UNDERSTANDING INFORMATION FRESHNESS IN REINFORCEMENT LEARNING UNDER DIFFUSIVE DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning relies on past interactions with the environment, but distribution shift can change how useful historical data remain for current decisions. Existing nonstationary-RL models often measure the cumulative magnitude of environmental change. We introduce *diffusive distribution shift*, a natural model for many real-world systems whose dynamics evolve continually through stochastic changes that may accumulate, cancel, or revert over time. We establish a simple, quantitative *information freshness principle*: historical data should be retained until the decision error caused by changing dynamics becomes comparable to the statistical uncertainty reduced by keeping them. This principle yields explicit expressions for how long historical data remain useful and for the unavoidable cost of tracking changing dynamics. Guided by this principle, we develop optimistic model-based RL algorithms for both known and unknown diffusion rates. We prove matching upper and lower bounds, up to logarithmic factors, for the additional regret caused by distribution shift, establishing the minimax optimality of freshness-guided learning. Moreover, without knowing the diffusion rate in advance, our adaptive algorithm achieves the same regret rate, up to logarithmic factors, as the algorithm tuned with the true rate. Controlled experiments validate the predicted freshness mechanism, and a wind-storage study illustrates its decision-level consequences.

Keywords. diffusive distribution shift; information freshness; nonstationary reinforcement learning; dynamic regret; minimax lower bounds.

1 INTRODUCTION

Reinforcement learning relies on past experience to improve future decisions. In a stationary environment, retaining more data typically reduces uncertainty about the dynamics and supports better decisions (Azar et al., 2017; Zhang et al., 2021). In real-world systems, however, the dynamics can change while the agent learns and acts. For example, robotic dynamics vary with operating conditions, user behavior shifts in interactive systems, and demand and supply patterns change in energy systems (Nagabandi et al., 2019; Huleihel et al., 2021; Dalal et al., 2016). Such changes are often not monotone or predictable in advance, and they may persist throughout learning (Slivkins and Upfal, 2008; Lecarpentier and Rachelson, 2019; Feng et al., 2022; Chen et al., 2023). A historical trajectory may therefore remain informative even when it no longer accurately represents the current dynamics (Chandak et al., 2020; Cheung et al., 2020; Xie et al., 2021). Hence, the agent must balance the statistical benefit of retaining more historical data against their mismatch with the current dynamics. We call this the *information freshness problem* in reinforcement learning: how much history should an agent retain when making decisions in the current environment?

Freshness is not determined by age alone. Using a short history limits exposure to outdated dynamics but may leave too few transition observations to estimate important transitions. A longer history provides more observations, yet it combines data collected under different past kernels. Many nonstationary-RL formulations summarize change through switching counts or cumulative variation (Ortner et al., 2020; Cheung et al., 2020; Domingues et al., 2021; Mao et al., 2021; Wei and Luo, 2021; Feng et al., 2023). These formulations accommodate general nonstationarity, but they do not directly quantify how well historical data match the current dynamics. In particular, cumulative

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.

variation adds every local movement along the kernel path. If the kernel moves away and later returns, both movements are counted even though data collected before the excursion may still describe the current kernel well. Exploiting this structure in RL is challenging: the agent observes transitions generated under its policies rather than the kernels themselves, and estimation errors can propagate through multi-step planning (Kearns and Singh, 2002; Azar et al., 2017). How much history to use therefore depends jointly on the available statistical information, the magnitude and persistence of environmental changes, and the planning horizon.

To address this problem, we develop a theory for episodic tabular RL under *diffusive distribution shift* in the transition dynamics. Our information freshness principle links the amount of history used for learning to the achievable dynamic regret: the useful memory scale balances statistical uncertainty against value mismatch with the current dynamics. We design algorithms that attain this balance and prove the regret upper bound and a matching lower bound to show its minimax optimality.

Our contributions are summarized as follows.

Information freshness under diffusive distribution shift. We introduce diffusive distribution shift to model persistent stochastic changes that may accumulate, cancel, or revert over time. We prove an information freshness principle: the useful lifetime of historical data is determined by balancing statistical uncertainty against mismatch with the current dynamics. In the main regime, the optimal lifetime satisfies $W_\nu^\star \asymp \sqrt{|\mathcal{S}||\mathcal{A}|}/(H\sqrt{\nu})$. Thus data become stale sooner as the diffusion rate ν or planning horizon H increases, whereas a larger state–action space $|\mathcal{S}||\mathcal{A}|$ requires more history.

Freshness-guided learning algorithms. Guided by the information freshness principle, we develop optimistic model-based algorithms that retain only the history expected to remain useful. When the diffusion rate is known, FRESH-RESTART periodically discards old data according to the predicted information lifetime. When the rate is unknown, ADA-FRESH uses observed trajectories to decide online when older data should be discarded, without estimating the diffusion rate directly.

Minimax-optimal regret guarantees. When the diffusion rate is known, we establish matching upper and lower bounds, up to logarithmic factors, for the dynamic-regret cost of distribution shift. When the rate is unknown, ADA-FRESH achieves the same regret rate without prior knowledge of ν. Together, these results show that the information freshness principle captures the fundamental cost of distribution shift and that our freshness-guided algorithms indeed attain the optimality.

Controlled experiments support the predicted information lifetime and regret rates, while an OPSD-based wind-storage study illustrates the method in a realistic control problem.

2 MDPs UNDER DIFFUSIVE DISTRIBUTION SHIFT

In this section, we introduce the model of episodic MDPs under diffusive distribution shift. Then we define the corresponding policy-learning objective.

2.1 MODEL

We consider episodic MDPs $M_k = (\mathcal{S}, \mathcal{A}, H, P_k, \{r_h\}_{h=1}^H, \rho)$ over episodes $k \in [K]$. The state and action spaces $\mathcal{S}$ and $\mathcal{A}$ are finite, H is the horizon, and ρ is the known initial-state distribution. The transition kernel P_k is fixed within episode k but may change between episodes. Within episode k, the transition kernel $P_k(\cdot \mid s, a)$ is the same at every stage h, so all visits to (s, a) provide observations of the same transition distribution. Across episodes, only P_k changes. Each reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is deterministic and fixed across episodes, and its value is observed when the corresponding state–action pair is visited. The initial-state distribution ρ is fixed and known.

Let $\mathcal{E}_k = \sigma(P_1, \ldots, P_k)$ be the history of transition kernels. The kernel evolving process may depend on its own history. We assume that it is exogenous: conditional on $\mathcal{E}_k$, future kernel changes are independent of the agent's policy and the trajectories. We impose the following condition on the distribution shift.

Assumption 2.1 (Diffusive distribution shift)**.** *There exists a diffusion rate $\nu \in [0, 1]$ such that, for every $k, \ell \geq 1$ with $k + \ell \leq K$,*

$$\mathbb{E}\Big[\|P_{k+\ell} - P_k\|_{\infty,\mathrm{TV}}^2 \mid \mathcal{E}_k\Big] \leq \min\{1, \nu\ell\}.$$

Here, $\|P - P'\|_{\infty,\mathrm{TV}} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mathrm{TV}}(P(\cdot \mid s, a), P'(\cdot \mid s, a))$ is the largest TV distance across all state–action pairs.

The assumption bounds the maximum TV distance between kernels ℓ episodes apart. Thus, ν measures how quickly historical transition data can become stale, while $\nu = 0$ recovers a stationary MDP. It covers a broad family of persistent stochastic shifts, including latent random walks, discretized diffusions, mean-reverting dynamics, and stochastic refreshes (Slivkins and Upfal, 2008; Choi et al., 1999; Lecarpentier and Rachelson, 2019; Xie et al., 2021; Chen et al., 2023; Feng et al., 2022). Representative examples are verified in sections J and L.

To illustrate how the diffusion rate ν relates to an underlying evolving system, consider a scalar latent process $\theta_{k+1} = \theta_k + \sigma\xi_{k+1}$ with $P_k = P_{\theta_k}$, where the innovations are independent, have mean zero, and have second moment at most one. If $\|P_\theta - P_{\theta'}\|_{\infty,\mathrm{TV}} \leq L|\theta - \theta'|$, then Assumption 2.1 holds with $\nu = (L^2\sigma^2) \wedge 1$. Thus, ν combines the speed σ of the latent evolution with the sensitivity L of the transition dynamics to that evolution.

Relation to standard nonstationarity models. Standard switching-count and variation-budget models measure how often or how much an MDP changes along its realized path (Ortner et al., 2020; Cheung et al., 2020; Domingues et al., 2021; Mao et al., 2021; Wei and Luo, 2021; Feng et al., 2023). Our condition instead bounds how far a past kernel can be from the current one. This matters when stochastic changes cancel or reverse: cumulative variation may be large even though the past and current kernels are close. By imposing this bound across any number of episodes, the model captures how stochastic changes in the transition dynamics spread over time.

2.2 Policy Learning Problem

Before each episode, the agent chooses a policy using only previously observed trajectories. Let Π denote the class of stage-dependent Markov policies $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$ for every state and stage. Under P_k, the value of policy π from state s at stage h is

$$V_{k,h}^\pi(s) := \mathbb{E}_{P_k,\pi}\left[\sum\nolimits_{t=h}^H r_t(s_t, a_t) \,\middle|\, s_h = s\right].$$

We write $V_{k,1}^\pi(\rho) := \mathbb{E}_{s_1\sim\rho}[V_{k,1}^\pi(s_1)]$ for its value at the start of an episode.

For each episode k, let $\pi_k^\star \in \arg\max_{\pi\in\Pi} V_{k,1}^\pi(\rho)$ be an optimal policy as a benchmark, and write $V_{k,1}^\star(\rho) := V_{k,1}^{\pi_k^\star}(\rho)$. The policy $\pi_k^\star$ is chosen with knowledge of the current kernel P_k and may change across episodes. In contrast, before episode k, the agent selects π_k using only trajectories collected in episodes $1, \ldots, k-1$, without observing P_k. We define the dynamic regret as

$$\mathrm{Reg}_K := \sum\nolimits_{k=1}^K \left[V_{k,1}^\star(\rho) - V_{k,1}^{\pi_k}(\rho)\right].$$

Because $\pi_k^\star$ is optimal for the current episode, Reg_K measures the cost of learning and tracking the evolving MDP under distribution shift. Our goal is to design algorithms that minimize $\mathbb{E}[\mathrm{Reg}_K]$ uniformly over all kernel processes satisfying Assumption 2.1, both when the diffusion rate ν is known and when it is unknown to the agent. The expectation is over the kernel process, sampled trajectories, and algorithm randomization.

3 The Information Freshness Principle

In this section, we first study how retaining more history reduces statistical uncertainty while increasing the mismatch between past data and the current dynamics. We then derive the information freshness principle by balancing these two effects.

3.1 Two Costs of Memory

Suppose the agent retains trajectories from the most recent W episodes. We quantify the statistical benefit and the kernel mismatch associated with this choice.

Statistical learning. To isolate the benefit of retaining more data, we consider the idealized case in which the transition kernel is fixed within a W-episode period. Standard optimistic tabular-RL algorithms incur $\widetilde{O}(H\sqrt{|\mathcal{S}||\mathcal{A}|W})$ regret over this period (Azar et al., 2017; Zhang et al., 2021). Hence, ignoring logarithmic factors, the average statistical cost per episode is

$$\mathsf{Stat}(W) := H\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{W}}.$$

This statistical learning cost decreases with W because a longer history provides more transition observations and can improve the policy performance.

Kernel mismatch. We next quantify the policy-value mismatch caused by distribution shift. Data collected in episode k are generated by P_k, but may be reused ℓ episodes later when the current kernel is $P_{k+\ell}$. We first bound how this kernel displacement changes policy values, and then use the bound to derive the mismatch cost of a W-episode period.

Lemma 3.1 (Value mismatch under diffusive distribution shift)**.** *Under Assumption 2.1, for every $k, \ell \geq 1$ with $k + \ell \leq K$,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|V^{\pi}_{k+\ell,1}(\rho) - V^{\pi}_{k,1}(\rho)\right| \mid \mathcal{E}_k\right] \leq \frac{H(H-1)}{2}\sqrt{\min\{1, \nu\ell\}}.$$

The factor $H(H-1)/2$ sums the remaining-horizon sensitivity over the $H-1$ transitions: a mismatch at stage h can affect up to $H-h$ subsequent rewards. The proof is given in section D. In a W-episode memory, every retained trajectory is at most $W-1$ episodes old. The lemma therefore gives the mismatch cost

$$\mathsf{Shift}_\nu(W) := H^2\sqrt{\nu W}.$$

Since the total reward in one episode is at most H, this mismatch is also at most H.

3.2 The Information Lifetime

We balance the statistical and mismatch costs through the following heuristic objective

$$\mathcal{J}_\nu(W) := \mathsf{Stat}(W) + \mathsf{Shift}_\nu(W) = H\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{W}} + H^2\sqrt{\nu W}.$$

The first statistical cost term decreases with W, while the second mismatch cost term increases with W. Their balance determines how much history should be retained.

Proposition 3.2 (Information freshness principle)**.** *For $\nu > 0$, the freshness objective $\mathcal{J}_\nu(W)$ is minimized over $W > 0$ at*

$$\overline{W}_\nu = \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{H\sqrt{\nu}}.$$

Restricting the memory length to the available range $[1, K]$ gives the optimal information lifetime $W_\nu = \min\{K, \max\{1, \overline{W}_\nu\}\}$. When $\nu = 0$, $W_0 = K$.

What determines information freshness. The information lifetime is not determined by the environment alone. For the same diffusion rate ν, a longer horizon H makes old data less useful because a transition mismatch can affect more future rewards. A larger state–action space $|\mathcal{S}||\mathcal{A}|$ has the opposite effect: more history is needed before the transition model can be estimated accurately. Information freshness is therefore a joint property of environmental evolution, statistical complexity, and the decision horizon.

When $1 < \overline{W}_\nu < K$, the statistical and mismatch costs are equal at the information lifetime. Therefore, the optimal objective satisfies

$$\mathcal{J}_\nu(\overline{W}_\nu) = 2H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4}.$$

Algorithm 1 FRESH-RESTART(W): learning with a fixed information lifetime

Require: Block length W, confidence level δ
Ensure: Policies $\pi_1, \ldots, \pi_K$
Partition $[K]$ into consecutive blocks of length at most W.
for each block B **do**
  Initialize a fresh MVP copy with empty row datasets $\{\mathcal{D}(s,a)\}_{s,a}$.
  for each global episode $k \in B$, with local index $q = 1, 2, \ldots$ **do**
    for each (s,a) with $N_{q-1}(s,a) := |\mathcal{D}(s,a)| > 0$ **do**
      Set $n_q(s,a) = 2^{\lfloor \log_2 N_{q-1}(s,a) \rfloor}$ and form $\widehat{P}_q(\cdot \mid s,a)$ from the first $n_q(s,a)$ samples in $\mathcal{D}(s,a)$.
    end for
    Compute $(\widetilde{Q}_{q,h}, \widetilde{V}_{q,h})_{h=1}^{H}$ backward using the optimistic MVP update in Appendix G, Eq. (1); use value one for an unseen row.
    Set $\pi_{k,h}(s) \in \arg\max_a \widetilde{Q}_{q,h}(s,a)$ and execute π_k.
    Append every observed transition $(s_{k,h}, a_{k,h}, s_{k,h+1})$ to the corresponding row dataset $\mathcal{D}(s_{k,h}, a_{k,h})$.
  end for
end for

Thus, with the other parameters fixed, the predicted per-episode cost scales as $\nu^{1/4}$. With different number of episodes K, the clipped lifetime also separates three regimes. When $\overline{W}_\nu \geq K$, all available history remains useful. When $1 < \overline{W}_\nu < K$, statistical uncertainty and kernel mismatch balance in the interior. When $\overline{W}_\nu \leq 1$, the model will get no benefit from retaining data across multiple episodes, and the per-episode loss is capped by H.

Note that, this calculation is exact for the heuristic freshness objective $\mathcal{J}_\nu$, but it does not yet establish an achievable or unavoidable regret rate. The following sections make the prediction rigorous. Specifically, we show that algorithms guided by this principle achieve the minimax-optimal regret rate, up to logarithmic factors, whether or not ν is known.

4 LEARNING WITH A KNOWN DIFFUSION RATE

In this section, we develop an optimistic learning algorithm for a known diffusion rate ν. We then prove its shift-dependent regret bound and a matching lower bound.

4.1 ALGORITHM DESIGN

When ν is known, the agent can directly compute the information lifetime W_ν from Proposition 3.2. Our algorithm, FRESH-RESTART, uses W_ν as its restart period. It runs a fresh copy of MVP for W_ν episodes, discards the accumulated data, and repeats. Thus the algorithm uses all observations within the current information lifetime but does not retain older trajectories.

MVP base learner. Monotonic Value Propagation (MVP) is an optimistic model-based RL algorithm (Zhang et al., 2021). It estimates each transition distribution from observed next states, constructs Bernstein bonuses for statistical uncertainty, and computes an optimistic policy by backward value propagation. The agent follows this policy, adds the resulting trajectory to the empirical model, and updates its value estimates as more data become available.

Freshness-guided restarts. FRESH-RESTART divides the K episodes into consecutive blocks of length W_ν. At the beginning of each block, it initializes a new MVP instance with empty transition counts. Within the block, MVP runs without modification. The restart prevents trajectories older than the information lifetime from influencing the current policy. Choosing W_ν balances the statistical benefit of retaining data against their growing mismatch with the current dynamics. Although the kernel continues to evolve within each block, our analysis accounts for the resulting mismatch in the regret bound.

4.2 Regret Guarantees and Minimax Optimality

We now state the regret guarantee obtained at the information lifetime W_ν and compare it with the minimax lower bound. A stronger bound for an arbitrary restart period is given in section H.

Theorem 4.1 (Known-rate regret upper bound)**.** *Under Assumption 2.1,* Fresh-Restart *with* $W = W_\nu$ *satisfies*

$$\mathbb{E}[\mathrm{Reg}_K] \le \widetilde{O}\left(H \min\{K, \sqrt{|\mathcal{S}||\mathcal{A}|K}\} + K \min\{H, H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4}\} + \frac{H|\mathcal{S}|^2|\mathcal{A}|K}{W_\nu} \right) \wedge HK.$$

The regret can be divided into the following three cost terms:

Cost of MDP Learning. The first term is present even in a stationary environment. When $\nu = 0$, we have $W_0 = K$, so Fresh-Restart runs a single MVP copy and recovers the standard $\widetilde{O}(H\sqrt{|\mathcal{S}||\mathcal{A}|K})$ dependence, together with the trivial HK cap. This matches the minimax dependence for stationary tabular episodic RL up to logarithmic factors (Azar et al., 2017) and is the leading term in the stationary MVP guarantee under our reward scale (Zhang et al., 2021). Thus the model introduces no additional learning cost when the distribution shift vanishes.

Cost of distribution shift. The second term is the additional price of making decisions under continually changing dynamics. In the interior regime, its per-episode contribution is $\widetilde{O}(H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4})$, exactly as predicted by the information freshness principle. Existing nonstationary-RL results measure change by cumulative variation, which adds the movement at every episode (Cheung et al., 2020; Mao et al., 2021). Our model instead controls how far two kernels can be after a given number of episodes. This distinction matters under stochastic evolution because successive changes may cancel or revert. It leads to a different regret law, whose dependence on the diffusion rate is $\nu^{1/4}$.

Cost of restarting MVP. The third term repeats the initialization cost of the base learner over roughly K/W_ν blocks. A single stationary MVP run has an additive $\widetilde{O}(H|\mathcal{S}|^2|\mathcal{A}|)$ term under our reward scale (Zhang et al., 2021). Setting $\nu = 0$ gives $W_\nu = K$ and recovers this stationary burn-in, so the unmatched term is inherited from the stationary base rather than caused by distribution shift. Eliminating polynomial burn-in while retaining a fully horizon-independent guarantee remains an open question in stationary time-homogeneous RL (Zhou et al., 2026). We therefore do not interpret this term as a fundamental cost of diffusive distribution shift.

Theorem 4.2 (Minimax lower bound)**.** *There is a universal constant* $c > 0$ *such that, for all* $|\mathcal{S}|, |\mathcal{A}| \ge 2$*,* $H \ge 2$*,* $K \ge 1$*, and* $\nu \in [0, 1]$*, every agent, even one given the diffusion rate* ν*, suffers minimax expected dynamic regret at least*

$$c\left[H \min\{K, \sqrt{|\mathcal{S}||\mathcal{A}|K}\} + K \min\{H, H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4}\} \right]$$

over MDPs satisfying Assumption 2.1. The same lower bound holds for a hard family in which every state–action pair has at most two possible next states.

Discussion. The lower bound matches the first two terms of the upper bound up to logarithmic factors: the cost of learning the MDP and the additional cost of distribution shift. The only unmatched term is $\widetilde{O}(H|\mathcal{S}|^2|\mathcal{A}|K/W_\nu)$, which comes from repeatedly initializing the stationary MVP base algorithm (Zhang et al., 2021). Thus the distribution-shift rate, including its dependence on ν, H, $|\mathcal{S}|$, and $|\mathcal{A}|$, is minimax optimal for unrestricted transition kernels. Under the bounded-successor condition introduced below in Corollary 4.3, the initialization term is absorbed and the complete upper and lower bounds match.

