005315 ± 0.000010
Full trajectory	128	22.82	0.005177	0.005180 ± 0.000010
Full trajectory	256	22.11	0.005045	0.005049 ± 0.000010
Full trajectory	512	20.77	0.004787	0.004788 ± 0.000010

Table 6: Feedback at 64 rare contexts. Expected survivors and return gaps; estimates use 20,000 histories and pointwise 95% intervals.

H.7 Numerical checks and reproduction

The extension runs 42 deterministic checks before generating data. They compare the untempered batch endpoints to the independently implemented original PS/WU recursions, enumerate small-horizon first-failure laws and conditional losses, enumerate all training and test bit patterns for a

small Bernoulli instance, and check the rare-context occupancy mean. Simulations provide additional checks of the general batch recursion, deterministic sums, allocation outputs, and both feedback kernels. We also verify at every interactive round that the expected potential decrease dominates cumulative deviation.

From the repository root, run `code/mixing_extensions.py` with Python 3 to regenerate the extension data, then `code/plot_mixing_extensions.py` for the plots and `code/export_mixing_tables.py` for the appendix feedback table. NumPy, SciPy, Matplotlib, and mpmath are the only third-party dependencies. For the original deployment study, run `code/mixing_experiment.py`, then `code/plot_mixing_deployment.py`. No data download or trained model is used.

H.8 Stochastic classes with separate informative contexts

We use the context-copy construction in the proof of Theorem 13. An episode starts at one of M uniformly chosen contexts. In context j, the expert chooses either safe action with equal probability, while candidate j puts probability one quarter on each safe action and one half on the failure action. Every other candidate agrees with the expert. Failure is absorbing, and all candidates agree there. The expert and all candidates start with equal prior weight.

Let b_j count informative labels previously observed in context j. Independent expert labels give odds $r_j = 2^{-b_j}$ against the expert. With weights frozen for the episode, the failure probability and the full-trajectory loss are

$$p_j = \frac{r_j}{2(1+\sum_i r_i)}, \qquad \ell(b_1,\dots,b_M) = \frac{1}{M}\sum_{j=1}^{M}\left[1-(1-p_j)^{H/2}\right].$$

The simulation samples the initial context and its truncated geometric active-state duration, then increments only that context's count. It evaluates the displayed loss before each update, integrating over the fresh context and trajectory. Thus only the learning history contributes sampling variance. This uses the coupled posterior over all candidates; running independent two-policy experiments would give a different process.

The grid has $M \in \{1,2,4,8,16,32\}$ and $H \in \{16,256,4096\}$, with 256 rounds and 20,000 independent histories at every setting. We report the finite-round cumulative loss without truncating a learning history or extrapolating its infinite-round tail. At horizon 256 the estimates for one and 32 non-experts are respectively 2.0931 and 36.5102, with pointwise 95% half-widths 0.0090 and 0.0614. Figure 6(b) uses horizon 256 and one, eight, or 32 non-experts. The single-context case is checked against an independent forward count recursion. For two contexts, a small finite count-vector recursion is also compared with simulation. These experiments examine the multipolicy construction at finite horizons, rather than identifying the unresolved uniform order of Theorem 13.

H.9 Plurality with genuine errors and a common query budget

The voting study uses the Bernoulli MDP of Appendix D with 64 non-experts and disagreement probability 0.6. One offline expert episode leaves each non-expert independently with probability 0.4. In a new episode, the initial state draws fresh independent disagreement bits. The expert plays action zero. Each surviving candidate plays its bit while active, and action zero after failure. Action one causes absorbing failure and loses the unit return. Plurality breaks ties by choosing action zero, yet strict wrong majorities give it initial expected gap 0.765005.

For K survivors and a fresh disagreement count $\ell \sim \mathrm{Bin}(K, 0.6)$, write $q = \ell/(K+1)$ and $v = \mathbf{1}\{2\ell > K+1\}$. Conditional deployment gaps are

$$g_{\mathrm{PS}} = 1-(1-q)^H, \qquad g_{\mathrm{PE}} = q, \qquad g_{\mathrm{vote}} = v.$$

Each rule queries exactly one state, at a uniformly chosen time in its own rollout, and updates only after the episode. A label at an active state removes the same disagreeing candidates, including when failure occurs at that state. The probabilities that this query is informative are

$$a_{\mathrm{PS}} = \frac{1}{H}\sum_{h=1}^{H}(1-q)^{h-1}, \qquad a_{\mathrm{PE}} = 1-q+\frac{q}{H}, \qquad a_{\mathrm{vote}} = 1-v+\frac{v}{H}.$$

The survivor count moves to $K - \ell$ with the corresponding probability, and otherwise stays at K. Summing over the binomial disagreement law gives a finite transition matrix. Propagating it from the offline survivor distribution gives exact expected current and averaged pre-update gaps, evaluated in double precision.

Independent simulations draw disagreement bits through their binomial count, the rollout's failure time, and an actual uniform query time. They do not sample this transition matrix. The grid uses horizons one, four, 16, and 64, with 256 queries and 20,000 histories per rule and horizon. The initial candidates, offline data law, environment, and one-query-per-episode budget are identical across rules. Each rule's deployed trajectory determines its own subsequent version space.

Horizon	PS average	PE average	Plurality average	Queried vote mistakes
1	0.008465	0.008465	0.007470	1.9124
4	0.034713	0.014987	0.016496	1.0557
16	0.140186	0.018567	0.024472	0.3916
64	0.537905	0.019747	0.028047	0.1122

Table 7: Exact averaged gaps over 256 queries and expected total queried plurality mistakes.

All twelve simulated average gaps lie within 1.75 standard errors of the corresponding recursions. Plurality improves substantially over PS in the longer episodes, while PE has a smaller average gap than plurality at horizons four, 16, and 64. A wrong plurality action at the queried state eliminates at least half the version space. The largest observed number of such mistakes in any voting history is four, below the pathwise bound $\lfloor \log_2 65 \rfloor = 6$. Query mistakes and unqueried deployment failures are different quantities; the last column counts only the former.

H.10 Coverage-study validation and execution

Run `code/mixing_coverage.py` with Python 3, then `code/plot_mixing_coverage.py`. The script uses master seed 20260927 and separate NumPy streams for every setting and rule. It saves all curves, standard errors, pointwise 95% intervals, exact voting references, source hashes, and timing metadata in `code/results/mixing_coverage.json`, with a companion summary CSV.

The 58 preliminary checks enumerate complete action trajectories for the Hellinger formula, enumerate context bits, policy draws, and query times for small voting instances, and compare finite count recursions. The two-context simulation check uses 20,000 additional histories, separate from the 600,000 histories in the main grid. Recorded execution took 14.4 seconds with peak process memory of approximately 125 MiB on an arm64 macOS machine, using Python 3.9.6, NumPy 1.26.4, and SciPy 1.13.1. These timings cover validation and data generation; figure rendering is separate.