Lower-bound intuition. The hard instance is a survival MDP in which a suboptimal action increases the probability of entering an absorbing state with no further reward. At random refresh times, the optimal action is redrawn, so the agent must identify it again from newly collected trajectories. The hardest refresh frequency leaves the agent with a constant probability of choosing a suboptimal action while satisfying the diffusion constraint. A mistake at stage h causes a value loss of order $\Delta(H - h)$,

giving a total loss of order $H^2\Delta$ per episode. Balancing the available information with the diffusion constraint yields

$$\Delta \asymp \left(\frac{|\mathcal{S}||\mathcal{A}|\nu}{H^2}\right)^{1/4},$$

and hence the lower-bound rate $H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4}$ per episode. This matches the dependence predicted by the information freshness principle.

Corollary 4.3 (Complete matching upper bound with bounded successors)**.** *Suppose every state–action pair has at most a fixed number d_0 of possible next states across all episodes almost surely. Then* FRESH-RESTART *at $W = W_\nu$ satisfies*

$$\mathbb{E}[\mathrm{Reg}_K] \le \widetilde{O}_{d_0}\Big(H\min\{K, \sqrt{|\mathcal{S}||\mathcal{A}|K}\} + K\min\{H, H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4}\}\Big) \wedge HK.$$

The agent need not know the successor sets.

Therefore, under bounded transition support, the complete finite-sample bounds match up to logarithmic factors.

5 Adaptive Learning with an Unknown Diffusion Rate

When the diffusion rate is unknown, the agent cannot compute the information lifetime W_ν. ADA-FRESH instead adapts its memory directly from observed trajectories. It compares predictions based on different amounts of history and restarts when long-memory predictions are no longer supported by recent returns. We first describe this mechanism and then show that it matches the known-rate regret guarantee up to logarithmic factors.

5.1 Adaptive Algorithm Design

Rather than estimate ν, ADA-FRESH maintains MVP instances that use different amounts of history. Long-running instances learn from more data, while shorter instances periodically focus on recent trajectories. If recent evidence shows that long-history predictions are no longer reliable, the algorithm discards the old data and restarts. Its memory therefore grows during stable periods and resets when historical data become stale.

The main difficulty is that different memory lengths cannot be evaluated simultaneously: each MVP instance proposes a policy, and its performance is observed only when that policy is executed. ADA-FRESH uses the multi-scale schedule and two consistency tests of MASTER (Wei and Luo, 2021, Algorithm 3) to decide when short instances should probe the current environment. The tests restart the algorithm when recent performance contradicts predictions based on a longer history. Algorithm 2 summarizes the procedure. The exact schedule and test thresholds are given in section K.

5.2 Adaptive Regret and Minimax Optimality

We compare unknown-rate ADA-FRESH with FRESH-RESTART tuned to the oracle lifetime W_ν.

Theorem 5.1 (Adaptive regret without knowing the diffusion rate)**.** *For every $\nu \in [0, 1]$, under Assumption 2.1,* ADA-FRESH *runs without knowing ν and satisfies*

$$\mathbb{E}[\mathrm{Reg}_K] \le \widetilde{O}\left(H\min\{K, \sqrt{|\mathcal{S}||\mathcal{A}|K}\} + K\min\{H, H^{3/2}(|\mathcal{S}||\mathcal{A}|)^{1/4}\nu^{1/4}\} + \frac{H|\mathcal{S}|^2|\mathcal{A}|K}{W_\nu}\right) \wedge HK.$$

Under Corollary 4.3, the final term is absent, up to constants depending on d_0.

No cost of adaptation. The bound matches Theorem 4.1, while Theorem 4.2 already grants ν. Thus adaptation adds no new term and the distribution-shift rate remains minimax optimal. Bounded successors give complete matching; unrestricted kernels retain only the same MVP initialization gap as in the known-rate case. Importantly, adaptation does not require estimating ν itself. ADA-FRESH only needs to select a memory scale whose statistical and mismatch costs are comparable to those at the oracle lifetime W_ν. Thus the diffusion rate may remain unknown even though its effect on decision quality is controlled at the minimax rate.

Algorithm 2 ADA-FRESH: adaptive memory selection

Require: MVP base algorithm, confidence level δ
Initialize the MASTER schedule and consistency tests as in section K.
Set the next episode $k \leftarrow 1$.
while $k \leq K$ **do**
  Start a new epoch at k; set $n \leftarrow 0$ and restart $\leftarrow$ false.
  while $k \leq K$ and restart $=$ false **do**
    Set $N \leftarrow 2^n$, $s \leftarrow k$, and $e \leftarrow \min\{s + N - 1, K\}$.
    Launch dyadic MVP copies according to the MASTER schedule in section K.
    for $t = s, \ldots, e$ **do**
      Run and update the launched copy of shortest length whose cell contains t.
      if either consistency test in equation (2) fires **then**
        Discard all copies; set $k \leftarrow t + 1$ and restart $\leftarrow$ true; **break**.
      end if
    end for
    if restart $=$ false **then**
      Set $k \leftarrow e + 1$ and $n \leftarrow n + 1$.
    end if
  end while
end while

6 EMPIRICAL VALIDATION

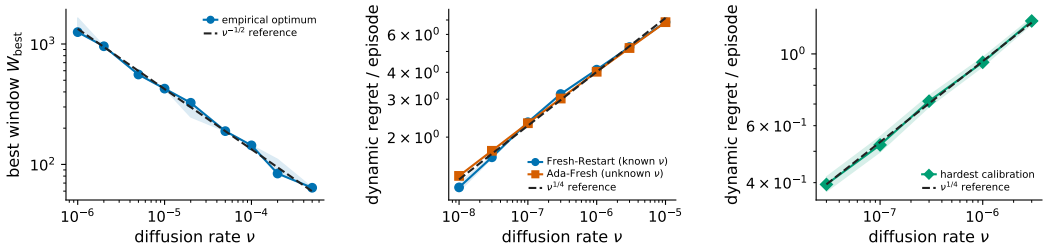
This section presents two complementary experiments. The first uses controlled evolving MDPs to test the predicted information lifetime, the regret rates of the analyzed algorithms, and the hard-instance scaling behind the lower bound. The second studies wind-storage control with real energy data to examine how finite memory affects policy regret and operating cost in a realistic decision problem.

Rate validation. We test the three rate predictions that drive the theory. A smooth latent diffusion is used to measure how the empirically best memory length changes with ν. We then run the analyzed FRESH-RESTART and ADA-FRESH algorithms on a recurrent survival MDP. Finally, a revealed-refresh testing problem isolates the hard family used in the lower bound. Across these experiments, we vary ν over prespecified grids while holding the other problem parameters fixed. The first panel selects the best fixed memory from a common candidate grid, while the algorithm panel reports full-run dynamic regret without removing startup episodes. All constants are fixed on separate development seeds before evaluation, and all reported curves use held-out seeds. Section M gives the environments, implementation details, and additional diagnostics.

The three log–log plots in figure 1 give a consistent picture. The fitted information-lifetime exponent is -0.491, close to the predicted $-1/2$. The regret exponents are 0.255 for FRESH-RESTART and 0.238 for ADA-FRESH, compared with the predicted $1/4$. The hardest lower-bound instances have exponent 0.253, with 95% bootstrap interval $[0.241, 0.263]$. Thus the lifetime, achievable regret, and hard-instance regret follow the rates predicted by the information freshness principle.

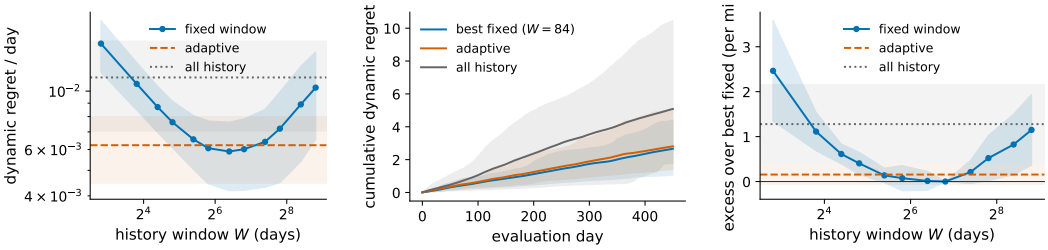
Wind-storage control with real energy data. We finally study day-ahead battery operation, where retaining too much history can make a controller slow to respond to changing wind, demand, and price conditions. We use 900 days of hourly data from the Denmark DK1 region in Open Power System Data (Open Power System Data, 2020). Each episode covers 24 hours. The MDP has 72 states and five charge or discharge actions. Its engineering model includes charge and discharge efficiency, self-discharge, finite power and energy, grid import and export, degradation, reliability and curtailment penalties, and a terminal reserve target. The agent knows these physical constraints and costs but must estimate the changing wind–load–price transition kernel from observed trajectories.

The experiment combines empirical operating regimes with controlled distribution shift. We estimate surplus and scarcity kernels from the lowest and highest 15% of historical scarcity scores. These endpoint kernels come from observed data; the battery model uses fixed engineering parameters. We then connect the endpoints through a smooth latent diffusion so that the displacement condition can



(a) Information lifetime. (b) Regret of the two algorithms. (c) Regret on the hard family.

Figure 1: Rate validation under controlled diffusive distribution shift. All panels use log–log axes; shaded bands are 95% bootstrap intervals and dashed lines show the predicted $\nu^{-1/2}$ or $\nu^{1/4}$ dependence.



(a) Freshness curve. (b) Cumulative control regret. (c) Realized operating cost.

Figure 2: Information freshness in OPSD-based wind-storage control. Empirical surplus and scarcity kernels are estimated from Denmark DK1 wind, load, and price data; only the smooth diffusive path between those kernels is simulated. Curves and 95% intervals are computed over ten held-out diffusion paths; one separate path was used to fix the adaptive constant before evaluation. The full system specification and displacement-bound checks appear in section M.2.

be checked. Only this path between empirical regimes is simulated. The resulting process satisfies the condition with $\nu = 1.52{\cdot}10^{-3}$, providing a setting in which the predicted lifetime can be compared with realized control performance.

Over ten held-out paths, the fixed-window curve has the predicted interior optimum: very short windows relearn too often, while all-history learning reacts too slowly as the kernel changes. The adaptive selector is within a factor 1.056 of the best fixed window, with 95% interval $[0.976, 1.178]$, despite not receiving ν. Relative to all-history learning, it reduces dynamic regret by 44.7% (95% interval $[20.7\%, 61.4\%]$) and removes 87.8% of the corresponding operating-cost gap. The study shows that the freshness tradeoff remains visible in a realistic control model and affects both policy regret and operational cost. Together with the controlled experiments above, it connects the theorem's mechanisms to an applied decision problem.

7 Conclusion

We studied how long historical data remain useful in RL under diffusive distribution shift. The information freshness principle identifies an optimal lifetime by balancing statistical uncertainty against mismatch with the current dynamics. Fresh-Restart attains the resulting rate when the diffusion rate is known, while Ada-Fresh achieves the same rate without knowing it. Matching lower bounds establish the minimax optimality of the distribution-shift cost. These results identify information lifetime as a fundamental quantity for learning in continually changing environments. This lifetime depends jointly on the rate of environmental evolution, the statistical complexity of the MDP, and the planning horizon. It therefore provides a concrete basis for deciding when historical data should be retained or discarded in an evolving decision problem.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, experimental code, mathematical checking, LaTeX editing, and prose revision. The authors are responsible for reviewing all AI-assisted text, mathematical statements, proofs, and code.

REFERENCES

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:263–272, 2017.

Chen, L. and Luo, H. Near-optimal goal-oriented reinforcement learning in non-stationary environments. *Advances in Neural Information Processing Systems*, 35:33973–33984, 2022.

Chen, Q., Golrezaei, N., and Bouneffouf, D. Non-stationary bandits with auto-regressive temporal dependency. *Advances in Neural Information Processing Systems*, 36:7895–7929, 2023.

Chandak, Y., Theocharous, G., Shankar, S., White, M., Mahadevan, S., and Thomas, P. S. Optimizing for the future in non-stationary MDPs. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:1414–1425, 2020.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:1843–1854, 2020.

Choi, S. P. M., Yeung, D.-Y., and Zhang, N. L. An environment model for nonstationary reinforcement learning. *Advances in Neural Information Processing Systems*, 12, 1999.

Dalal, G., Gilboa, E., and Mannor, S. Hierarchical decision making in electricity grid management. In *Proceedings of the 33rd International Conference on Machine Learning*, PMLR 48:2197–2206, 2016.

Darwiche Domingues, O., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, PMLR 130:3538–3546, 2021.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33, 2020.

Feng, F., Huang, B., Zhang, K., and Magliacane, S. Factored adaptation for non-stationary reinforcement learning. *Advances in Neural Information Processing Systems*, 35:31957–31971, 2022.

Feng, S., Yin, M., Huang, R., Wang, Y.-X., Yang, J., and Liang, Y. Non-stationary reinforcement learning under general function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:9976–10007, 2023.

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

Gerogiannis, A., Huang, Y.-H., and Veeravalli, V. Is prior-free black-box non-stationary reinforcement learning feasible? In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, PMLR 258:2692–2700, 2025.

Huleihel, W., Pal, S., and Shayevitz, O. Learning user preferences in non-stationary environments. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, PMLR 130:1432–1440, 2021.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002. doi:10.1023/A:1017984413808.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Lecarpentier, E. and Rachelson, E. Non-stationary Markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Basar, T. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:7447–7458, 2021.

Maurer, A. and Pontil, M. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Nagabandi, A., Clavera, I., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., and Finn, C. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2019.

Ortner, R., Gajane, P., and Auer, P. Variational regret bounds for reinforcement learning. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, PMLR 115:81–90, 2020.

Open Power System Data. *Data Package Time Series*. Version 2020-10-06, 2020. doi:10.25832/time_series/2020-10-06.

Slivkins, A. and Upfal, E. Adapting to a changing environment: The Brownian restless bandits. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 343–354, 2008.

Wei, C.-Y. and Luo, H. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of Thirty Fourth Conference on Learning Theory*, PMLR 134:4300–4354, 2021.

Xie, A., Harrison, J., and Finn, C. Deep reinforcement learning amidst continual structured non-stationarity. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:11393–11403, 2021.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? A near-optimal algorithm escaping the curse of horizon. In *Proceedings of Thirty Fourth Conference on Learning Theory*, PMLR 134:4528–4531, 2021. Full version: arXiv:2009.13503.

Zhang, Z., Ji, X., and Du, S. Horizon-free reinforcement learning in polynomial time: The power of stationary policies. In *Proceedings of Thirty Fifth Conference on Learning Theory*, PMLR 178:3858–3904, 2022.

Zhao, P., Li, L.-F., and Zhou, Z.-H. Dynamic regret of online Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:26865–26894, 2022.

Zhou, R., Zhang, Z., and Du, S. S. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:42878–42914, 2023.

Zhou, R., Zhang, Z., Fazel, M., and Du, S. S. Asymptotically optimal regret for reinforcement learning without horizon dependence. *arXiv preprint arXiv:2607.19854*, 2026.

A Appendix Organization

The appendix begins with related work and is then organized by proof dependency. We first collect two results from earlier work and define the histories and high-probability events used later. We then prove the freshness result, the interval-diameter bound, and the MVP guarantee for data from changing kernels. These facts give the known-ν upper bound and its bounded-support refinement. The lower bound follows one hard family from its construction to the final regret calculation. The unknown-ν proof comes next and adds the multiscale replay argument. The last sections give latent-process examples and experimental details.

Both upper bounds use the same two facts. The interval-diameter result controls how much the kernel can change over the data being reused. The changing-kernel MVP result then turns this kernel change into regret. The lower bound uses a separate argument.

B Related Work

This section places our model, freshness principle, and algorithms within four related lines of work: stationary episodic RL, nonstationary RL, adaptive data reuse, and stochastic dynamics models.

Stationary episodic RL. Finite-horizon tabular RL provides the stationary benchmark for our problem. Existing work gives sharp guarantees for model-based optimism, model-free Q-learning, horizon-free learning, and variance-dependent methods (Azar et al., 2017; Jin et al., 2018; Zhang et al., 2021; 2022; Zhou et al., 2023). When $\nu = 0$, our setting reduces to this stationary problem, and our algorithms use the MVP planner and bonuses of Zhang et al. (2021). The main difference is that, when $\nu > 0$, trajectories retained by the agent may have been generated by different kernels, while performance is evaluated under the current kernel. We extend the stationary analysis to separate ordinary estimation error from this kernel mismatch. The support-dependent guarantee of MVP-V (Zhou et al., 2023, Theorem 7) also motivates our bounded-successor refinement, under which the complete finite-sample upper and lower bounds match.

Nonstationary reinforcement learning. Nonstationary RL commonly measures environmental change by a switching count or a cumulative variation budget. This framework has been studied for continuing MDPs (Ortner et al., 2020; Cheung et al., 2020), episodic MDPs (Domingues et al., 2021; Mao et al., 2021), stochastic shortest paths (Chen and Luo, 2022), and function approximation (Feng et al., 2023); related online-MDP work allows changing rewards with fixed or known transitions (Fei et al., 2020; Zhao et al., 2022). Two close rate comparisons are the sliding-window guarantee of Cheung et al. (2020), with $(B_r + B_p)^{1/4}T^{3/4}$-type dependence in continuing MDPs, and the model-free guarantee of Mao et al. (2021), with $\Delta^{1/3}T^{2/3}$-type dependence for arbitrary episodic variation. These models charge the changes accumulated along the realized path. Our condition instead controls the expected displacement between a historical kernel and the current kernel across every episode gap. This distinction matters when stochastic movements cancel or reverse: cumulative variation can be large even when an earlier kernel remains close to the current one. The resulting $\nu^{-1/2}$ information lifetime and $\nu^{1/4}$ regret dependence therefore characterize a different minimax problem rather than a sharper bound for the same variation-budget class.

Forgetting and adaptive memory. Restarts, sliding windows, discounting, and temporal kernels are standard ways to limit the use of stale data (Cheung et al., 2020; Domingues et al., 2021; Mao et al., 2021; Feng et al., 2023). BORL tunes a window through a bandit meta-algorithm (Cheung et al., 2020), while MASTER provides a multi-scale reduction for adapting to unknown nonstationarity and has also been used for stochastic shortest paths (Wei and Luo, 2021; Chen and Luo, 2022). Our contribution is not the restart or multi-scale schedule itself. For known ν, we derive the restart period from the information freshness principle and prove that it gives the minimax diffusion-dependent rate. For unknown ν, we combine the changing-kernel base algorithm with the two MASTER tests and show that adaptation loses no additional polynomial factor. Recent work shows that MASTER's tests may fail to trigger at practical horizons in piecewise-stationary bandits (Gerogiannis et al., 2025). This observation motivates our evaluation of the executable adaptive wrapper in addition to its asymptotic guarantee.

Stochastic and latent dynamics. The closest modeling precedents describe nonstationarity through hidden Markov modes (Choi et al., 1999), Brownian or random-walk arm means (Slivkins and Upfal, 2008), autoregressive rewards (Chen et al., 2023), and adaptive, latent, or factorized dynamics in RL (Nagabandi et al., 2019; Xie et al., 2021; Feng et al., 2022). Smoothly evolving MDPs have also been studied for robust planning when the current model is revealed (Lecarpentier and Rachelson, 2019). The Brownian restless bandit of Slivkins and Upfal (2008) has a related stochastic evolution mechanism, but it does not involve transition learning or multi-stage error propagation. The dynamic-parameter MDP of Xie et al. (2021) infers an unobserved latent process from experience and uses the inferred representation for off-policy learning, but does not characterize dynamic regret. Our model instead states the evolution condition directly on transition kernels, does not require estimating a latent representation, and covers both smooth evolution and stochastic refreshes. Matching upper and lower bounds then characterize how long historical transition data remain useful and the minimax cost of tracking the evolving MDP.

C Imported Results and Basic Tools

C.1 Imported results

We use two results from earlier work. We state them in our notation and identify their exact sources. We do not repeat the original stationary and scheduling proofs. The function $C(\cdot)$ and the notation $\widetilde{O}(\cdot)$ absorb constants and logarithmic factors. Every step that uses Assumption 2.1 is proved in this appendix.

Lemma C.1 (Stationary optimistic guarantees; UCBVI/MVP)**.** *For any fixed-kernel stage-homogeneous finite-horizon tabular MDP with normalized episode return in $[0,1]$, a prefix-uniform optimistic model-based algorithm such as UCBVI with Bernstein bonuses or MVP can be run so that, with probability at least $1-\delta$, simultaneously for all local rounds t and all prefixes $n \le K$,*

$$\widetilde{f}_t \ge f^\star, \qquad \sum_{u=1}^{n}(\widetilde{f}_u - f(\pi_u)) \le c\, C_{\mathrm{stat}}(n), \qquad \sum_{u=1}^{n}(\widetilde{f}_u - Y_u) \le c\, C_{\mathrm{stat}}(n).$$

For MVP one can take $C_{\mathrm{stat}}(n) = \widetilde{O}((\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|^2|\mathcal{A}|) \wedge n)$, where the logarithms depend on $|\mathcal{S}|, |\mathcal{A}|, H, n, 1/\delta$; Bernstein-style UCBVI gives the analogous guarantee with its standard lower-order terms.

Source and adaptation. For MVP, the displayed result is Theorem 1 in the full version accompanying Zhang et al. (2021), after normalizing episode returns to $[0,1]$. Lemmas 3–4 of that paper prove concentration and optimism. Lemma 7 gives the regret decomposition, and Lemmas 8–10 bound its three error terms. A union bound over $n \le K$ makes these bounds hold for every prefix. Adding the episode-return martingale and applying Freedman's inequality gives the version with Y_u (Freedman, 1975). The corresponding UCBVI-BF results are Theorem 2 and supplementary Lemmas 1 and 19 of Azar et al. (2017). Our extension to data from changing kernels is proved in Lemmas C.5 and G.10. □

Lemma C.2 (MASTER single-block scheduling and charging)**.** *Consider an online decision problem with expected normalized returns $f_t(\pi) \in [0,1]$ and observed returns $Y_t \in [0,1]$. Let C be nondecreasing and let $\eta(n) = C(n)/n$ be dyadically nonincreasing. Let B be the realized part of one dyadic MASTER block of length N, possibly cut short by a restart. Suppose a partition $\mathcal{P}_B$ of B satisfies the uniform prefix bounds required by the two MASTER tests. Then, on the base and scheduling event, the normalized observed regret on B satisfies*

$$\sum_{t\in B}(f_t^\star - Y_t) \le \widetilde{O}\left[\sum_{I\in\mathcal{P}_B} C(|I|) + \sum_{q\in\{1,2,4,\ldots,N\}} \frac{\eta(q)}{\eta(N)} C(q)\right].$$

Source and adaptation. This is Lemma 4 of Wei and Luo (2021), written with $\rho(n) = \eta(n)$ and $C(n) = n\eta(n)$. Their Assumption 1 and Eq. (1) give the bounds on each stable interval. Their Lemma 3 handles copies that pause and resume, and Lemmas 16–17 bound replay. We keep the replay term explicit because it may repeatedly pay the startup part of $C(n) = (a\sqrt{n} + b) \wedge n$. Later

we verify the stable-interval bounds for changing kernels and show that a costly replay requires a large kernel change. □

C.2 Information structure and proof events

Let $\mathcal{F}_k$ contain everything available at the end of episode k: the kernel history, the agent's random choices, and the observed trajectories. Thus $\mathcal{E}_k \subseteq \mathcal{F}_k$. Exogeneity means that the agent's random choices and trajectory noise reveal no extra information about future kernels once $\mathcal{E}_k$ is known. The displacement bound therefore remains valid after conditioning on $\mathcal{F}_k$. Lemma F.1 extends this fact to random block starting times. For concentration within one transition row, $(\mathcal{H}_i)_i$ denotes the history of observations from that row.

We use two high-probability events. The event Ω_{MVP} contains all concentration, variance, Bellman, and return bounds needed by one MVP copy. The event Ω_{ad} contains the same bounds for every copy created by MASTER, together with the scheduling and replay bounds. There are only polynomially many events, so a union bound changes only logarithmic factors. Below, "on the joint event" means the relevant one of these two events.

C.3 Basic deterministic identities

We use two elementary identities throughout the proofs. The first turns total-variation distance into value error. The second compares two policies by summing their one-step value gaps.

Lemma C.3 (TV-span inequality)**.** *For probability measures p, q and any bounded function v,*

$$|(p-q)v| \le d_{\mathrm{TV}}(p,q)\,\mathrm{span}(v), \qquad \mathrm{span}(v) := \max_x v(x) - \min_x v(x).$$

Proof. Let $m = (\max v + \min v)/2$ and $g = v - m$. Then $(p-q)m = 0$ and $\|g\|_\infty = \mathrm{span}(v)/2$. Thus $|(p-q)v| = |(p-q)g| \le \|p-q\|_1 \|g\|_\infty = d_{\mathrm{TV}}(p,q)\,\mathrm{span}(v)$. □

The next identity is the finite-horizon performance-difference lemma (Kakade and Langford, 2002), stated with the trajectory generated by the second policy.

Lemma C.4 (Finite-horizon performance difference)**.** *Fix an MDP with transition kernel P and policies π, π'. Then*

$$V_1^{\pi'}(\rho) - V_1^{\pi}(\rho) = \mathbb{E}_{\pi',P}\left[\sum_{h=1}^{H}\left(V_h^{\pi'}(S_h) - Q_h^{\pi'}(S_h,A_h)\right)\right].$$

Proof. For a trajectory generated by π, write

$$V_h^{\pi'}(S_h) - V_h^{\pi}(S_h) = V_h^{\pi'}(S_h) - Q_h^{\pi'}(S_h,A_h) + P\left(V_{h+1}^{\pi'} - V_{h+1}^{\pi}\right)(S_h,A_h).$$

Taking conditional expectations and telescoping over h gives the identity. □

C.4 Concentration around the sample mixture

For the support-dependent refinements, define $\mathcal{N}(s,a) := \bigcup_{k\in[K]} \mathrm{supp}\, P_k(\cdot \mid s,a)$ and $d_{\mathrm{succ}} := \max_{s,a} |\mathcal{N}(s,a)|$. This notation formalizes the bounded-successor condition in Corollary 4.3.

Several parts of the MVP proof use the same concentration event, so we state it once. We need one bound for value vectors that do not depend on the current transition samples, and another bound for vectors computed from those samples. The support-dependent version adds no further assumption.

Lemma C.5 (Concentration around the sample mixture)**.** *Fix a row (s,a), and order its next-state indicators as $X_i \in \{e_1,\dots,e_{|\mathcal{S}|}\}$. Put $\mu_i = \mathbb{E}[X_i \mid \mathcal{H}_{i-1}]$, $\widehat{P}_N = N^{-1}\sum_{i=1}^N X_i$, and $\overline{P}_N = N^{-1}\sum_{i=1}^N \mu_i$. If observation i comes from episode u_i, then $\mu_i = P_{u_i}(\cdot \mid s,a)$; episode labels are nondecreasing and may repeat. Let $L = \log(c|\mathcal{S}|^2|\mathcal{A}|HK^6/\delta)$. With probability at least $1-\delta$, the bounds below hold for all copies, rows, destination states, and dyadic N. Let $\mathcal{V} \subset [0,1]^{|\mathcal{S}|}$ contain at most KH vectors determined by the kernel path but not by transition noise. If the samples and*

comparison episode u lie in an interval J, the bounds hold for every $v \in \mathcal{V}$ and every data-dependent z with $\operatorname{span}(z) \le B$*:*

$$|\widehat{P}_N(x) - \overline{P}_N(x)| \le c\sqrt{\frac{\overline{P}_N(x)L}{N}} + c\frac{L}{N}, \qquad x \in \mathcal{S},$$

$$|(\widehat{P}_N - P_u)v| \le c\sqrt{\frac{\operatorname{Var}_{\widehat{P}_N}(v)L}{N}} + c\frac{L}{N} + cD(J), \qquad v \in \mathcal{V},$$

$$|(\widehat{P}_N - \overline{P}_N)z| \le c\sqrt{\frac{|\mathcal{S}|\operatorname{Var}_{\overline{P}_N}(z)L}{N}} + c\frac{|\mathcal{S}|BL}{N}, \qquad \operatorname{span}(z) \le B,$$

$$|(\widehat{P}_N - \overline{P}_N)z| \le c\sqrt{\frac{d_{\mathrm{succ}}\operatorname{Var}_{\overline{P}_N}(z)L}{N}} + c\frac{d_{\mathrm{succ}}BL}{N}, \quad \operatorname{span}(z) \le B.$$

Proof. The proof has four steps. First, we control each next-state probability. Second, we use a coupling to handle value vectors that are independent of the current transition samples. Third, we turn the coordinatewise bound into a bound that is valid for sample-dependent vectors. Fourth, we repeat the same argument using only the possible next states.

Step 1: Establish the coordinatewise event.

Fix a destination state x. The error $Z_i = \mathbf{1}\{X_i = e_x\} - \mu_i(x)$ has conditional mean zero and is bounded by one. Its total conditional variance is at most $\sum_{i=1}^{N} \mu_i(x) = N\overline{P}_N(x)$. Freedman's inequality (Freedman, 1975) gives the desired bound for one x and one N. Dyadic peeling and a union bound make it hold for all required choices at once.

Step 2: Couple to the comparison kernel.

We now handle a value vector v that does not depend on the current transition samples. Condition on the complete kernel path. This makes every vector in $\mathcal{V}$ fixed, while exogeneity preserves the martingale structure of the transition samples. For each $X_i \sim \mu_i$, construct an auxiliary sample $Y_i \sim P_u(\cdot \mid s, a)$ by maximal coupling. Use a fresh coupling randomizer at every step. Conditional on the past, each Y_i has the same law $P_u(\cdot \mid s, a)$, so the Y_i's are conditionally i.i.d. Moreover,

$$\mathbb{P}(X_i \neq Y_i \mid \mathcal{H}_{i-1}) = d_{\mathrm{TV}}(\mu_i, P_u(\cdot \mid s, a)) \le D(J).$$

Let M_N count the pairs for which $X_i \neq Y_i$, and let $\widehat{Q}_N = N^{-1}\sum_i Y_i$. Freedman's inequality bounds the fraction of mismatched pairs:

$$\frac{M_N}{N} \le D(J) + c\sqrt{\frac{D(J)L}{N}} + c\frac{L}{N} \le cD(J) + c\frac{L}{N}.$$

For a fixed $v \in \mathcal{V}$, the empirical-Bernstein inequality for the auxiliary sample (Maurer and Pontil, 2009, Theorem 4) gives

$$|(\widehat{Q}_N - P_u)v| \le c\sqrt{\frac{\operatorname{Var}_{\widehat{Q}_N}(v)L}{N}} + c\frac{L}{N}.$$

The two empirical distributions differ only on the mismatched pairs, so $d_{\mathrm{TV}}(\widehat{P}_N, \widehat{Q}_N) \le M_N/N$. Expanding their variances also gives

$$\left|\operatorname{Var}_{\widehat{P}_N}(v) - \operatorname{Var}_{\widehat{Q}_N}(v)\right| \le 3\frac{M_N}{N}.$$

We also have $|(\widehat{P}_N - \widehat{Q}_N)v| \le M_N/N$. Combining these bounds and using $\sqrt{xy} \le (x+y)/2$ gives

$$|(\widehat{P}_N - P_u)v| \le c\sqrt{\frac{\operatorname{Var}_{\widehat{P}_N}(v)L}{N}} + c\frac{L}{N} + cD(J).$$

The logarithm L is large enough for a union bound over all vectors, comparison episodes, stages, rows, copies, and dyadic prefixes. Later, $\mathcal{V}$ will be the set of normalized optimal value vectors $\{\bar{V}^{\star}_{u,h} : u \in [K], h \in [H]\}$. These vectors depend on the kernel path but not on the transition samples used by the empirical kernel. We do not use this coupling argument for the data-dependent optimistic values.

Step 3: Make the Bellman input data dependent.

We next allow z to depend on the transition samples. Set $m = \overline{P}_N z$. Since $\widehat{P}_N - \overline{P}_N$ has total mass zero, replacing z by $z - m\mathbf{1}$ does not change the quantity being bounded. The coordinatewise event and Cauchy–Schwarz then give

$$|(\widehat{P}_N - \overline{P}_N)z| \le c\sqrt{\frac{L}{N}} \sum_x \sqrt{\overline{P}_N(x)}|z(x) - m| + c\frac{L}{N}\sum_x |z(x) - m|$$
$$\le c\sqrt{\frac{|\mathcal{S}| \operatorname{Var}_{\overline{P}_N}(z) L}{N}} + c\frac{|\mathcal{S}|BL}{N}.$$

The coordinatewise event holds before z is chosen. Therefore, the resulting inequality remains valid even when z is computed from the same row samples. This is the form needed for the Bellman residual.

Step 4: Restrict to the possible next states.

Fix a row (s, a). Every distribution μ_i is supported on $\mathcal{N}(s, a)$, so the same is true of $\overline{P}_N$. Every observed next state also belongs to this set. Hence $\widehat{P}_N - \overline{P}_N$ is zero outside $\mathcal{N}(s, a)$. We may restrict the preceding sums to at most d_{succ} states. Cauchy–Schwarz then gives

$$\sum_{x \in \mathcal{N}(s,a)} \sqrt{\overline{P}_N(x)}|z(x) - m| \le \sqrt{d_{\mathrm{succ}} \operatorname{Var}_{\overline{P}_N}(z)}, \qquad \sum_{x \in \mathcal{N}(s,a)} |z(x) - m| \le d_{\mathrm{succ}} B.$$

Substituting these two bounds proves the support-dependent inequality. The set $\mathcal{N}(s, a)$ is used only in the proof; the agent does not need to know it. □

D Proof of Lemma 3.1

We first establish a deterministic value-sensitivity inequality and then apply the conditional displacement bound in Assumption 2.1.

Step 1: Establish deterministic value sensitivity.

Fix two stage-homogeneous kernels P, Q and a policy π, and let $\delta_h = \left\| V^{\pi}_{P,h} - V^{\pi}_{Q,h} \right\|_{\infty}$. The Bellman equations and Lemma C.3 give

$$\delta_h \le \delta_{h+1} + \sup_{s,a} \left|(P - Q)V^{\pi}_{Q,h+1}(s, a)\right| \le \delta_{h+1} + (H - h)\left\|P - Q\right\|_{\infty,\mathrm{TV}},$$

because $\operatorname{span}(V^{\pi}_{Q,h+1}) \le H - h$. Since $\delta_{H+1} = 0$, summation gives

$$\delta_1 \le \sum_{h=1}^{H-1} (H - h)\left\|P - Q\right\|_{\infty,\mathrm{TV}} = \frac{H(H-1)}{2}\left\|P - Q\right\|_{\infty,\mathrm{TV}}.$$

Step 2: Apply the displacement condition.

The preceding inequality is uniform over π. Setting $P = P_{k+\ell}$ and $Q = P_k$, conditional Jensen and Assumption 2.1 yield

$$\mathbb{E}\Big[\left\|P_{k+\ell} - P_k\right\|_{\infty,\mathrm{TV}} \mid \mathcal{E}_k\Big] \le \sqrt{\mathbb{E}\Big[\left\|P_{k+\ell} - P_k\right\|^2_{\infty,\mathrm{TV}} \mid \mathcal{E}_k\Big]} \le \sqrt{\min\{1, \nu\ell\}}.$$

Taking conditional expectations in the deterministic inequality proves the policy-uniform conclusion. Finally, $|\sup_\pi x_\pi - \sup_\pi y_\pi| \le \sup_\pi |x_\pi - y_\pi|$ gives the same bound for optimal values. This proves Lemma 3.1. □

Corollary D.1 (Normalized value diameter)**.** *For normalized returns* $f_P(\pi) = H^{-1}V^{\pi}_{P,1}(\rho)$*,*

$$\sup_{\pi} |f_P(\pi) - f_Q(\pi)| \le \frac{H-1}{2} \|P - Q\|_{\infty,\mathrm{TV}}.$$

Consequently, if $u, v \in I$*, then* $\sup_\pi |f_u(\pi) - f_v(\pi)| \le (H-1)D(I)/2$*.*

Proof. Divide the deterministic value-sensitivity inequality established above by H. □

E Proof of Proposition 3.2

For $\nu > 0$, differentiating the freshness objective gives

$$\mathcal{J}'_\nu(W) = -\frac{H\sqrt{|\mathcal{S}|\,|\mathcal{A}|}}{2W^{3/2}} + \frac{H^2\sqrt{\nu}}{2W^{1/2}}.$$

The derivative is negative for $W < \sqrt{|\mathcal{S}|\,|\mathcal{A}|}/(H\sqrt{\nu})$ and positive above this value, proving the unique unconstrained minimizer. Substituting this minimizer into the two terms gives $\mathsf{Stat}(\overline{W}_\nu) = \mathsf{Shift}_\nu(\overline{W}_\nu) = H^{3/2}(|\mathcal{S}|\,|\mathcal{A}|)^{1/4}\nu^{1/4}$, and hence the stated value of $\mathcal{J}_\nu(\overline{W}_\nu)$. Restricting to $[1, K]$ clips that minimizer at the endpoints; rounding to an adjacent integer changes each square-root term by at most a universal factor. When $\nu = 0$, $\mathcal{J}_0(W) = H\sqrt{|\mathcal{S}|\,|\mathcal{A}|/W}$ is decreasing, so the minimizer is K. □

F Interval diameter under diffusive distribution shift

The freshness calculation compares two episodes. The algorithms reuse a whole interval of data, so we must control the largest kernel difference inside that interval. For $I \subseteq [K]$, define $D(I) := \sup_{u,v\in I} \|P_u - P_v\|_{\infty,\mathrm{TV}}$ and $\Gamma(I) := \min\{1, HD(I)\}$.

Lemma F.1 (Interval diameter under diffusive distribution shift)**.** *Under Assumption 2.1, there is a universal constant* c_D *such that every deterministic interval* $I \subseteq [K]$ *satisfies the first two bounds below. Moreover, after padding the path by* $P_t = P_K$ *for* $t > K$*, every bounded* $(\mathcal{F}_k)$*-stopping time* τ*, length* N*, and* $x \in (0, 1]$ *satisfy the last two bounds, where* $I_{\tau,N} := \{\tau, \dots, \tau + N - 1\}$*:*

$$\begin{aligned}
\mathbb{E}[D(I)] &\le c_D\sqrt{\nu|I|}\log(2|I|),\\
\mathbb{E}[\Gamma(I)] &\le c_D \min\left\{1, H\sqrt{\nu|I|}\log(2|I|)\right\},\\
\mathbb{E}[D(I_{\tau,N})^2 \mid \mathcal{F}_\tau] &\le c_D\nu N\log^2(2N),\\
\mathbb{P}(\Gamma(I_{\tau,N}) \ge x \mid \mathcal{F}_\tau) &\le c_D \min\left\{1, \frac{H^2\nu N\log^2(2N)}{x^2}\right\}.
\end{aligned}$$

The proof has three steps. First, we show that the displacement bound remains valid when a block starts at a random time chosen from past data. Second, we bound all kernel differences inside such a block. Third, we specialize the result to deterministic intervals.

Step 1: Lift the displacement bound to stopping times.

Fix k. Once $\mathcal{E}_k$ is given, the agent's randomization and trajectory noise carry no information about future kernels. Conditioning on the larger history $\mathcal{F}_k$ therefore does not change the conditional law of those kernels. Thus Assumption 2.1 also holds given $\mathcal{F}_k$.

For a bounded stopping time τ, take any $A \in \mathcal{F}_\tau$. Since $A \cap \{\tau = k\} \in \mathcal{F}_k$, summing the deterministic-time bound over k gives

$$\mathbb{E}\left[\mathbf{1}_A \|P_{\tau+\ell} - P_\tau\|^2_{\infty,\mathrm{TV}}\right] \le \nu\ell\,\mathbb{P}(A).$$

Because this inequality holds for every $A \in \mathcal{F}_\tau$, it is exactly the required conditional bound. If the block extends past K, constant padding can only reduce the number of episodes over which the kernel changes.

Step 2: Control the diameter of an adaptive block.

First let $N = 2^J$. The claim is trivial for $J = 0$. For $r = 0, \dots, J-1$, define

$$M_r := \max_{0 \le j < 2^{J-r}-1} \left\| P_{\tau+(j+1)2^r} - P_{\tau+j2^r} \right\|_{\infty,\mathrm{TV}}.$$

At scale r, the maximum squared increment is at most the sum of all squared increments at that scale. Step 1 and the tower property then give $\mathbb{E}[M_r^2 \mid \mathcal{F}_\tau] \le \nu N$. Next, write the offset of any episode in binary. This connects its kernel to P_τ using at most one increment from each scale. Hence $D(I_{\tau,N}) \le 2\sum_r M_r$. Cauchy–Schwarz gives the stated second-moment bound. If N is not a power of two, place the block inside a dyadic block of length less than $2N$. This changes only the constant. The probability bound now follows from conditional Markov's inequality and $\Gamma(I) \le HD(I)$.

Step 3: Recover the deterministic expectation bound.

A deterministic interval is a special case of Step 2. Jensen's inequality gives the bound on $\mathbb{E}[D(I)]$. Combining it with $\Gamma(I) \le 1$ and $\Gamma(I) \le HD(I)$ gives the second bound. This proves Lemma F.1.□

G Proof of Proposition G.1: MVP with Data from Changing Kernels

This section proves that MVP keeps its stationary learning rate when it reuses data from changing kernels, up to an additional interval-diameter term. We first define the guarantee and the executable base algorithm, then prove optimism and control its cumulative Bellman error.

Throughout this section, $\Gamma(I) := \min\{1, HD(I)\}$.

Base guarantee. Let $C : \mathbb{N} \to [1, \infty)$ be nondecreasing, with $C(n) \le n$, and let $\eta(n) := C(n)/n$ decrease over dyadic n. An algorithm has the C-base guarantee if the following two bounds hold on every prefix J of the interval where it runs. The algorithm outputs a policy π_t and a predicted normalized return $\widetilde{f}_t \in [0,1]$ in each episode:

$$\widetilde{f}_u \ge \min_{v \in J} f_v^\star - c_0\Gamma(J) \quad \text{for all } u \in J, \qquad \sum_{u \in J} (\widetilde{f}_u - f_u(\pi_u)) \le c_0\{C(|J|) + |J|\Gamma(J)\}.$$

These bounds hold with probability at least $1-\delta$. After increasing the logarithmic factor in C, the second bound also holds with the observed return Y_u in place of its mean $f_u(\pi_u)$. MASTER uses this observed-return form.

Proposition G.1 (MVP guarantee for data from changing kernels)**.** *The MVP algorithm below satisfies the C_{MVP}-base guarantee on the unrestricted model class defined by Assumption 2.1. If $d_{\mathrm{succ}} \le d_0$ almost surely for a deterministic d_0, it also satisfies the $C_{\deg,0}$-base guarantee, where*

$$C_{\mathrm{MVP}}(n) = \widetilde{O}\Big((\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|^2|\mathcal{A}|) \wedge n\Big),$$
$$C_{\deg,0}(n) = \widetilde{O}\Big((\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + d_0|\mathcal{S}|\,|\mathcal{A}|) \wedge n\Big).$$

The base algorithm does not need to know d_0 or the next-state sets $\mathcal{N}(s,a)$.

We prove the stronger result in Proposition G.2, which also allows an MVP copy to pause and resume. Three points need care. The estimate uses transitions from earlier kernels, but its value prediction must still be optimistic for the current kernel. Its statistical error should keep the usual square-root rate, with only an added $n\Gamma(J)$ term for kernel change. Finally, the few episodes that cross a count threshold, together with trajectory noise, must fit inside the stated lower-order terms.

G.1 Algorithm and proof map

This subsection defines the MVP copy used by both algorithms and states the stronger guarantee needed when a copy pauses and resumes. It also introduces the safe-prefix device used only in the proof.

Executable base. We work throughout in normalized units. Put $\bar{r}_h = r_h/H$, and write $\bar{V}^\star_{u,h} = V^\star_{u,h}/H$; hence every normalized value lies in $[0,1]$ and $\sum_h \bar{r}_h(S_h, A_h) \le 1$ on every trajectory. A base copy pools transition observations across stages, which is valid because P_u is stage homogeneous.

At the beginning of local episode q, let $N_{q-1}(s,a)$ be the number of observations of row (s,a) previously received by this copy. If the row is nonempty, freeze the dyadic snapshot

$$n_q(s,a) = 2^{\lfloor \log_2 N_{q-1}(s,a) \rfloor}$$

and form $\widehat{P}_q(\cdot \mid s,a)$ from the first $n_q(s,a)$ observations of that row. If the row is empty, its backup is set to one. For a nonempty row define

$$b(p,v,n) = c_b\left(\sqrt{\frac{\mathrm{Var}_p(v)L}{n}} + \frac{L}{n}\right), \qquad L = \log\frac{c|\mathcal{S}|^2|\mathcal{A}|HK^6}{\delta},$$

and compute

$$\begin{aligned} \widetilde{V}_{q,H+1} &= 0, \\ \widetilde{Q}_{q,h}(s,a) &:= \min\left\{1, \bar{r}_h(s,a) + \widehat{P}_q\widetilde{V}_{q,h+1}(s,a) + b(\widehat{P}_q, \widetilde{V}_{q,h+1}, n_q(s,a))\right\}, \\ \widetilde{V}_{q,h}(s) &:= \max_a \widetilde{Q}_{q,h}(s,a). \end{aligned} \tag{1}$$

The copy predicts $\widetilde{f}_q = \widetilde{V}_{q,1}(\rho)$, follows a greedy policy, and updates only when it is active. The proof stops its Bellman calculation at the first visit to a row that was unseen at the start of the episode or that reaches the next dyadic count threshold. Call this stage σ_q, and set $\sigma_q = H+1$ if no such visit occurs. The stages before σ_q form the *safe prefix*. We call the episode *cut* when $\sigma_q \le H$. Whether a stage belongs to the safe prefix is known before its transition is sampled.

This stopping rule exists only in the proof. The algorithm keeps every transition and updates its snapshots before its next active episode. When the proof drops the suffix after σ_q, it pays at most one unit of normalized return. The average kernel $\overline{P}_q$ is also used only in the proof.

The display assumes known fixed rewards. If a reward is initially unknown, assign the optimistic value $1/H$ until its first visit and then store the observed reward. These reward charges sum to at most $|\mathcal{S}|\,|\mathcal{A}|$, which both complexity bounds absorb.

The next proposition is stated for a predictable subsequence of episodes, so it also covers MVP copies that pause under MASTER. Its proof separates three costs. Variance bounds control sampling error. The interval diameter controls the difference between the sampled kernels and the current kernel. The number of cut episodes controls the remaining startup cost.

Proposition G.2 (MVP guarantee on a changing-kernel interval)**.** *Run the MVP instance above on any predictable subsequence $\tau_1 < \tau_2 < \cdots$, and update it only on active episodes. On one high-probability event, every prefix of n active episodes satisfies the bounds below. Here $J = [\tau_1, \tau_n]$ is the smallest enclosing time interval, and d_0 is any deterministic upper bound on d_{succ}.*

$$\begin{aligned} C_{\mathrm{MVP}}(n) &:= \widetilde{O}\Big(\min\{n, \sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|^2|\mathcal{A}|\}\Big), \\ C_{\deg}(n) &:= \widetilde{O}\Big(\min\{n, \sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|\}\Big), \\ C_{\deg,0}(n) &:= \widetilde{O}\Big(\min\{n, \sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + d_0|\mathcal{S}|\,|\mathcal{A}|\}\Big), \\ \widetilde{f}_q &\ge \min_{v\in J} f^\star_v - c_1\Gamma(J) \quad (q \in [n]), \\ \sum_{q=1}^{n}(\widetilde{f}_q - f_{\tau_q}(\pi_q)) &\le c_1\left(C_{\mathrm{MVP}}(n) + n\Gamma(J)\right). \end{aligned}$$

After enlarging the logarithmic factor, the same sum bound holds with Y_{τ_q} in place of $f_{\tau_q}(\pi_q)$. The conclusions also hold with $C_{\deg}$ in place of C_{MVP}. The agent need not know the next-state sets or d_{succ}. If $d_{\mathrm{succ}} \le d_0$, it is enough to tune the wrapper with $C_{\deg,0}$.

The remainder of the section proves this stronger statement. The main proposition then follows by taking the active episodes to be contiguous.

G.2 Proof of the changing-kernel interval guarantee

The proof needs three ingredients. First, a monotone bonus transfers optimism from the estimated model to the current MDP. Second, a one-step Bellman bound separates sampling error from kernel change. Third, a variance recursion sums the sampling errors over an episode. The final calculation also adds the cut episodes and the return martingale.

Cut-episode count. The usual doubling-count argument for MVP (Zhang et al., 2021, Section 6.3) shows that a base copy with n active episodes has at most

$$|\mathcal{S}|\,|\mathcal{A}|\big(1+\lceil\log_2(nH)\rceil\big)$$

cut episodes. To see this, charge a cut episode to the row that caused it. The cause is either the first observation of that row or the first visit after a new dyadic count threshold. Each row can be charged once per threshold, and no row is visited more than nH times.

We next use the monotone Bernstein construction from Zhang et al. (2021, Lemma 14). We state and verify the scaled version because our value vectors may have span B_0.

Lemma G.3 (Monotone Bernstein envelope)**.** *For every fixed universal width $B_0 \ge 1$ and sufficiently large universal constant a_0, define F by the first line below. On every box $[m, m+B_0]^{|\mathcal{S}|}$, F is coordinatewise nondecreasing in v and is invariant under constant shifts: $F(p, v-c\mathbf{1}, n) = F(p, v, n) - c$. For any prescribed universal constant c_{conf}, the constants a_0 and c_b in equation (1) can be chosen so that*

$$F(p,v,n) := pv + \max\left\{ a_0\sqrt{\frac{\mathrm{Var}_p(v)L}{n}}, a_0^2 B_0 \frac{L}{n} \right\},$$

$$pv + b(p,v,n) \ge F(p,v,n) \ge pv + c_{\mathrm{conf}}\left(\sqrt{\frac{\mathrm{Var}_p(v)L}{n}} + \frac{L}{n}\right).$$

Proof. Adding a constant to every coordinate changes the mean by that constant and leaves the variance unchanged. This proves the shift identity. We next check monotonicity. When the linear term $a_0^2 B_0 L/n$ is active, the derivative with respect to $v(x)$ is $p(x)$. When the square-root term is active, the derivative is

$$p(x)\left[1 + \frac{a_0(v(x)-pv)}{\sqrt{n\,\mathrm{Var}_p(v)/L}}\right].$$

On this branch, $\sqrt{n\,\mathrm{Var}_p(v)/L} \ge a_0 B_0$. Since $v(x) - pv \ge -B_0$, the derivative is nonnegative. Continuity gives the same conclusion where the two branches meet. Finally, the maximum of two nonnegative numbers is at least half their sum. Choosing a_0 large enough, and then increasing c_b, proves the two comparison inequalities. □

Lemma G.4 (Variance comparison)**.** *Assume the coordinate bounds in Lemma C.5. Suppose the row samples and the episode u all lie in J. Then the first two bounds below hold for any, possibly data-dependent, vector v with $\mathrm{span}(v) \le B$. The second bound also uses the next-state support size. The last bound holds for any distributions p, q and bounded vector z:*

$$\mathrm{Var}_{\widehat{P}_N}(v) \le \frac{3}{2}\mathrm{Var}_{P_u}(v) + c|\mathcal{S}|B^2 D(J) + c\frac{|\mathcal{S}|B^2 L}{N},$$

$$\mathrm{Var}_{\widehat{P}_N}(v) \le \frac{3}{2}\mathrm{Var}_{P_u}(v) + c d_{\mathrm{succ}} B^2 D(J) + c\frac{d_{\mathrm{succ}} B^2 L}{N},$$

$$|\mathrm{Var}_p(z) - \mathrm{Var}_q(z)| \le 3\,\mathrm{span}(z)^2 d_{\mathrm{TV}}(p,q).$$

Proof. The coordinatewise concentration bound and $\big|\overline{P}_N(x) - P_u(x)\big| \le D(J)$ imply

$$\widehat{P}_N(x) \le \tfrac{3}{2}P_u(x) + cD(J) + cL/N.$$

Let $m = P_u v$. The empirical variance is at most $\sum_x \widehat{P}_N(x)(v(x)-m)^2$. Insert the preceding coordinate bound and use $|v(x)-m| \le B$. This proves the first inequality. For the last inequality,

shift z so that its values lie in $[0, B]$, where $B = \mathrm{span}(z)$. Then $\left|(p-q)z^2\right| \le B^2 d_{\mathrm{TV}}(p,q)$ and $\left|(pz)^2 - (qz)^2\right| \le 2B^2 d_{\mathrm{TV}}(p,q)$.

For the support-dependent version, both P_u and $\widehat{P}_N$ are zero outside $\mathcal{N}(s,a)$. The same calculation therefore sums over at most d_{succ} states instead of $|\mathcal{S}|$, which proves the second inequality. □

Proposition G.5 (Approximate optimism for the current MDP)**.** *Let $\tau_1 < \cdots < \tau_n$ be the active episodes of one copy and let $J_q = [\tau_1, \tau_q]$. On the sample-mixture event, if $HD(J_q) \le 1$, then every local episode q, stage h, and state s satisfy*

$$\widetilde{V}_{q,h}(s) \ge \bar{V}^{\star}_{\tau_q,h}(s) - c_\star(H-h)D(J_q).$$

For every J_q, including the case $HD(J_q) > 1$, $\widetilde{f}_q \ge f^{\star}_{\tau_q} - c_\star \Gamma(J_q)$, where $f^{\star}_u = V^{\star}_{u,1}(\rho)/H$.

Proof. If $HD(J_q) > 1$, the scalar claim is immediate from $\widetilde{f}_q \ge 0$, $f^{\star}_{\tau_q} \le 1$, and $c_\star \ge 1$. It remains to consider $HD(J_q) \le 1$. We prove the value bound by backward induction. Both value functions are zero at stage $H+1$. Assume the claim holds at stage $h+1$, and write $u = \tau_q$, $d = D(J_q)$, and $e_{h+1} = c_\star(H-h-1)_+ d$. Choose $c_\star$ larger than the diameter constant in the concentration bound and choose $B_0 > 1 + c_\star$. For a row that has been observed, the shift identity and monotonicity of the envelope give

$$\begin{aligned}
\widehat{P}_q \widetilde{V}_{q,h+1} + b_q &\ge F(\widehat{P}_q, \widetilde{V}_{q,h+1}, n_q) \\
&= F(\widehat{P}_q, \widetilde{V}_{q,h+1} + e_{h+1}\mathbf{1}, n_q) - e_{h+1} \\
&\ge F(\widehat{P}_q, \bar{V}^{\star}_{u,h+1}, n_q) - e_{h+1}.
\end{aligned}$$

The induction hypothesis says that $\widetilde{V}_{q,h+1} + e_{h+1}\mathbf{1} \ge \bar{V}^{\star}_{u,h+1}$. Both vectors lie in a box of width at most $1 + c_\star$, so monotonicity applies. The Bernstein bound in Lemma C.5, with the current kernel as the comparison kernel, then gives

$$F(\widehat{P}_q, \bar{V}^{\star}_{u,h+1}, n_q) \ge P_u \bar{V}^{\star}_{u,h+1} - cd.$$

Add the normalized reward and maximize over actions. Clipping at one does not hurt the lower bound because every normalized optimal Q-value is at most one. An unseen row also has backup one, so it is optimistic. If rewards are initially unknown, the algorithm uses $1/H$ before the first visit and the exact observed reward afterward; this is optimistic as well. A sufficiently large $c_\star$ now closes the induction. At stage one, the error is at most $c_\star HD(J_q) = c_\star \Gamma(J_q)$. □

The next lemma sums conditional variances by telescoping squared values. It is the standard argument from Zhang et al. (2021, Lemmas 13 and 16), written for kernels that may change between episodes and for recursions that stop at a predictable stage.

Lemma G.6 (Total-variance recursion)**.** *For each episode u, let $Z_{u,h} : \mathcal{S} \to [0, B]$ be predictable, with $Z_{u,H+1} = 0$. Let $\sigma_u \in [H+1]$ be a stopping stage such that $\{h < \sigma_u\}$ is known before the stage-h transition. Write $\mathcal{I} = \{(u,h) : h < \sigma_u\}$. If the first line below holds on the visited pairs in $\mathcal{I}$, then the second line holds on an anytime event of probability at least $1-\delta$:*

$$Z_{u,h}(S_{u,h}) \le a_{u,h} + P_u Z_{u,h+1}(S_{u,h}, A_{u,h}), \qquad a_{u,h} \ge 0,$$

$$\sum_{(u,h)\in\mathcal{I}} \mathrm{Var}_{P_u}(Z_{u,h+1})(S_{u,h}, A_{u,h}) \le cB \sum_{(u,h)\in\mathcal{I}} a_{u,h} \quad + cB^2 \left(N_{\mathrm{cut}} + \log \frac{cKH}{\delta} \right),$$

where $N_{\mathrm{cut}} = \sum_u \mathbf{1}\{\sigma_u \le H\}$. The kernels may differ from episode to episode.

Proof. For one transition, write $z = Z_{u,h}(S_{u,h})$, $m = P_u Z_{u,h+1}(S_{u,h}, A_{u,h})$, and $z' = Z_{u,h+1}(S_{u,h+1})$. Add and subtract $(z')^2$ and z^2 to obtain

$$\mathrm{Var}_{P_u}(Z_{u,h+1}) = \{P_u Z^2_{u,h+1} - (z')^2\} + \{(z')^2 - z^2\} + \{z^2 - m^2\}$$

Summing the middle term over a stopped episode gives

$$Z_{u,\sigma_u}(S_{u,\sigma_u})^2 - Z_{u,1}(S_{u,1})^2 \le B^2 \mathbf{1}\{\sigma_u \le H\};$$

If $\sigma_u = H+1$, the endpoint is zero because $Z_{u,H+1} = 0$. The recursion also gives $z^2 - m^2 \le 2Ba_{u,h}$. The first brace has conditional mean zero and conditional variance at most $4B^2 \operatorname{Var}_{P_u}(Z_{u,h+1})$. Multiplying by the predictable indicator $\mathbf{1}\{h < \sigma_u\}$ preserves these facts. Freedman's inequality controls its sum. The inequality $cB\sqrt{V\ell} \le V/2 + c'B^2\ell$ then moves half of the variance term to the left-hand side. A dyadic union bound makes the result hold for every prefix. □

Lemma G.7 (One-step Bellman error under changing kernels)**.** *Fix a safe stage $h < \sigma_q$ of local episode q, put $u = \tau_q$, and let $d_q = D(J_q)$. On its visited row there is a nonnegative predictable $\beta_{q,h}$. Writing $Z_{q,h} = \widetilde{V}_{q,h+1} - \bar{V}^{\star}_{u,h+1}$, its Bellman, unrestricted-support charge, and support-dependent charge inequalities are, respectively,*

$$\widetilde{V}_{q,h}(S_{u,h}) \le \bar{r}_h(S_{u,h}, A_{u,h}) + P_u \widetilde{V}_{q,h+1}(S_{u,h}, A_{u,h}) + \beta_{q,h},$$

$$\beta_{q,h} \le \widetilde{O}\left[\sqrt{\frac{\operatorname{Var}_{P_u}(\widetilde{V}_{q,h+1})}{n_q}} + \sqrt{\frac{|\mathcal{S}| \operatorname{Var}_{P_u}(Z_{q,h})}{n_q}} + \frac{|\mathcal{S}|}{n_q} + \sqrt{\frac{|\mathcal{S}| d_q}{n_q}} + d_q\right],$$

$$\beta_{q,h} \le \widetilde{O}\left[\sqrt{\frac{\operatorname{Var}_{P_u}(\widetilde{V}_{q,h+1})}{n_q}} + \sqrt{\frac{d_{\mathrm{succ}} \operatorname{Var}_{P_u}(Z_{q,h})}{n_q}} + \frac{d_{\mathrm{succ}}}{n_q} + \sqrt{\frac{d_{\mathrm{succ}} d_q}{n_q}} + d_q\right].$$

Here row arguments and logarithmic factors are suppressed. For unknown fixed rewards, $\beta^r_{q,h}$ is added to and henceforth absorbed into the definition of $\beta_{q,h}$.

Proof. The proof has three steps. We first separate the Bellman error into a fixed-vector part, a sample-dependent part, and kernel mismatch. We then add the algorithmic bonus. Finally, we repeat the same bounds over the smaller set of possible next states.

Step 1: Control the exogenous comparison term.

Clipping can only reduce the upper Bellman error, so it is enough to study the unclipped backup. Add and subtract the current optimal value vector and the average sampled kernel:

$$(\widehat{P}_q - P_u)\widetilde{V}_{q,h+1} = (\widehat{P}_q - P_u)\bar{V}^{\star}_{u,h+1} + (\widehat{P}_q - \overline{P}_q)Z_{q,h} + (\overline{P}_q - P_u)Z_{q,h}.$$

The first term contains the fixed vector $\bar{V}^{\star}_{u,h+1}$. Apply the empirical-Bernstein bound in Lemma C.5. Then use $\bar{V}^{\star}_{u,h+1} = \widetilde{V}_{q,h+1} - Z_{q,h}$ and the variance comparison in Lemma G.4. This gives

$$\left|(\widehat{P}_q - P_u)\bar{V}^{\star}_{u,h+1}\right| \le \widetilde{O}\left[\sqrt{\frac{\operatorname{Var}_{P_u}(\widetilde{V}_{q,h+1})}{n_q}} + \sqrt{\frac{\operatorname{Var}_{P_u}(Z_{q,h})}{n_q}} + \sqrt{\frac{|\mathcal{S}| d_q}{n_q}} + \frac{|\mathcal{S}|}{n_q} + d_q\right].$$

We used $\operatorname{Var}_p(x+y) \le 2\operatorname{Var}_p(x) + 2\operatorname{Var}_p(y)$ to separate the two value vectors.

Step 2: Control the data-dependent terms and bonus.

The second term contains the sample-dependent vector $Z_{q,h}$, so we use the uniform bound in Lemma C.5. The variance comparison

$$\operatorname{Var}_{\overline{P}_q}(Z_{q,h}) \le \operatorname{Var}_{P_u}(Z_{q,h}) + cD(J_q).$$

together with Lemma C.3 gives

$$\left|(\widehat{P}_q - \overline{P}_q)Z_{q,h}\right| + \left|(\overline{P}_q - P_u)Z_{q,h}\right| \le \widetilde{O}\left[\sqrt{\frac{|\mathcal{S}|\operatorname{Var}_{P_u}(Z_{q,h})}{n_q}} + \sqrt{\frac{|\mathcal{S}|d_q}{n_q}} + \frac{|\mathcal{S}|}{n_q} + d_q\right],$$

Here $\operatorname{span}(Z_{q,h}) \le 2$. It remains to bound the bonus used by the algorithm. Applying Lemma G.4 directly to that bonus gives

$$b(\widehat{P}_q, \widetilde{V}_{q,h+1}, n_q) \le \widetilde{O}\left[\sqrt{\frac{\operatorname{Var}_{P_u}(\widetilde{V}_{q,h+1})}{n_q}} + \sqrt{\frac{|\mathcal{S}|d_q}{n_q}} + \frac{|\mathcal{S}|}{n_q}\right].$$

The fixed-vector error, the sample-dependent error, and the bonus together give the stated bound on $\beta_{q,h}$. If rewards are initially unknown, add the reward charge $\beta^r_{q,h}$ described above.

Step 3: Restrict the argument to the possible next states.

For the support-dependent result, use the versions of Lemmas G.4 and C.5 that sum only over possible next states. The fixed-vector term then satisfies

$$\left|(\widehat{P}_q - P_u)\bar{V}^{\star}_{u,h+1}\right| \le \widetilde{O}\Bigg[\sqrt{\frac{\operatorname{Var}_{P_u}(\widetilde{V}_{q,h+1})}{n_q}} + \sqrt{\frac{\operatorname{Var}_{P_u}(Z_{q,h})}{n_q}} + \sqrt{\frac{d_{\mathrm{succ}}d_q}{n_q}} + \frac{d_{\mathrm{succ}}}{n_q} + d_q\Bigg].$$

The sample-dependent term and the kernel-mismatch term satisfy

$$\left|(\widehat{P}_q - \overline{P}_q)Z_{q,h}\right| + \left|(\overline{P}_q - P_u)Z_{q,h}\right| \le \widetilde{O}\left[\sqrt{\frac{d_{\mathrm{succ}}\operatorname{Var}_{P_u}(Z_{q,h})}{n_q}} + \sqrt{\frac{d_{\mathrm{succ}}d_q}{n_q}} + \frac{d_{\mathrm{succ}}}{n_q} + d_q\right].$$

The same support-dependent variance comparison gives the following bonus bound:

$$b(\widehat{P}_q, \widetilde{V}_{q,h+1}, n_q) \le \widetilde{O}\left[\sqrt{\frac{\operatorname{Var}_{P_u}(\widetilde{V}_{q,h+1})}{n_q}} + \sqrt{\frac{d_{\mathrm{succ}}d_q}{n_q}} + \frac{d_{\mathrm{succ}}}{n_q}\right].$$

Because $d_{\mathrm{succ}} \ge 1$, the residual-variance term without this factor is no larger than the weighted term. Adding the three bounds proves the support-dependent claim. This refinement changes only the analysis; the algorithm does not need to know $\mathcal{N}(s,a)$. □

The next summation is the standard MVP doubling argument underlying the bonus bound in Zhang et al. (2021, Lemma 9). We restate it for safe prefixes and arbitrary predictable weights.

Lemma G.8 (Dyadic count summation)**.** *Index the transitions in safe prefixes by $t = (q,h)$, and let n_t be the active dyadic snapshot of their row. For arbitrary nonnegative predictable weights w_t, simultaneously over local prefixes,*

$$\sum_t \frac{1}{n_t} \le \widetilde{O}(|\mathcal{S}|\,|\mathcal{A}|),$$

$$\sum_t \sqrt{\frac{w_t}{n_t}} \le \widetilde{O}\left(\sqrt{|\mathcal{S}|\,|\mathcal{A}|\sum_t w_t}\right),$$

$$\sum_t \frac{1}{\sqrt{n_t}} \le \widetilde{O}(\sqrt{|\mathcal{S}|\,|\mathcal{A}|nH}).$$

Proof. Group visits by row and dyadic level $n_t = 2^j$. After cut episodes are stopped at their predictable trigger, a fixed row-level group contains at most 2^j safe visits before the next snapshot

becomes active. Its contribution to the inverse-count sum is therefore at most one. There are $O(|\mathcal{S}|\,|\mathcal{A}|\log(nH))$ groups. Within a group G, Cauchy–Schwarz gives

$$\sum_{t\in G}\sqrt{w_t/2^j} \le \sqrt{\sum_{t\in G} w_t}.$$

A second Cauchy–Schwarz inequality over row-level groups proves the weighted statement. Taking $w_t = 1$ and using at most nH transitions proves the last claim. □

For a local prefix, let $\mathcal{T}_{\text{safe}} := \{(q,h) : h < \sigma_q\}$ and define

$$M_2 := \sum_{(q,h)\in\mathcal{T}_{\text{safe}}} \beta_{q,h},$$
$$M_4 := \sum_{(q,h)\in\mathcal{T}_{\text{safe}}} \mathrm{Var}_{P_{\tau_q}}(\widetilde{V}_{q,h+1})(S_{\tau_q,h}, A_{\tau_q,h}),$$
$$M_5 := \sum_{(q,h)\in\mathcal{T}_{\text{safe}}} \mathrm{Var}_{P_{\tau_q}}(\widetilde{V}_{q,h+1} - \bar{V}^{\star}_{\tau_q,h+1})(S_{\tau_q,h}, A_{\tau_q,h}).$$

The next argument follows Zhang et al. (2021, Lemma 16). We apply the variance recursion once to the optimistic values and once to the gap between optimistic and optimal values.

Lemma G.9 (Cumulative variance bounds)**.** *For every local prefix of length n with hull J, in the nontrivial regime $HD(J) \le 1$,*

$$M_4 \le \widetilde{O}(n + M_2 + |\mathcal{S}|\,|\mathcal{A}|), \qquad M_5 \le \widetilde{O}(M_2 + n\Gamma(J) + |\mathcal{S}|\,|\mathcal{A}|).$$

Proof. The first step applies the variance recursion to the optimistic values. The second step shifts the optimism gap so that it is nonnegative and predictable, then applies the same recursion again.

Step 1: Control the optimistic-value variance M_4.

Condition on the complete kernel path. Let $\mathcal{H}_{q,0}$ be the history after the algorithm selects episode τ_q and before it observes the trajectory. Extend this history one transition at a time. The next state at each stage is drawn from P_{τ_q}, while the active copy and its value functions were fixed before the episode began. They are therefore predictable for this trajectory.

The first inequality in Lemma G.7 has exactly the form required by Lemma G.6. Apply that lemma with $B = 1$, $a_{q,h} = \bar{r}_h(S_{\tau_q,h}, A_{\tau_q,h}) + \beta_{q,h}$, and stopping stage σ_q. Normalized rewards sum to at most one per episode, and the number of cut episodes is $\widetilde{O}(|\mathcal{S}|\,|\mathcal{A}|)$. Substituting these two facts proves the bound on M_4.

Step 2: Control the shifted-gap variance M_5.

To control M_5, we shift the optimism gap so that it is nonnegative. The shift must be known before the episode starts, so we define it using only the prefix up to episode q:

$$d_q := D(J_q), \qquad e_{q,h} := c_\star(H-h)_+ d_q, \qquad G_{q,h} := \widetilde{V}_{q,h} - \bar{V}^{\star}_{\tau_q,h} + e_{q,h}\mathbf{1}.$$

The interval $J_q = [\tau_1, \tau_q]$ is known before episode τ_q. Thus d_q and $G_{q,h}$ are predictable. We also have $d_q \le D(J)$ and $HD(J) \le 1$. Approximate optimism from Proposition G.5 makes $G_{q,h}$ nonnegative and bounded by a universal constant. Combining the optimal Bellman inequality with Lemma G.7 gives

$$G_{q,h}(S_{\tau_q,h}) \le P_{\tau_q} G_{q,h+1}(S_{\tau_q,h}, A_{\tau_q,h}) + \beta_{q,h} + e_{q,h} - e_{q,h+1}.$$

The shift changes by at most $c_\star d_q$ between adjacent stages, and adding this constant does not change variance. Apply Lemma G.6 again, now with $a_{q,h} = \beta_{q,h} + c_\star d_q$. We obtain

$$M_5 \le \widetilde{O}\left(M_2 + H\sum_{q=1}^{n} d_q + |\mathcal{S}|\,|\mathcal{A}|\right) \le \widetilde{O}\left(M_2 + nHD(J) + |\mathcal{S}|\,|\mathcal{A}|\right).$$

Because $HD(J) \le 1$, the last term is $n\Gamma(J)$. The shifts $e_{q,h}$ are used only in the proof. Most importantly, d_q depends only on the history available before episode q; it does not use the final prefix length n. The same predictable construction therefore works for every prefix. Removing the conditioning on the kernel path preserves the probability of the event. □

The final accounting follows the stationary MVP decomposition in Zhang et al. (2021, Lemmas 7–10), with the additional diameter and support-dependent terms proved above.

Lemma G.10 (Cumulative MVP error)**.** *For every local prefix of n active episodes with enclosing interval J, the raw safe-prefix Bellman charge satisfies the unrestricted-support and support-dependent bounds*

$$M_2 \le \widetilde{O}\left(\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|^2|\mathcal{A}| + n\Gamma(J)\right),$$
$$M_2 \le \widetilde{O}\left(\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}| + n\Gamma(J)\right).$$

Both hold whenever $\Gamma(J) < 1$, and are simultaneous over all maintained copies and local prefixes. No $\wedge n$ truncation is asserted for this raw charge.

Proof. We first sum the one-step error bounds over all safe transitions. The resulting right-hand side contains M_2 through the variance terms. Young's inequality moves these terms back to the left. We then repeat the calculation using the smaller next-state support.

Step 1: Aggregate the unrestricted-support one-step charges.

Sum the charge in Lemma G.7 and apply Lemma G.8. With $d = D(J)$,

$$M_2 \le \widetilde{O}\Big(\sqrt{|\mathcal{S}|\,|\mathcal{A}|M_4} + \sqrt{|\mathcal{S}|^2|\mathcal{A}|M_5} + |\mathcal{S}|^2|\mathcal{A}| + |\mathcal{S}|\sqrt{|\mathcal{A}|nHd} + nHd + \sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|\,|\mathcal{A}|\Big).$$

The fourth term comes from summing $\sqrt{|\mathcal{S}|d/n_t}$. The fifth comes from the direct diameter penalty d at each of at most nH transitions.

Step 2: Close the unrestricted-support self-bound.

Insert the two variance bounds from Lemma G.9. The $|\mathcal{S}|\,|\mathcal{A}|$ terms created by stopped prefixes contribute at most $|\mathcal{S}|\,|\mathcal{A}| + |\mathcal{S}|^{3/2}|\mathcal{A}| \le 2|\mathcal{S}|^2|\mathcal{A}|$, so the startup term absorbs them. For the terms that contain M_2, Young's inequality gives

$$\sqrt{|\mathcal{S}|\,|\mathcal{A}|M_2} + \sqrt{|\mathcal{S}|^2|\mathcal{A}|M_2} \le \tfrac{1}{2}M_2 + \widetilde{O}(|\mathcal{S}|^2|\mathcal{A}|)$$

after adjusting constants, while

$$|\mathcal{S}|\sqrt{|\mathcal{A}|nHd} \le \tfrac{1}{2}|\mathcal{S}|^2|\mathcal{A}| + \tfrac{1}{2}nHd, \qquad \sqrt{|\mathcal{S}|^2|\mathcal{A}|n\Gamma(J)} \le \tfrac{1}{2}|\mathcal{S}|^2|\mathcal{A}| + \tfrac{1}{2}n\Gamma(J).$$

The other two applications of Young's inequality absorb the mixed mismatch terms. Move the resulting half of M_2 to the left-hand side. This proves the unrestricted-support bound. All variances in this calculation are taken under the actual episode kernel P_{τ_q}.

Step 3: Repeat the accounting with bounded support.

For the support-dependent bound, summing the corresponding one-step inequality gives

$$M_2 \le \widetilde{O}\big(\sqrt{|\mathcal{S}|\,|\mathcal{A}|M_4} + \sqrt{d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|M_5} + d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}| + \sqrt{d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|nHd} + nHd + \sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|\,|\mathcal{A}|\big).$$

The weighted dyadic-count inequality gives the first two terms. The inverse-count terms sum to $\widetilde{O}(d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|)$, while

$$\sqrt{d_{\mathrm{succ}}d}\sum_t n_t^{-1/2} \le \widetilde{O}\Big(\sqrt{d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|nHd}\Big).$$

Substituting Lemma G.9 and using

$$\sqrt{|\mathcal{S}|\,|\mathcal{A}|M_2} + \sqrt{d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|M_2} \le \tfrac{1}{2}M_2 + \widetilde{O}(d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|),$$
$$\sqrt{d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|n\Gamma(J)} \le \tfrac{1}{2}d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}| + \tfrac{1}{2}n\Gamma(J),$$

Since $d_{\mathrm{succ}} \ge 1$ and $Hd \le \Gamma(J)$ in this regime, these inequalities absorb all remaining mixed terms. The stopped-prefix terms contribute at most $|\mathcal{S}|\,|\mathcal{A}| + |\mathcal{S}|\,|\mathcal{A}|\sqrt{d_{\mathrm{succ}}} \le 2d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}|$. Moving the remaining fraction of M_2 to the left proves the support-dependent bound. □

G.3 Assembly of Proposition G.2

We now combine the preceding ingredients. Approximate optimism controls the gap to the current optimal value, while the Bellman and martingale bounds control the difference between the prediction and the realized return.

Step 1: Establish approximate optimism.

Fix a local prefix and condition on Ω_{MVP}. Approximate optimism follows from Proposition G.5. Monotonicity in the interval then gives the first bound.

Step 2: Control prediction error against observed returns.

First consider $\Gamma(J) < 1$. For each episode q, define

$$\xi_{q,0} = \widetilde{V}_{q,1}(\rho) - \widetilde{V}_{q,1}(S_{\tau_q,1}), \qquad \xi_{q,h} = P_{\tau_q}\widetilde{V}_{q,h+1} - \widetilde{V}_{q,h+1}(S_{\tau_q,h+1}),$$

where the second expression is evaluated at the visited state-action pair. Both the initial-state term and the stopped transition terms are martingale differences. Summing the Bellman inequality in Lemma G.7 over the safe prefix gives

$$\widetilde{f}_q - Y_{\tau_q} \le \mathbf{1}\{\sigma_q \le H\} + \sum_{h<\sigma_q} \beta_{q,h} + \xi_{q,0} + \sum_{h<\sigma_q} \xi_{q,h}.$$

If the episode is not cut, this is the usual full-horizon telescoping sum. If it is cut, the sum stops at stage $\sigma_q - 1$. The remaining value is at most one, and the omitted future rewards are nonnegative. The same bound holds when $\sigma_q = 1$; then only $\xi_{q,0}$ and the unit cut charge remain.

The transition part of the martingale has predictable variance at most M_4, and the initial-state part contributes at most n. Freedman's inequality, Lemma G.9, and Young's inequality yield

$$\sum_{q=1}^{n}\xi_{q,0} + \sum_{(q,h)\in\mathcal{T}_{\mathrm{safe}}} \xi_{q,h} \le \widetilde{O}(\sqrt{n + M_4} + 1) \le \tfrac{1}{4}M_2 + \widetilde{O}(\sqrt{n} + \sqrt{|\mathcal{S}|\,|\mathcal{A}|} + 1).$$

Combining this with Lemma G.10 and the unit terminal charge for each of the $\widetilde{O}(|\mathcal{S}|\,|\mathcal{A}|)$ cut episodes proves the raw observed bound

$$\sum_{q=1}^{n}(\widetilde{f}_q - Y_{\tau_q}) \le \widetilde{O}(\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + |\mathcal{S}|^2|\mathcal{A}| + n\Gamma(J)).$$

The support-dependent part of Lemma G.10, together with the same martingale and cut episode bounds, yields on the same event

$$\sum_{q=1}^{n}(\widetilde{f}_q - Y_{\tau_q}) \le \widetilde{O}\Big(\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + d_{\mathrm{succ}}|\mathcal{S}|\,|\mathcal{A}| + n\Gamma(J)\Big).$$

Each summand is at most one, so the sum is also at most n. Taking the smaller bound gives $\widetilde{O}(C_{\mathrm{MVP}}(n) + n\Gamma(J))$. When $\Gamma(J) = 1$, the bound follows directly from this per-episode cap. Applying the same truncation to the support-dependent display gives $\widetilde{O}(C_{\deg}(n) + n\Gamma(J))$.

Step 3: Transfer to expected returns and all copies.

Finally, $Y_{\tau_q} - f_{\tau_q}(\pi_q)$ is a bounded martingale difference because the active copy and its policy are fixed before the trajectory is sampled. Freedman's inequality contributes $\widetilde{O}(\sqrt{n})$ on every prefix, which is absorbed by either complexity bound. A union bound covers all copies created by MASTER. Pausing a copy changes only the waiting time between its updates; it does not change its local data or filtration. This proves Proposition G.2. A contiguous interval is a special case, so the same result proves Proposition G.1. The algorithm uses only empirical counts and values and does not need the sets $\mathcal{N}(s,a)$. □

H Proof of Theorem 4.1

The proof follows the algorithm. We first bound regret inside one restart block. We then sum over blocks, average the amount of kernel change, and choose the window length.

Step 1: Bound regret on each restart block.

Fix a block B. The base guarantee controls two gaps. Approximate optimism gives $f_k^\star - \widetilde{f}_k \le c\Gamma(B)$ for every $k \in B$, while prediction error satisfies

$$\sum_{k\in B}(\widetilde{f}_k - f_k(\pi_k)) \le c\,(C(|B|) + |B|\Gamma(B))\,.$$

Add these two gaps, sum over all blocks, and multiply by H to return to the original reward scale. This gives

$$\mathrm{Reg}_K \le cH \sum_{B\in\mathcal{B}_W} (C(|B|) + |B|\Gamma(B))\,.$$

Step 2: Average the block diameters under the displacement condition.

There are at most $K/W + 1$ blocks. By Lemma F.1, each block satisfies $\mathbb{E}[\Gamma(B)] \le \widetilde{O}(\min\{1, H\sqrt{\nu W}\})$. We run every copy with failure probability K^{-3}. Taking expectations and applying a union bound adds only $O(H/K)$, so

$$\mathbb{E}[\mathrm{Reg}_K] \le \widetilde{O}\left(H\left(\frac{K}{W}+1\right)C(W) + HK\min\{1, H\sqrt{\nu W}\}\right).$$

Step 3: Apply the caps and conclude.

Regret is always at most HK. Use this cap, replace $\min\{1, H\sqrt{\nu W}\}$ by $H\sqrt{\nu W}$, and substitute C_{MVP} from Proposition G.1. This gives the stronger intermediate bound for any W.

To choose W, balance the statistical term $HK\sqrt{|\mathcal{S}|\,|\mathcal{A}|/W}$ with the mismatch term $H^2K\sqrt{\nu W}$. Their balance point is $\sqrt{|\mathcal{S}|\,|\mathcal{A}|}/(H\sqrt{\nu})$, clipped to $[1,K]$, which is exactly W_ν. In the interior regime, each term equals $KH^{3/2}(|\mathcal{S}|\,|\mathcal{A}|)^{1/4}\nu^{1/4}$. The $|\mathcal{S}|^2|\mathcal{A}|$ part of C_{MVP} contributes the startup term $\widetilde{O}(H|\mathcal{S}|^2|\mathcal{A}|K/W_\nu)$. This proves Theorem 4.1. □

I Proof of Corollary 4.3

The corollary follows by inserting the support-dependent MVP complexity into the known-rate block argument. The only point to check is that its additive term is absorbed by the square-root term at every sample size.

Proof. If $d_{\mathrm{succ}} \le d_0$, the support-dependent bound in Proposition G.2 gives

$$C_{\deg}(n) = \widetilde{O}\Big((\sqrt{|\mathcal{S}|\,|\mathcal{A}|n} + d_0|\mathcal{S}|\,|\mathcal{A}|) \wedge n\Big) = \widetilde{O}_{d_0}\Big(n \wedge \sqrt{|\mathcal{S}|\,|\mathcal{A}|n}\Big)\,.$$

The second equality follows by splitting at $n = |\mathcal{S}|\,|\mathcal{A}|$: below that scale the cap n applies, and above it $d_0|\mathcal{S}|\,|\mathcal{A}| \le d_0\sqrt{|\mathcal{S}|\,|\mathcal{A}|n}$. This dependence parallels the transition-support dependence in MVP-V

(Zhou et al., 2023, Theorem 7); Proposition G.2 proves the corresponding changing-kernel guarantee. Applying the three-step block argument in section H with this complexity function and choosing $W = W_\nu$ gives the stated upper bound. The matching lower bound is Theorem 4.2, whose hard kernels have next-state support size two. □

J Proof of Theorem 4.2

Both terms in the lower bound come from the same survival MDP. In this MDP, a wrong action makes absorption more likely and therefore loses many future rewards. We first state the testing fact used to hide the best action. We then construct the MDP, verify Assumption 2.1, and bound how much data the agent sees between refreshes. Random refreshes give the shift-dependent term. Setting the refresh probability to zero gives the stationary term.

Step 1: Record the testing inequality.

The next lemma is the standard one-good-action testing argument, but it allows the transcript to stop at a random time. It follows from the adaptive KL decomposition and Pinsker's inequality (Lattimore and Szepesvári, 2020, Lemma 15.1 and Eq. (14.12)). We include the short proof to show that revealed side information and the stopping rule do not leak the hidden action under the reference model.

Lemma J.1 (One-good-action testing). *Let $J \sim \mathrm{Unif}([|\mathcal{A}|])$, and let an adaptive agent observe a transcript of random length N. Conditional on $J = j$, action j produces observations from Q_1, while every other action produces observations from Q_0. Under the reference law $\mathbb{P}_0$, every action produces observations from Q_0. The agent may use arbitrary side information and an arbitrary stopping rule. Under $\mathbb{P}_0$, both must be independent of J. If $D_{\mathrm{KL}}(Q_0\|Q_1) \le \alpha$, then every transcript-measurable estimator $\widehat{J}$ satisfies*

$$\mathbb{P}(\widehat{J} = J) \le \frac{1}{|\mathcal{A}|} + \sqrt{\frac{\alpha \mathbb{E}_0[N]}{2|\mathcal{A}|}}.$$

Proof. Let $\mathbb{P}_j$ be the law of the transcript when $J = j$, and let N_j count how often action j is chosen. The adaptive KL chain rule gives $D_{\mathrm{KL}}(\mathbb{P}_0\|\mathbb{P}_j) \le \alpha \mathbb{E}_0[N_j]$. Pinsker's inequality then gives

$$\begin{aligned}\mathbb{P}(\widehat{J} = J) &\le \frac{1}{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{A}|} \Big(\mathbb{P}_0(\widehat{J} = j) + d_{\mathrm{TV}}(\mathbb{P}_j, \mathbb{P}_0)\Big) \\ &\le \frac{1}{|\mathcal{A}|} + \sqrt{\frac{1}{2|\mathcal{A}|} \sum_{j=1}^{|\mathcal{A}|} D_{\mathrm{KL}}(\mathbb{P}_0\|\mathbb{P}_j)} \le \frac{1}{|\mathcal{A}|} + \sqrt{\frac{\alpha \mathbb{E}_0[N]}{2|\mathcal{A}|}},\end{aligned}$$

where we used Pinsker, Jensen, and $\sum_j N_j = N$. □

Step 2: Construct the survival MDP and verify the model.

Let $n = |\mathcal{S}| - 1$ and $m = n|\mathcal{A}|$. The MDP has n alive states $x_1, \dots, x_n$ and one absorbing bad state b. At the start of episode k, choose an alive state uniformly: $X_k \sim \mathrm{Unif}([n])$ and $S_{k,1} = x_{X_k}$. Each alive state x has a hidden best action $U_{k,x} \in [|\mathcal{A}|]$. Set $p = 1/(8H)$ and choose $0 < \Delta \le 1/(128H)$. The transitions are

$$P_k(b \mid x, a) = p + \Delta \mathbf{1}\{a \ne U_{k,x}\}, \qquad P_k(x \mid x, a) = 1 - p - \Delta \mathbf{1}\{a \ne U_{k,x}\},$$

and b is absorbing. Rewards are fixed: $r_h(x, a) = 1$ and $r_h(b, a) = 0$.

Lemma J.2 (Action gap and survival). *For the survival MDP, the optimal alive-state value is $V_h^\star(x) = \sum_{t=0}^{H-h}(1-p)^t$. For every wrong action $a \ne U_{k,x}$,*

$$Q_h^\star(x, U_{k,x}) - Q_h^\star(x, a) = \Delta V_{h+1}^\star(x) \ge \frac{7}{8}\Delta(H - h).$$

Moreover, under every agent and every stage, $\mathbb{P}(S_{k,h} \ne b) \ge 111/128$.

Proof. The optimal policy always chooses the hidden best action and therefore moves to b with probability p. Summing its survival probabilities gives the formula for $V_h^\star(x)$. A wrong action adds Δ to the probability of immediate absorption, so it loses $\Delta V_{h+1}^\star(x)$. Also,

$$V_{h+1}^\star(x) \geq (H-h)(1-p)^H \geq (H-h)(1-Hp) = \frac{7}{8}(H-h),$$

which proves the gap bound. Under any action sequence, the hazard is at most $p + \Delta \leq 17/(128H)$. Bernoulli's inequality then gives $\mathbb{P}(S_{k,h} \neq b) \geq (1-p-\Delta)^H \geq 111/128$. □

Initially, draw every $U_{1,x}$ independently and uniformly from $[|\mathcal{A}|]$. Before each later episode, draw one refresh indicator $B_k \sim \mathrm{Ber}(q)$. If $B_k = 0$, keep all best actions unchanged. If $B_k = 1$, redraw every best action independently and uniformly. We reveal all refresh times to the agent. This extra information can only make the problem easier.

Lemma J.3 (Validity under diffusive distribution shift)**.** *The global-refresh survival process satisfies*

$$\mathbb{E}\Big[\|P_{k+\ell} - P_k\|_{\infty,\mathrm{TV}}^2 \mid \mathcal{E}_k\Big] \leq q\Delta^2\ell.$$

Thus it satisfies Assumption 2.1 whenever $q\Delta^2 \leq \nu$.

Proof. If no refresh occurs during the next ℓ episodes, then $P_{k+\ell} = P_k$. If a refresh occurs, each transition row changes by at most Δ in TV. A union bound places the probability of at least one refresh below $\min\{1, q\ell\}$. Therefore

$$\mathbb{E}\Big[\|P_{k+\ell} - P_k\|_{\infty,\mathrm{TV}}^2 \mid \mathcal{E}_k\Big] \leq \Delta^2 \min\{1, q\ell\} \leq \min\{1, q\Delta^2\ell\}.$$

This is at most $\min\{1, \nu\ell\}$ whenever $q\Delta^2 \leq \nu$. □

Step 3: Bound the information available between refreshes.

Lemma J.4 (Expected number of informative observations)**.** *Fix episode k, stage h, and condition on $X_k = x$. Under the null reference law in which every action has hazard $p + \Delta$, let $N_{k,h,x}$ be the number of hazard observations from context x since its current global-refresh time and strictly before action $A_{k,h}$. Then*

$$\mathbb{E}_0[N_{k,h,x}] \leq \frac{H}{nq} + H.$$

Proof. Data from an earlier episode $t < k$ describe the current best action only if no refresh occurs from episode $t+1$ through k. This has probability $(1-q)^{k-t}$. Episode t starts from state x with probability $1/n$ and supplies at most H hazard observations. Summing over past episodes gives at most $H/(nq)$ observations in expectation. Earlier stages of the current episode add at most H. □

Lemma J.5 (Constant action-error probability before absorption)**.** *There is a universal constant $c_{\mathrm{err}} > 0$ such that the following holds. Choose $q = 2048H^2\Delta^2/m$ with $\Delta \leq 1/(128H)$. Then, at every episode and every stage,*

$$\mathbb{P}(S_{k,h} \neq b,\ A_{k,h} \neq U_{k,X_k}) \geq c_{\mathrm{err}}.$$

Proof. We reduce action choice at one state and stage to the one-good-action test. The information bound limits the probability of finding the hidden best action; the survival bound then converts this testing error into a constant probability of taking a wrong action before absorption.

Condition on $X_k = x$ and write $J := U_{k,x}$, but do not condition on the age of the current refresh segment. The agent knows the refresh times and everything observed before this segment. Under the reference model, this side information is independent of the newly drawn J. Earlier segments also have the same law for every possible J. Thus only the $N_{k,h,x}$ hazard observations from the current segment can identify the current best action. Its expectation already averages over the random age of the segment.

In the reference model, every action has hazard $p + \Delta$. In model j, action j instead has hazard p. The KL divergence is at most the chi-square divergence (Lattimore and Szepesvári, 2020, Eq. (14.16)).

One observation therefore contributes at most $\Delta^2/[p(1-p)] \le 16H\Delta^2$. Combining this fact with Lemmas J.1 and J.4 gives

$$\epsilon := \sqrt{\frac{\alpha \mathbb{E}_0[N_{k,h,x}]}{2|\mathcal{A}|}} \le \sqrt{\frac{8H^2\Delta^2}{mq} + \frac{8H^2\Delta^2}{|\mathcal{A}|}} \le \sqrt{\frac{1}{256} + \frac{1}{2048|\mathcal{A}|}} < 0.07.$$

Define $\widehat{J} = A_{k,h}$ unless the agent has already reached b. On the absorbed event, let $\widehat{J}$ be an independent uniform action. Then $\mathbb{P}(\widehat{J} = J) \le 1/|\mathcal{A}| + \epsilon$, while the absorbed event contributes exactly $1/|\mathcal{A}|$ success probability. Therefore

$$\mathbb{P}(S_{k,h} \ne b, A_{k,h} \ne J) \ge \mathbb{P}(S_{k,h} \ne b)\left(1 - \frac{1}{|\mathcal{A}|}\right) - \epsilon \ge \frac{111}{256} - 0.07,$$

which is a positive universal constant. □

Step 4: Derive the shift-dependent term.

If $\nu = 0$, the shift-dependent term is zero. For $\nu > 0$, set

$$\Delta_{\mathrm{dyn}} := \min\left\{\frac{1}{128H}, \left(\frac{m\nu}{2048H^2}\right)^{1/4}\right\}, \qquad q := 2048\frac{H^2\Delta_{\mathrm{dyn}}^2}{m}.$$

If the second branch is active, then $q\Delta_{\mathrm{dyn}}^2 = \nu$ and $\Delta_{\mathrm{dyn}} \le 1/(128H)$ implies $q \le 1/(8m) \le 1$. If the cap is active, then $q = 1/(8m)$, and the cap-activation inequality gives $q\Delta_{\mathrm{dyn}}^2 \le \nu$. Thus Lemma J.3 applies. By Lemmas J.2, C.4 and J.5, the regret in any episode is at least

$$c_{\mathrm{err}} \sum_{h=1}^{H-1} \frac{7}{8}\Delta_{\mathrm{dyn}}(H-h) \ge cH^2\Delta_{\mathrm{dyn}}.$$

Summing over episodes and substituting the definition of Δ_{dyn} gives

$$\mathbb{E}[\mathrm{Reg}_K] \ge cK \min\left\{H, H^{3/2}m^{1/4}\nu^{1/4}\right\}.$$

Step 5: Derive the stationary term.

Set $q = 0$, so the MDP is stationary and belongs to the model class for every ν. Choose

$$\Delta_{\mathrm{stat}} := \min\left\{\frac{1}{128H}, \sqrt{\frac{m}{2048H^2K}}\right\}.$$

Fix episode k, stage h, and current context x. Under the null reference law, the transcript before this action contains at most $H(k-1)/n + H \le HK/n + H$ hazard observations in expectation. With $\alpha \le 16H\Delta_{\mathrm{stat}}^2$, the testing error satisfies

$$\epsilon_{\mathrm{stat}}^2 \le 8H^2\Delta_{\mathrm{stat}}^2\left(\frac{K}{m} + \frac{1}{|\mathcal{A}|}\right).$$

If the cap is active, then $K \le 8m$, so

$$\epsilon_{\mathrm{stat}}^2 \le \frac{K}{2048m} + \frac{1}{2048|\mathcal{A}|} \le \frac{1}{256} + \frac{1}{4096}.$$

If the square-root branch is active, then $K \ge 8m$, so

$$\epsilon_{\mathrm{stat}}^2 \le \frac{1}{256} + \frac{m}{256|\mathcal{A}|K} \le \frac{1}{256} + \frac{1}{2048|\mathcal{A}|}.$$

In both cases $\epsilon_{\mathrm{stat}} < 0.07$. The same randomization-after-absorption argument as in Lemma J.5, together with Lemma J.2, gives a universal lower bound on the probability of an action error before absorption at every stage. Thus the expected regret per episode is at least $cH^2\Delta_{\mathrm{stat}}$, and multiplying by K gives $cH\min\{K, \sqrt{mK}\}$.

Step 6: Combine the two terms.

Both constructions satisfy Assumption 2.1 with parameter at most ν. The worst-case regret is therefore at least the larger of the shift-dependent and stationary bounds, which is at least half their sum. Since $m = (|\mathcal{S}|-1)|\mathcal{A}| \geq |\mathcal{S}|\,|\mathcal{A}|/2$, we may replace m by $|\mathcal{S}|\,|\mathcal{A}|$ after changing the universal constant. The agent was given ν and every refresh time, so the same lower bound holds when ν is known. Each nonabsorbing row has only the two possible next states $\{x, b\}$, and the absorbing row has one. Thus $d_{\mathrm{succ}} = 2$. This proves Theorem 4.2. □

K Proof of Theorem 5.1

We prove Theorem 5.1 for both unrestricted kernels and the bounded-successor class. If ν were known, the agent would use one window length. MASTER does not know this length, so it tries several lengths by running short copies inside longer blocks. These extra copies create a replay cost. We first compare the algorithm with any fixed window. We then show that a replay can be costly only when the kernel changes enough to justify it. The displacement condition controls how often this happens. We use the scheduling rule from Lemma C.2; all arguments about changing kernels are proved below.

Fix either complexity bound from Proposition G.2. Choose one deterministic version of its logarithmic factors and write

$$C(n) := C_{\mathrm{base}}(n) = (a\sqrt{n} + b) \wedge n,$$
$$\eta(n) := \frac{C(n)}{n} = \min\left\{1, \frac{a}{\sqrt{n}} + \frac{b}{n}\right\},$$
$$a = \widetilde{\Theta}(\sqrt{|\mathcal{S}|\,|\mathcal{A}|}), \qquad b = \begin{cases} \widetilde{\Theta}(|\mathcal{S}|^2|\mathcal{A}|), & \text{for } C_{\mathrm{MVP}}, \\ \widetilde{\Theta}(d_0|\mathcal{S}|\,|\mathcal{A}|), & \text{for } C_{\mathrm{deg},0}. \end{cases}$$

We fix the logarithmic factors once for the whole proof. With this choice, C is nondecreasing and η is nonincreasing over dyadic scales. MASTER uses a slightly larger threshold $\widehat{\eta}(n)$, differing only by logarithmic factors. In the bounded-support case, the wrapper knows the numerical bound d_0, but it does not know the support sets $\mathcal{N}(s,a)$.

K.1 Executable adaptive wrapper

For a block $B = [s, s+N-1]$, let $\mathcal{D}_N = \{1, 2, 4, \ldots, N\}$ and partition B into aligned cells $I_{q,j} = [s+jq, s+(j+1)q-1]$, where $q \in \mathcal{D}_N$. A fresh MVP copy starts on each cell independently with probability $p_{N,q} = \eta(N)/\eta(q)$; the length-$N$ copy always starts. In each episode, the shortest available copy supplies the policy and prediction.

Let $g_t \in [0,1]$ be the active copy's predicted value, let $Y_t = H^{-1}\sum_{h=1}^{H} r_h(s_{t,h}, a_{t,h})$ be its realized return, and set $U_t = \min_{s \leq r \leq t} g_r$. A completed length-$q$ copy on $[u, u+q-1]$ and the current block prefix $[s,t]$ trigger the two MASTER tests

$$\begin{aligned} \frac{1}{q}\sum_{r=u}^{u+q-1} Y_r &\geq U_{u+q-1} + c_1\widehat{\eta}(q), \quad \text{(short-window comparison)}, \\ \frac{1}{t-s+1}\sum_{r=s}^{t}(g_r - Y_r) &\geq c_2\widehat{\eta}(t-s+1), \quad \text{(prediction-error test)}, \end{aligned} \tag{2}$$

where c_1, c_2 are universal constants. Either test ends the current epoch. These are the tests used by Algorithm 2; the schedule is Algorithm 3 of Wei and Luo (2021), instantiated with independent MVP copies. We prove the guarantee of this wrapper for either complexity bound above.

K.2 Comparison bound and theorem assembly

Fix any window length W for the analysis. The algorithm does not know or use W. We compare its regret with three costs: learning inside each window, changes of the kernel inside each window, and

the extra replays used to adapt. The first two are the same costs as in the known-rate proof. We only need to show that the replay cost is no larger. The next three steps give the comparison bound, take expectations over the random kernel path, and evaluate the replay sum.

Step 1: Prove the simultaneous comparison bound.

At one episode and one dyadic scale, at most one aligned cell can start. The algorithm therefore creates at most $K(1+\log_2 K)$ copies. It also tests only polynomially many intervals, block scales, and prefixes. A union bound over all of them changes only the logarithmic factors in L and $\widehat{\eta}$. After increasing their constants, $\mathbb{P}(\Omega_{\mathrm{ad}}) \geq 1-\delta$.

Let $\mathfrak{B}$ be the random collection of realized MASTER blocks, and let N_B be the planned length of B. Define

$$\mathcal{R}_{\mathrm{replay}} := \sum_{B\in\mathfrak{B}} \sum_{q\in\{1,2,4,\ldots,N_B\}} \frac{q\eta(q)^2}{\eta(N_B)} \mathbf{1}\{\Gamma(\overline{B}) \geq c_{\mathrm{loc}}\eta(q)\}, \qquad \mathcal{D}_n := \{1,2,4,\ldots,2^{\lfloor \log_2 n \rfloor}\}.$$

The following comparison holds on Ω_{ad}.

Lemma K.1 (Window oracle bound)**.** *For every deterministic comparison window $W \in [K]$, on Ω_{ad},*

$$\sum_{t=1}^{K}(f_t^\star - Y_t) \leq \widetilde{O}\left[\left(\frac{K}{W}+1\right)C(W) + \sum_{B_0\in\mathcal{B}_W} |B_0|\Gamma(B_0) + \mathcal{R}_{\mathrm{replay}}\right],$$

where $\mathcal{B}_W$ is the consecutive length-W partition.

Run the algorithm with $\delta = K^{-3}$. For the analysis, fix any deterministic dyadic window W. By Lemma K.1, only two random quantities remain: the kernel diameter inside the comparison blocks and the replay cost.

Step 2: Average diameter and replay charges under diffusive distribution shift.

Extend the kernel path past episode K by setting $P_t = P_K$ and $\mathcal{F}_t = \mathcal{F}_K$. For a random starting time τ, let $I_{\tau,N} = \{\tau,\ldots,\tau+N-1\}$. The stopping-time part of Lemma F.1 bounds both the expected diameter of such an interval and the chance that this diameter is large. In particular, every comparison block satisfies $\mathbb{E}[\Gamma(B_0)] \leq \widetilde{O}(\min\{1, H\sqrt{\nu W}\})$. Define $v := c_D H^2 \nu \log^2(2K)$, where c_D is the constant from that lemma. The next lemma uses this bound to control all replay charges.

Lemma K.2 (Expected replay cost)**.** *With*

$$\mathfrak{S}_K(a,b,v) := \sum_{N\in\mathcal{D}_K} \frac{1}{N\eta(N)} \sum_{q\in\mathcal{D}_N} q\min\{\eta(q)^2, vN\},$$

the replay cost satisfies $\mathbb{E}[\mathcal{R}_{\mathrm{replay}}] \leq \widetilde{O}(K\mathfrak{S}_K(a,b,v))$.

Take expectations in Lemma K.1. The observed return satisfies $\mathbb{E}[Y_t \mid \pi_t, P_t, \mathrm{past}] = f_t(\pi_t)$, and the total normalized regret on the failure event is at most $K\delta$. Hence the expected normalized regret is at most

$$\widetilde{O}\left[\left(\frac{K}{W}+1\right)C(W) + K\min\{1, H\sqrt{\nu W}\} + K\mathfrak{S}_K(a,b,v)\right]. \tag{3}$$

Step 3: Evaluate the replay sum and choose the window.

Define $W_v := \min\{K, \max\{1, a/\sqrt{v}\}\}$, with $a/\sqrt{0} = +\infty$. The next lemma evaluates the replay sum for the concrete function $C(n) = (a\sqrt{n}+b)\wedge n$.

Lemma K.3 (Evaluation of the replay cost)**.** *If $a/\sqrt{W_v} + b/W_v + \sqrt{vW_v}$ is below a sufficiently small universal constant, then*

$$\mathfrak{S}_K(a,b,v) \leq \widetilde{O}\left(\frac{a}{\sqrt{W_v}} + \frac{b}{W_v} + \sqrt{vW_v}\right).$$

Outside this regime, the normalized regret theorem follows from the trivial cap K.

The two replay lemmas bound the expected replay cost at window W_v. Since $v = c_D H^2 \nu \log^2(2K)$, the lengths W_v and W_ν differ only by logarithmic factors. Therefore

$$\mathbb{E}[\mathcal{R}_{\text{replay}}] \le \widetilde{O}\left[K\left(\frac{a}{\sqrt{W_\nu}} + \frac{b}{W_\nu} + H\sqrt{\nu W_\nu}\right)\right].$$

If the bracket is larger than a universal constant, the normalized regret cap K is already enough. Otherwise, the cap in $C(W_\nu)$ is inactive. In that case, $K(a/\sqrt{W_\nu} + b/W_\nu) = (K/W_\nu)C(W_\nu)$. The replay cost is therefore of the same order as the learning and mismatch costs at $W = W_\nu$. Multiplying by H returns to the original reward scale. The final two subsections insert the two choices of C.

K.3 Paused MVP copies

This subsection shows that the MVP guarantee survives the MASTER schedule. Although short copies interrupt longer ones, each copy still sees a valid ordered subsequence of episodes, and the total learning cost of all active copies remains of order $C(n)$.

Inside a MASTER block of length N, each aligned dyadic cell of length $q \le N$ starts a fresh MVP copy with probability

$$p_{N,q} := \frac{\eta(N)}{\eta(q)}.$$

The length-N copy always starts. In each episode, the algorithm runs the shortest available copy that covers that episode. Any longer copy pauses while the shorter copy runs and resumes afterward. Each launch coin is drawn when its cell begins and is independent of future kernel changes.

For a maintained copy A, let $\mathrm{supp}(\mathsf{A}) = [\mathsf{A}.s, \mathsf{A}.e]$, and let $n_{\mathsf{B}}(J)$ be the number of rounds in J on which copy B is active.

The first bound below is Freedman's inequality for bounded episode returns (Freedman, 1975). The second is the launch-count bound proved inside the proof of Lemma 3 of Wei and Luo (2021), written for the complexity function C.

Lemma K.4 (Uniform return and scheduling event)**.** *With probability at least $1-\delta$, the following bounds hold simultaneously. The first holds for every chronological interval J. The second holds for every length-N parent block, every maintained copy A, and every length-n prefix $J = [\mathsf{A}.s, t] \subseteq \mathrm{supp}(\mathsf{A})$ that starts with the copy:*

$$\left|\sum_{u\in J}\{Y_u - f_u(\pi_u)\}\right| \le \widetilde{O}(C(|J|)), \qquad \sum_{\mathsf{B}:n_{\mathsf{B}}(J)>0} C(n_{\mathsf{B}}(J)) \le \widetilde{O}(C(n)).$$

Proof. Apply Freedman's inequality to the episode-return martingale and take a union bound over the at most K^2 intervals. This proves the first statement. Its $\widetilde{O}(\sqrt{|J|})$ term is no larger than $C(|J|)$.

For the second statement, fix N, A, J and a dyadic length $q \le N$. At most $\lceil n/q \rceil + 2$ aligned cells of length q can update inside J. Each starts a copy with probability $p_{N,q}$. Bernstein's inequality and a union bound give

$$Z_{J,q} \le \widetilde{O}\left(\left(\frac{n}{q} + 2\right)\frac{\eta(N)}{\eta(q)} + 1\right).$$

Each such copy contributes at most $C(\min\{q, n\})$. When $q \le n$, the mean contribution is at most $2C(n)$, and the Bernstein remainder is at most $C(n)$ times a logarithmic factor. When $q > n$, only a constant number of cells meet the prefix, and each contributes at most $C(n)$. Summing over dyadic q and adding the root copy proves the claim. □

Let $g_t \in [0,1]$ denote the pre-episode optimistic scalar output by the active copy, recorded before observing or using episode t's trajectory.

Lemma K.5 (MVP bounds for paused copies)**.** *On the joint MVP and scheduling event, every maintained copy* A *and every length-n prefix $J = [\mathsf{A}.s, t]$ satisfy*

$$g_r \geq \min_{u \in J} f^\star_u - c\Gamma(J), \quad (r \in J),$$
$$\sum_{r \in J} (g_r - Y_r) \leq \widetilde{O}(C(n) + n\Gamma(J)),$$
$$\sum_{r \in J} (f^\star_r - Y_r) \leq \widetilde{O}(C(n) + n\Gamma(J)).$$

Proof. At round $r \in J$, the active copy started no earlier than $\mathsf{A}.s$. All episodes used by that copy therefore lie in J. Approximate optimism from Proposition G.2, together with monotonicity of the interval diameter, proves the first claim.

Split J into the rounds assigned to each active copy. For any one copy, these rounds form a consecutive prefix of its updates, so Proposition G.2 applies. The scheduling bound in Lemma K.4 makes all learning costs sum to $\widetilde{O}(C(n))$. Since exactly one copy is active in each round, the change costs sum to $n\Gamma(J)$. This proves the second bound. Finally, the optimal values within J differ by at most $c\Gamma(J)$, which gives the third bound. □

K.4 Fresh intervals and the MASTER tests

This subsection identifies intervals on which the kernels change too little to justify a restart. We show that neither MASTER test fires inside such an interval, and then partition every comparison block into fresh pieces with controlled total learning cost.

Set $c_g = 1$. Any other universal choice would only change the constants in the tests in equation (2).

We call an interval I *uniformly fresh* if $\Gamma(I) \leq c_g \eta(|I|)$. In words, the kernels change no more than the base algorithm's average error on that interval. Every subinterval $J \subseteq I$ is fresh as well, because $\Gamma(J) \leq \Gamma(I)$ and $\eta(|J|) \geq \eta(|I|)$. Therefore, Lemma K.5 gives average error $\widetilde{O}(\eta(|J|))$ for every copy prefix and replay interval contained in I.

Fresh intervals do not cause restarts. Suppose a MASTER block starts inside a uniformly fresh interval I. Take any prefix J that remains inside I. Since $\Gamma(J) \leq \Gamma(I) \lesssim \eta(|I|) \leq \eta(|J|)$, the second bound in Lemma K.5 keeps the average prediction error below the threshold of Test 2.

Now consider Test 1 and a completed copy of length q inside I. Its average return is at most its average optimal value plus $\widetilde{O}(\eta(q))$. The optimal values inside I differ by at most $c\Gamma(I)$. The first bound in Lemma K.5 also gives $U_t \geq \min_{r \in [s,t]} f^\star_r - c\Gamma(I)$. Because $q \leq |I|$, freshness makes both diameter errors no larger than the test tolerance. Thus Test 1 cannot fire either. A block that starts in I cannot restart before it leaves I.

Lemma K.6 (Partition into fresh intervals)**.** *Every interval B of length n admits a partition $\mathcal{P}(B)$ into uniformly fresh intervals such that*

$$\sum_{I \in \mathcal{P}(B)} C(|I|) \leq \widetilde{O}(C(n) + n\Gamma(B)).$$

Proof. Write $\gamma = \Gamma(B)$. If $\gamma \leq c_g \eta(n)$, then B is already fresh, so no split is needed. Otherwise, choose the largest dyadic length $w \leq n$ for which $\gamma \leq c_g \eta(w)$, and split B into consecutive pieces of length at most w. Each piece is fresh because its diameter is at most γ, while its learning error is at least $\eta(w)$.

It remains to bound the total cost of these pieces. If $2w \leq n$, the maximal choice of w, together with regularity between adjacent dyadic scales, gives $\eta(w) \lesssim \gamma$. There are at most $n/w + 1$ pieces, so their total cost is at most $(n/w + 1)C(w) \lesssim n\gamma + C(n)$. If $2w > n$, the partition has at most two pieces. Monotonicity of C then gives the same bound. □

K.5 CHARGING REPLAY TO KERNEL CHANGE

This subsection explains why adaptation does not create an uncontrolled replay cost. A large prediction error first provides evidence of a large kernel change. The replay schedule then detects the error and restarts the block, which limits how long the error can continue.

Consider a planned length-N MASTER block starting at s. Let $B = [s, e]$ be the episodes actually run before a restart or the end of learning, and let $\overline{B} = \{s, \dots, s + N - 1\}$ be the full planned block.

Lemma K.7 (Large prediction error requires kernel change)**.** *If $t \in B$ and $f_t^\star - g_t \geq c_w \widehat{\eta}(q)$, then*

$$\Gamma(\overline{B}) \geq c_{\mathrm{loc}} \eta(q)$$

for a universal $c_{\mathrm{loc}} > 0$.

Proof. Let $J_t \subseteq \overline{B}$ contain all episodes used by the active copy up to time t. The certificate in Proposition G.2 says that g_t is no smaller than $\min_{u \in J_t} f_u^\star - c\Gamma(J_t)$. The simulation bound says that the current optimal value differs from this minimum by at most $c\Gamma(\overline{B})$. Since $\Gamma(J_t) \leq \Gamma(\overline{B})$, these two facts give $f_t^\star - g_t \leq c'\Gamma(\overline{B})$. Therefore, an error of size $c_w \widehat{\eta}(q)$ requires $\Gamma(\overline{B}) \geq c_{\mathrm{loc}} \eta(q)$. □

Dyadic error decomposition. We next sort prediction errors by dyadic scale. This is the deterministic decomposition in Wei and Luo (2021, Lemma 16). Let $\{I_i\}_i$ be any time-ordered partition of the episodes run in a planned length-N block. For each dyadic $q \leq N$, let $\tau_i(q)$ be the first time in I_i at which $f_t^\star - g_t \geq c_w \widehat{\eta}(q)$. Let $\xi_i(q) = [\max I_i - \tau_i(q) + 1]_+$ count the rounds from that first crossing to the end of I_i. If the threshold is never crossed, set $\xi_i(q) = 0$. Then

$$\sum_{t \in B} (f_t^\star - g_t) \leq \widetilde{O}\left(C(|B|) + \sum_{q \in \{1,2,4,\dots,N\}} \eta(q) \sum_i \xi_i(q) \right).$$

To prove the bound, write $d_t = [f_t^\star - g_t]_+ \leq 1$. Small errors, meaning $d_t < c_w \widehat{\eta}(N)$, contribute at most $\widetilde{O}(|B|\eta(N)) = \widetilde{O}(C(|B|))$. Here and below, the fixed logarithmic factor in $\widehat{\eta}$ is absorbed into $\widetilde{O}(\cdot)$.

Now take a round $t \in I_i$ with a larger error. Let $q_t \leq N$ be the smallest dyadic scale whose threshold is crossed. Since $\tau_i(q_t) \leq t$, this round is counted by $\xi_i(q_t)$. If $q_t > 1$, the threshold at scale $q_t/2$ was not crossed. Regularity between adjacent scales therefore gives

$$d_t < c_w \widehat{\eta}(q_t/2) \leq \widetilde{O}(\eta(q_t));$$

For $q_t = 1$, we instead use $d_t \leq 1 = \eta(1)$. Thus each large error can be charged to one scale at which the corresponding threshold has already been crossed. Summing these charges, and then allowing all later rounds to be counted as well, gives the displayed bound. Finally, $f_t^\star - g_t \leq d_t$.

The next lemma controls how long such a threshold crossing can remain unpaid. A replay eventually detects it and restarts the block. By Lemma K.7, this replay charge can occur only when the kernel diameter is large enough. The argument adapts Wei and Luo (2021, Lemma 17).

Lemma K.8 (Replay coverage after a large change)**.** *Condition on the kernel path and fix a planned length-N block. If the algorithm reaches it, let B be the episodes actually run. Partition $\overline{B}$ into uniformly fresh intervals $\{G_i\}_i$, using only the kernel path. Let $I_i = G_i \cap B = [s_i, e_i]$ denote the nonempty intersections, and define $\tau_i(q)$ and $\xi_i(q)$ as above. On the scheduling event,*

$$\eta(q) \sum_i [\xi_i(q) - 4q]_+ \leq \widetilde{O}\left(\frac{q\eta(q)^2}{\eta(N)} \mathbf{1}\{\Gamma(\overline{B}) \geq c_{\mathrm{loc}} \eta(q)\} \right).$$

Proof. We prove the claim in four steps. First, a long suffix creates many chances to launch a replay. Second, only a logarithmic number of these launches can fail. Third, the first successful replay restarts the block. Finally, a large prediction error ties this restart to a large kernel diameter.

Step 1: Count eligible replay starts.

Fix an interval I_i with a nonzero long suffix. Consider the aligned length-q cells that start after $\tau_i(q)$ and remain inside the fresh interval G_i. We may lose at most $2q$ rounds at each boundary. Let A_i be the number of remaining cell starts reached before the block ends. The dyadic grid then gives

$$A_i \geq \frac{[\xi_i(q) - 4q]_+}{q}$$

unless an earlier eligible replay was already launched. After a successful launch, either the block ends during the next q rounds or the replay reaches its endpoint and triggers the restart proved in Step 3. The $4q$ boundary allowance covers both cases.

Step 2: Control predictable launch failures.

Order all eligible starts by time. Whether a start is eligible is known before its launch coin is drawn: the threshold crossing is in the past, the cell is fixed, and the block has survived all earlier tests. Conditional on this past and on the kernel path, the launch probability is $p_{N,q} = \eta(N)/\eta(q)$. The probability that L eligible launches all fail is therefore at most $e^{-p_{N,q}L}$. Set

$$L_{N,q} = \left\lceil p_{N,q}^{-1} \log \frac{cK^2(1 + \log_2 K)}{\delta} \right\rceil$$

There are at most K^2 choices of a start and a block length, and at most $1 + \log_2 K$ scales. A union bound therefore shows that every reached block has at most $\widetilde{O}(p_{N,q}^{-1})$ eligible failures before the first successful launch. Reaching a block depends only on earlier coins, so it does not change the probabilities of the later launch coins.

Step 3: Use the first successful replay to restart.

Consider the first replay that is launched. If the block ends before this replay finishes, there are no later rounds to charge. Otherwise, its whole length-q support lies inside one fresh interval G_i, and it starts after $\tau_i(q)$. Since the running baseline is a minimum, $U_{\mathsf{A}.e} \leq g_{\tau_i(q)}$. Also, $|G_i| \geq q$, so Corollary D.1 gives, for every replay round t,

$$f_t^\star \geq f_{\tau_i(q)}^\star - c\Gamma(G_i) \geq f_{\tau_i(q)}^\star - \widetilde{O}(\eta(q)).$$

The return bound in Lemma K.5, applied to this completed replay, gives

$$\frac{1}{q} \sum_{t=\mathsf{A}.s}^{\mathsf{A}.e} Y_t \geq \frac{1}{q} \sum_{t=\mathsf{A}.s}^{\mathsf{A}.e} f_t^\star - \widetilde{O}(\eta(q)) \geq f_{\tau_i(q)}^\star - \widetilde{O}(\eta(q)).$$

At time $\tau_i(q)$, the threshold crossing says that $f_{\tau_i(q)}^\star - g_{\tau_i(q)} \geq c_w \widehat{\eta}(q)$. Choose c_w larger than c_1 and the constants in the two error bounds above. The replay's average return then exceeds $U_{\mathsf{A}.e} + c_1\widehat{\eta}(q)$. Test 1 fires at the replay endpoint, so the block ends no later than this first successful replay. Together with Step 2, this gives $\sum_i [\xi_i(q) - 4q]_+ \leq \widetilde{O}(q/p_{N,q})$.

Step 4: Restore the displacement condition.

If the left-hand side is nonzero, at least one threshold crossing $\tau_i(q)$ exists. By Lemma K.7, such a crossing requires $\Gamma(\overline{B}) \geq c_{\mathrm{loc}}\eta(q)$. The bound from Step 3 can therefore be multiplied by the indicator in the lemma statement. □

Lemma K.9 (Regret on one MASTER block)**.** *Let B be the episodes run in a planned length-N block. For every partition $\mathcal{P}_B$ obtained by intersecting B with a kernel-path-measurable partition into uniformly fresh intervals,*

$$\sum_{t \in B} (f_t^\star - Y_t) \leq \widetilde{O}\left[\sum_{I \in \mathcal{P}_B} C(|I|) + \sum_{q \in \{1,2,4,\ldots,N\}} \frac{q\eta(q)^2}{\eta(N)} \mathbf{1}\{\Gamma(\overline{B}) \geq c_{\mathrm{loc}}\eta(q)\} \right].$$

Proof. Write each regret term as $f_t^\star - Y_t = (f_t^\star - g_t) + (g_t - Y_t)$, and let $n = |B|$. We first bound the second part. If Test 2 does not fire on the last round, the definition of the test gives

$\sum_{t\in B}(g_t - Y_t) < c_2 n\widehat{\eta}(n)$. If Test 2 fires on that round, the preceding $n-1$ rounds passed the test, and the final increment is at most one. Hence

$$\sum_{t\in B}(g_t - Y_t) \le c_2(n-1)\widehat{\eta}(n-1) + 1 \le \widetilde{O}(C(n)).$$

The same bound holds if Test 1 fires on the final round. Thus $\sum_{t\in B}(g_t - Y_t) = \widetilde{O}(C(|B|))$. Because η is nonincreasing and $\mathcal{P}_B$ partitions B, this cost is at most $\sum_{I\in\mathcal{P}_B} C(|I|)$.

It remains to bound $\sum_{t\in B}(f_t^\star - g_t)$. Apply the dyadic decomposition above. At each scale, the first $4q$ rounds after a threshold crossing contribute at most $4q\eta(q)$; summed across the fresh pieces and scales, these terms are $\widetilde{O}(\sum_{I\in\mathcal{P}_B} C(|I|))$. The longer suffixes are bounded by Lemma K.8. Adding the two parts proves the lemma. □

K.6 Assembly of the comparison bound

We now sum the one-block bounds over the full run. The first lemma controls ordinary learning and replay costs for any comparison window, and the second averages the replay indicators under diffusive distribution shift.

K.6.1 Proof of Lemma K.1

We first split the comparison blocks into fresh intervals. We then show that MASTER creates only a logarithmic number of pieces inside each such interval and sum the one-block bound over those pieces.

Step 1: Build and count fresh fragments.

Apply Lemma K.6 to every coarse comparison block. The resulting pieces form a fresh partition $\mathcal{P}$ of $[K]$. For bookkeeping near the horizon, extend the kernel path to time $2K$ by setting $P_t = P_K$, and append the zero-diameter interval $[K+1, 2K]$. Denote the extended partition by $\mathcal{P}^+$.

Fix one fresh interval $I \in \mathcal{P}^+$. At most one MASTER block can enter I from the left. If a new epoch starts a block inside I, the fresh-interval argument prevents another restart before that block leaves I. Thus only a constant number of epochs can meet I. Each epoch contains at most $1 + \lfloor \log_2 K \rfloor$ planned blocks, so I is cut into only $O(\log K)$ intersections. Every intersection $J \subseteq I$ is still fresh because $\Gamma(J) \le \Gamma(I) \le \eta(|I|) \le \eta(|J|)$.

Step 2: Sum the one-block regret bounds.

The partition $\mathcal{P}^+$ depends only on the kernel path. We may therefore condition on that path and apply Lemma K.9 to every realized MASTER block. If $J = I \cap B$ is nonempty, then $C(|J|) \le C(|I|)$. Since Step 1 creates only $O(\log K)$ such intersections for each I,

$$\sum_{B\in\mathfrak{B}} \sum_{I\in\mathcal{P}: I\cap B\neq\varnothing} C(|I\cap B|) \le \widetilde{O}\left(\sum_{I\in\mathcal{P}} C(|I|)\right).$$

The replay terms remain indexed by their MASTER blocks, and their sum is $\mathcal{R}_{\mathrm{replay}}$. The ordinary learning terms are controlled by the last display. Finally, summing Lemma K.6 over the coarse comparison blocks gives

$$\sum_{I\in\mathcal{P}} C(|I|) \le \widetilde{O}\left[\left(\frac{K}{W}+1\right) C(W) + \sum_{B_0\in\mathcal{B}_W} |B_0|\Gamma(B_0)\right],$$

Combining these two bounds proves Lemma K.1. □

K.6.2 Proof of Lemma K.2

We count how many blocks of each length can be reached and apply the stopping-time diameter bound at their random starting times. Summing the resulting tail probabilities gives the expected replay cost.

Fix a dyadic block length N. Before an epoch can reach a length-N block, it must complete earlier blocks whose total length is $N-1$. Therefore, at most cK/N such blocks can be reached. When $N=1$, the trivial bound K gives the same order.

Let $L_N = \lceil cK/N \rceil$. Write $E_{N,j}$ for the event that the j-th length-N block is reached, and let $\tau_{N,j}$ be its starting time. If this block is never reached, set $\tau_{N,j} = K+1$. Because the kernel path is held constant after time K, $\tau_{N,j}$ is a bounded stopping time and $E_{N,j}$ is known at that time. We can therefore apply Lemma F.1 conditionally at $\tau_{N,j}$:

$$\mathbb{E}\big[\mathbf{1}_{E_{N,j}}\mathbf{1}\{\Gamma(I_{\tau_{N,j},N}) \geq c_{\mathrm{loc}}\eta(q)\} \mid \mathcal{F}_{\tau_{N,j}}\big] \leq \widetilde{O}\left(\mathbf{1}_{E_{N,j}} \min\left\{1, \frac{H^2\nu N}{\eta(q)^2}\right\}\right).$$

Multiply this probability by the replay charge at scale q, then sum over j, q, and N. This is exactly the replay sum in the lemma statement. Notice that no independence is needed between the number of reached blocks and their diameters: the diameter bound is applied at each random starting time. This proves Lemma K.2. □

K.7 Evaluation of the multiscale replay sum

The final calculation uses one elementary dyadic bound. For every dyadic N and every $z \in [0,1]$,

$$\sum_{q \in \mathcal{D}_N} q \min\{\eta(q)^2, z\} \leq \widetilde{O}(a^2 + b\sqrt{z}).$$

The case $z=0$ is immediate, so assume $z>0$. Write $x_q = a/\sqrt{q}$ and $y_q = b/q$. Up to a universal constant, $\eta(q)^2 \leq \min\{1, x_q^2\} + \min\{1, y_q^2\}$. We may bound the two parts separately because $\min\{u+v, z\} \leq \min\{2u, z\} + \min\{2v, z\}$. It remains to control

$$\sum_{q \in \mathcal{D}_N} \min\{qz, a^2\} + \sum_{q \in \mathcal{D}_N} \min\left\{qz, \frac{b^2}{q}\right\}.$$

Every term in the first sum is at most a^2, so that sum is $O(a^2 \log(2N))$. For the second sum, split the dyadic scales at $q_0 = b/\sqrt{z}$. The scales below q_0 contribute $O(q_0 z)$, and the scales above it contribute $O(b^2/q_0)$. Both terms are $O(b\sqrt{z})$, which proves the estimate.

K.7.1 Proof of Lemma K.3

The definition of W_v creates three cases: a stationary path, a lifetime clipped at K, and an interior lifetime. We bound the replay sum in each case.

Step 1: Remove the trivial regime.

If $v=0$, every replay indicator is zero and there is nothing to prove. In the remaining nontrivial regime, the assumption gives $W_v \gg a^2$. Hence the cap in $C(W_v)$ is inactive.

Step 2: Bound the boundary case $W_v = K$.

Suppose first that $W_v = K$. Then $v \leq a^2/K^2$. For every $q \leq N \leq K$, monotonicity gives $\eta(q)^2 \geq vN$. Thus the contribution of a fixed scale N is at most $vN/\eta(N)$. The scales with $N < a^2$ contribute $O(va^2)$, which is at most $O(\sqrt{vK})$ because $\sqrt{v} \leq a/K$ and $K \gg a^2$. The scales with $N \geq a^2$ contribute at most $vK^{3/2}/a \leq \sqrt{vK}$.

Step 3: Bound the interior case $1 < W_v < K$.

Now suppose $1 < W_v < K$. The definition of W_v gives $v \asymp a^2/W_v^2$, and hence $\sqrt{vW_v} \asymp a/\sqrt{W_v}$. We first sum over $N \leq W_v$. The scales below a^2 contribute at most $va^2 \leq a/\sqrt{W_v}$, while the remaining scales contribute at most $vW_v^{3/2}/a = a/\sqrt{W_v}$.

For $N > W_v$, apply the dyadic estimate above with $z = \min\{1, vN\}$. Since $C(N) \geq a\sqrt{N}$, the contribution at a fixed scale N is at most

$$\widetilde{O}\left(\frac{a}{\sqrt{N}} + \frac{b\sqrt{v}}{a}\right).$$

The first term forms a geometric sum starting at W_v, so its total is $\widetilde{O}(a/\sqrt{W_v})$. The second term occurs at only logarithmically many scales and equals b/W_v up to logarithmic factors. Finally, the nontrivial condition and $a \geq 1$ rule out $W_v = 1$. This proves Lemma K.3. □

K.8 Unrestricted-kernel specialization of Theorem 5.1

For unrestricted kernels, set $a = \widetilde{\Theta}(\sqrt{|\mathcal{S}|\,|\mathcal{A}|})$ and $b = \widetilde{\Theta}(|\mathcal{S}|^2|\mathcal{A}|)$ in (3). In the interior regime, $W_\nu \asymp \sqrt{|\mathcal{S}|\,|\mathcal{A}|}/(H\sqrt{\nu})$. At this window length, the statistical and mismatch terms both have order

$$\widetilde{O}\Big(KH^{3/2}(|\mathcal{S}|\,|\mathcal{A}|)^{1/4}\nu^{1/4}\Big).$$

If $W_\nu = K$, the stationary term $H\min\{K, \sqrt{|\mathcal{S}|\,|\mathcal{A}|K}\}$ is already large enough to cover both terms. If $W_\nu = 1$, regret is at most HK. The remaining b-term contributes $\widetilde{O}(H|\mathcal{S}|^2|\mathcal{A}|K/W_\nu)$. This proves the unrestricted-kernel bound in Theorem 5.1.

K.9 Bounded-next-state-support specialization

For kernels with at most d_0 possible next states per row, set $a = \widetilde{\Theta}(\sqrt{|\mathcal{S}|\,|\mathcal{A}|})$ and $b = \widetilde{\Theta}(d_0|\mathcal{S}|\,|\mathcal{A}|) = \widetilde{O}_{d_0}(a^2)$. For every $n \geq 1$,

$$(a\sqrt{n} + b) \wedge n \leq \widetilde{O}_{d_0}\big(n \wedge a\sqrt{n}\big).$$

When $n \leq a^2$, the cap by n gives the right-hand side. When $n \geq a^2$, the additive term satisfies $b \leq \widetilde{O}_{d_0}(a\sqrt{n})$. Thus, up to logarithmic factors and constants depending on d_0, the base complexity has the square-root form used in the lifetime calculation.

Apply (3) with $W_\nu = \min\{K, \max\{1, a/(H\sqrt{\nu})\}\}$. In the interior regime, the statistical and diameter terms are both

$$\widetilde{O}_{d_0}\Big(KH^{3/2}a^{1/2}\nu^{1/4}\Big) = \widetilde{O}_{d_0}\Big(KH^{3/2}(|\mathcal{S}|\,|\mathcal{A}|)^{1/4}\nu^{1/4}\Big).$$

If $W_\nu = K$, the stationary term covers both contributions. If $W_\nu = 1$, use the regret cap HK. This proves the upper bound. The construction in Theorem 4.2 has next-state support size two, and its lower bound still holds when ν is unknown. Hence the rates match for every fixed $d_0 \geq 2$. This completes the proof of Theorem 5.1. □

Revealing ν can only help the agent, so the same lower bound applies when ν is unknown. Comparing the bound above with Theorem 4.1 shows that learning ν adds no polynomial factor.

L Notes on Smooth Latent Diffusions

Smooth latent processes provide natural examples of diffusive distribution shift. Consider a latent random walk with natural filtration $\mathcal{G}_k$. If $\theta_{k+1} = \theta_k + \sigma\xi_{k+1}$, $\mathbb{E}[\xi_{k+1} \mid \mathcal{G}_k] = 0$, $\mathbb{E}[\|\xi_{k+1}\|_2^2 \mid \mathcal{G}_k] \leq d$, and P_θ is L-Lipschitz in TV, martingale orthogonality gives

$$\mathbb{E}\left[\left\|P_{\theta_{k+\ell}} - P_{\theta_k}\right\|_{\infty,\mathrm{TV}}^2 \mid \mathcal{G}_k\right] \lesssim L^2 d\sigma^2 \ell.$$

Since the kernel filtration is contained in $\mathcal{G}_k$, the tower property verifies Assumption 2.1 with $\nu \asymp L^2 d\sigma^2$, after truncating the right side at one.

For a mean-reverting AR(1) latent process $\theta_{k+1} = \rho\theta_k + \sigma\xi_{k+1}$, with $\rho \in [0, 1)$ and the same martingale-difference noise conditions, the conditional change also contains deterministic drift:

$$\mathbb{E}[\|\theta_{k+\ell} - \theta_k\|_2^2 \mid \mathcal{G}_k] \leq 2(1-\rho^\ell)^2\,\|\theta_k\|_2^2 + 2\sigma^2 d\sum_{j=0}^{\ell-1}\rho^{2j}.$$

If $\|\theta_k\|_2 \leq R$ almost surely, then $(1-\rho^\ell)^2 \leq \ell(1-\rho)/(1+\rho)$. The induced kernels satisfy the condition with $\nu \lesssim L^2\{R^2(1-\rho)/(1+\rho) + d\sigma^2\}$. Our lower bound uses geometric refreshes and applies to the full kernel-level class. Smooth subclasses may admit faster rates under extra margin, observability, or mixing assumptions.

M Empirical Validation Details

The experiments test the freshness predictions and the algorithms used in the upper bounds. They are not a broad benchmark.

M.1 Controlled Theorem-Algorithm Validation

This experiment has two goals. It checks the lifetime and fourth-root predictions in controlled environments, and it evaluates the exact known- and unknown-rate algorithms analyzed in the paper. We first describe the code correspondence and then explain each diagnostic.

Algorithm-to-code correspondence. The MVP code uses the dyadic row snapshots and empirical-Bernstein backup in equation (1). It normalizes values and pools transitions across stages. Fresh-Restart starts a new copy in each block. Ada-Fresh launches copies with probability $\eta(N)/\eta(q)$, runs the shortest covering copy, and applies both tests in Algorithm 2. Safe-prefix stopping is only a proof device; the implementation retains every transition. Constants are chosen on development seeds and fixed before evaluation. The empirical-diameter rule appears only as an ablation.

Diffusive survival benchmark. We use one live state, one absorbing state, and two actions. In episode k, the good action survives with probability 0.99, while the other survives with probability $0.99 - \Delta$. With probability q, the good action is redrawn uniformly before the next episode. Thus a kernel change moves every affected row by at most Δ in total variation, and over every ℓ-episode interval,

$$\mathbb{E}\Big[\|P_{k+\ell} - P_k\|_{\infty,\mathrm{TV}}^2 \mid \mathcal{E}_k\Big] \le q\Delta^2\ell.$$

We set $\nu = q\Delta^2$ and $q = H^2\Delta^2/(|\mathcal{S}|\,|\mathcal{A}|)$. For $|\mathcal{S}| = |\mathcal{A}| = 2$, this makes the mean refresh lifetime equal to the freshness comparator:

$$\frac{1}{q} = \frac{\sqrt{|\mathcal{S}|\,|\mathcal{A}|}}{H\sqrt{\nu}} = W_\nu.$$

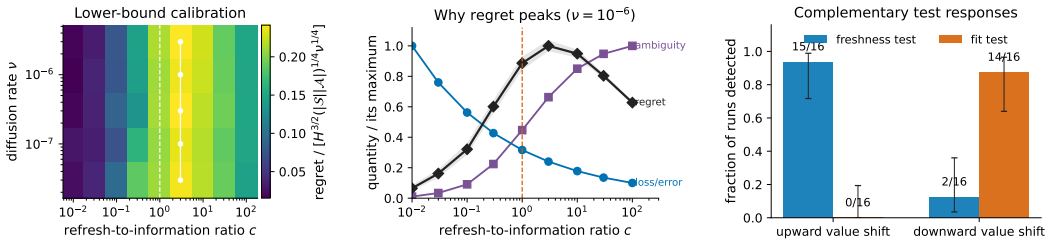
The benchmark therefore follows the same least-favorable calibration as the survival lower bound, up to universal constants.

For the main rate experiment, we use $H = 60$, $K = 8000$, good-action survival probability 0.995, and seven values $\nu \in \{10^{-8}, 3{\cdot}10^{-8}, \dots, 10^{-5}\}$. This gives at least 24 predicted information lifetimes even at the smallest diffusion rate. Results use twelve held-out seeds; separate development seeds set all constants. The complexity is $C(n) = \min\{n, \sqrt{n} + 2\}$, matching the theorem's $(a\sqrt{n} + b) \wedge n$ form. The MVP bonus uses the exact logarithmic factor and multiplier 0.1, and regret includes all episodes. The additional fixed-window and component diagnostics use $H = 20$, $K = 6000$, and survival probability 0.99.

Smooth-path freshness diagnostic. To isolate freshness from policy margins, let $\theta_{k+1} = \theta_k + \sigma\epsilon_{k+1}$, with $p_{k,0} = 1/2 + b\sin\theta_k$ and $p_{k,1} = 1/2 + b\cos\theta_k$. This map is b-Lipschitz in maximum TV, so $\sigma = \sqrt{\nu}/b$ gives diffusion rate ν. Each episode provides one balanced probe from each row. We evaluate all windows after a common initialization using 40 seeds and 20,000 episodes.

Revealed-refresh hardness diagnostic. The hardness scan isolates the binary test in the lower bound. We set $H = 20$, $|\mathcal{S}| = |\mathcal{A}| = 2$, and $m = H^2/(|\mathcal{S}|\,|\mathcal{A}|) = 100$. A hidden best action $J_k \in \{0, 1\}$ is redrawn with probability q, and the agent observes each refresh. Choosing action a yields m Bernoulli probes with mean $1/2 + \Delta/2$ if $a = J_k$, and $1/2 - \Delta/2$ otherwise. The agent resets its log likelihood after each refresh and chooses the posterior maximum. This rule is Bayes optimal because either action gives the same testing information.

For diffusion rate ν and calibration c, set $\Delta = (|\mathcal{S}|\,|\mathcal{A}|\nu/(cH^2))^{1/4}$ and $q = cH^2\Delta^2/(|\mathcal{S}|\,|\mathcal{A}|)$. A wrong action costs $\min\{H, H^2\Delta\}$. Thus $\nu = q\Delta^2$, and c compares refresh and information rates. We test nine values of c from 10^{-2} to 100, five diffusion rates from $3{\cdot}10^{-8}$ to $3{\cdot}10^{-6}$, 20,000 episodes, and sixteen held-out seeds. Each bootstrap replicate repeats the maximization over c.



(a) Hardness across calibrations; circles mark maxima.

(b) Testing ambiguity and loss per error.

(c) Responses of the two adaptive tests.

Figure 3: Additional mechanism diagnostics. The revealed-refresh experiment has an interior hardest calibration because statistical ambiguity increases as the value lost per error decreases. The final panel verifies that the two ADA-FRESH tests respond to complementary directions of change.

Directional freshness-test stress test. To exercise both MASTER tests, we use upward and downward survival-kernel shifts. The low-value survival probabilities are $(0.30, 0.10)$, and the high-value probabilities are $(0.70, 0.99)$. After a 1,100-episode stable prefix, a one-time shift occurs with hazard $q = \nu/D^2$, where $\nu = 2{\cdot}10^{-3}$ and $D = 0.89$. Conditional on any history, crossing the shift within ℓ episodes has probability at most $q\ell$. Hence the expected squared kernel change is at most $qD^2\ell = \nu\ell$.

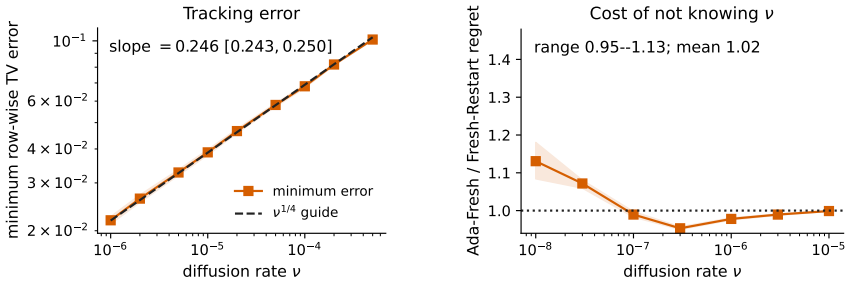
We run all 6,000 episodes with ADA-FRESH and the MVP base algorithm. Development seeds set the two MASTER constants to 0.1 and 24; results use sixteen held-out seeds. Under the upward shift, the short-window test fires in $15/16$ runs and the prediction-error test in none. Under the downward shift, the counts are $2/16$ and $14/16$. The experiment tests the distinct roles of the two tests.

Lower-bound tradeoff and uncertainty. In figure 3, regret is maximized at $c = 3$ for every tested diffusion rate. The log–log slope of the resulting maximal-regret curve is 0.2531, with a 95% bootstrap confidence interval across seeds of $[0.2412, 0.2632]$. Panel (a) reports held-out-seed means, panel (b) shades the 95% bootstrap interval for expected regret, and panel (c) uses Wilson intervals for the event that a test fires at least once. This experiment supports the testing-error/value-loss tradeoff behind the lower bound; the lower bound itself remains the analytic statement of Theorem 4.2.

What the experiment supports. The lifetime experiment in figure 1a gives a fitted best-window exponent of -0.491, with a 95% bootstrap confidence interval across seeds of $[-0.498, -0.474]$. Panel (a) of figure 4 gives a fitted minimum estimation-error exponent of 0.246, with interval $[0.243, 0.250]$. These measurements directly test the separate predictions $W_\nu \asymp \nu^{-1/2}$ and error $\asymp \nu^{1/4}$; neither axis is normalized by an empirically selected window.

For the online-RL rate experiment, a log–log fit of full-run regret against ν gives exponent 0.2551 for FRESH-RESTART with a 95% paired-seed bootstrap interval of $[0.2514, 0.2592]$, and 0.2383 for ADA-FRESH with interval $[0.2380, 0.2387]$. ADA-FRESH's regret divided by the tuned algorithm's regret ranges from 0.953 to 1.131, with mean 1.016. These full-run results support the predicted fourth-root dependence and show similar empirical scaling for known and unknown diffusion rates; they do not discard startup episodes or fit a selected subset of diffusion rates.

Diagnostics and scope. The cumulative prediction-error test triggers on average between 5.25 and 12.42 times as ν increases across the tested range. The short-window comparison test does not trigger on this particular family. This is structurally expected: a redraw changes which action is optimal but leaves the optimal value unchanged, so the one-sided upward-value comparison in Test 1 has no signal; Test 2 detects the stale policy through optimistic prediction error. The largest aggregate empirical mean-squared kernel change divided by the analytic displacement bound is 0.737 for the survival family and 0.840 for the smooth diagnostic. Fixed-restart sweeps, the kernel-change audit, and the empirical-diameter heuristic remain in the reproducibility package as diagnostics, but are not used as primary empirical evidence for either upper-bound theorem.



(a) Minimum tracking error.

(b) Cost of adapting to unknown ν.

Figure 4: Additional controlled diagnostics. The smooth-path estimation error follows the predicted $\nu^{1/4}$ law, while the regret ratio of ADA-FRESH to FRESH-RESTART remains order one. Bands are 95% intervals over held-out seeds; the dashed curve in panel (a) is an anchored theoretical slope.

Finally, the least-favorable survival calibration couples the policy gap to ν. A fourth-root slope on that family alone would not establish a minimax lower bound: even a nearly constant action-error rate can inherit that slope from the instance gap. The smooth-path diagnostics therefore report the lifetime and estimation-error laws separately. The lower bound in Theorem 4.2, including the unknown-ν case, rests on the testing argument above.

M.2 OPSD-ANCHORED WIND-STORAGE CONTROL

This experiment tests finite memory in a realistic battery-control problem. It combines observed energy regimes with a fixed engineering model and a controlled diffusive path.

Energy data and storage MDP. We use the first 900 complete days of the Denmark DK1 series in Open Power System Data (Open Power System Data, 2020). Each episode has $H = 24$ hours. Wind, load, and day-ahead price are each divided into two empirical-quantile bins. Combining these eight exogenous states with nine battery levels gives $|\mathcal{S}| = 72$. Five actions range from full discharge to full charge. The model includes battery and power limits, efficiency, self-discharge, grid trading, reliability and curtailment penalties, degradation, and a terminal reserve target. The agent knows these physics and costs and estimates only the exogenous kernel.

Empirical regimes and displacement bound. For each day, we form a scarcity score from normalized load plus price minus wind. The lowest and highest 15% of scores define empirical surplus and scarcity endpoint kernels P^{sur} and P^{sca}. Let D be their largest TV distance across state–action pairs. We then evolve the kernel according to

$$\theta_{k+1} = \theta_k + \sigma\epsilon_{k+1}, \qquad \alpha_k = 0.5 + 0.45\sin(\theta_k), \qquad P_k = (1-\alpha_k)P^{\mathrm{sur}} + \alpha_k P^{\mathrm{sca}}.$$

Since sine is 1-Lipschitz, the process satisfies Assumption 2.1 with $\nu = (0.45D\sigma)^2$. Here $D = 0.8665$, $\sigma = 0.10$, and $\nu = 1.5203{\cdot}10^{-3}$. Across all audited episode separations and test paths, the largest empirical-to-analytic mean-squared-change ratio is 0.462. Only the path between regimes is simulated; the endpoints come from data, and the physical model stays fixed.

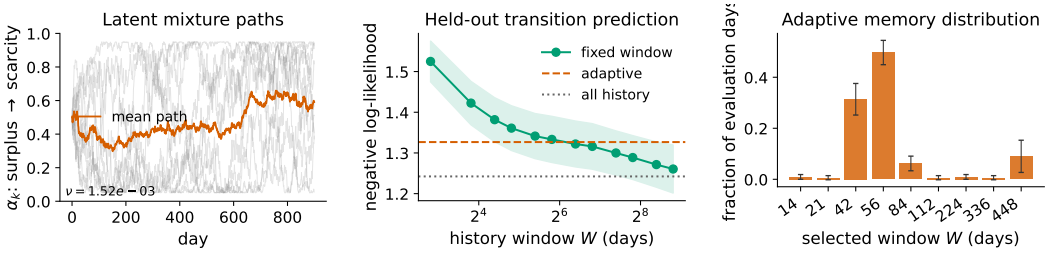
Controllers and evaluation. We compare all-history estimation, twelve fixed windows from 7 to 448 days, and an adaptive multi-scale rule. The adaptive rule minimizes an empirical uncertainty-plus-mismatch score and uses neither ν nor the current kernel. Dynamic programming under the current kernel defines regret. Evaluation runs from day 448 through day 898. One development path sets the score weight; ten independent paths give the reported results. This selector tests the application-level principle, while the preceding subsection evaluates ADA-FRESH with MVP.

Decision-level findings. The fixed-window regret curve is U-shaped: short histories lack data, while long histories average over old regimes. Adaptive regret is 1.056 times the best fixed-window regret, with 95% interval $[0.976, 1.178]$. Relative to all-history learning, it reduces regret by 44.7% (95% interval $[20.7\%, 61.4\%]$) and wins on all ten paths. The raw cost reduction is 0.107%, but

Table 1: Wind-storage control on ten held-out diffusion paths. The best fixed window is selected separately for each metric.

Metric	Best fixed	Adaptive	All history
Dynamic regret/day	0.005898 ($W = 84$)	0.006231	0.011267
Transition NLL	1.260531 ($W = 448$)	1.326729	1.242318
Operating cost/day	4.492430 ($W = 112$)	4.493103	4.497935

it removes 87.8% of the controllable gap (95% interval $[66.3\%, 94.9\%]$). All-history learning has 6.80% better one-step likelihood but worse regret and cost. Predictive fit and decision value therefore favor different memory lengths.



(a) Latent paths between empirical regimes.

(b) Held-out transition prediction.

(c) Distribution of selected memory lengths.

Figure 5: Additional wind-storage diagnostics. The latent process moves repeatedly between OPSD-derived surplus and scarcity regimes. All-history learning gives the best one-step likelihood, while the adaptive controller mainly selects finite memory lengths. Shading and error bars, where shown, are 95% intervals over ten held-out paths.