NEAR-OPTIMAL REGRET FOR CONSTRAINED MDPS WITH HARD STOPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Constrained Markov Decision Processes with Hard Stopping, which generalize Contextual Bandits with Knapsacks (CBwK) and can be viewed as a hard-constraint variant of Constrained Markov Decision Processes (CMDPs). In each of T episodes, the agent interacts with the environment through an H-step MDP, where at each step it takes an action based on the current state and receives stochastic reward and cost. The objective is to maximize cumulative reward until the end of T episodes or the cumulative cost exceeds a budget B. We propose Agressive Online Gradient Descent (AOGD) and prove an upper regret bound of $\tilde{O}\big((1+\frac{1}{\sqrt{b}})H^2S\sqrt{AT}\big)$, where $b = B/T$ is the average budget per episode, S, A are the numbers of states and actions. Up to our knowledge, this is the first algorithm that can achieve sublinear regret guarantee as long as the budget $B = \tilde{\omega}(1)$. We also establish a lower bound of $\Omega\big(\sqrt{H^4SAT/b}\big)$, matching our upper bound up to a factor of $\sqrt{S}$, and showing our bounds are near optimal in the key parameters T and b. Further, we also show that with a sliding-window design, our algorithm would also have strong performance under a non-stationary setting.

1 INTRODUCTION

In Constrained Markov Decision Processes (CMDPs) (Efroni et al., 2020; Liu et al., 2021; Ding et al., 2021), an agent runs an H-step MDP in each of T episodes and maximizes cumulative reward subject to cost constraints. Such models arise in resource-limited marketing, scheduling, and healthcare (Xiao et al., 2019; Fountoulakis et al., 2023; Yao et al., 2022).

Existing work on CMDPs generally assumed the existence of a stationary policy that satisfies the cost constraints throughout Efroni et al. (2020); Liu et al. (2021); Wei et al. (2022b); Zhou et al. (2023). The goal is then to minimize regret with respect to both total reward and constraint violations.

A related special case is Contextual Bandits with Knapsacks (CBwK) (Agrawal et al., 2016), corresponding to $H = 1$. Unlike soft-constraint CMDPs, CBwK stops interaction once total cost exceeds the budget B and retains only previously accumulated reward. Some previous work assumes the existence of null action which always yields zero reward and zero cost, or allows skipping episode (Agrawal et al., 2016; Slivkins et al., 2023). This assumption does not always hold in reality, however, especially when the fairness constraints as well as non-stationary cases come into sight (Chzhen et al., 2023).

The existing work on CBwK would also assume that there is a stationary policy that satisfies the budget constraint throughout Han et al. (2023); Chzhen et al. (2023); Guo and Liu (2025), and defined the benchmark to be the optimal total reward by such policies.

Null actions or skipping episode can nevertheless be impractical, and full-horizon feasibility can be impossible, as in critical-care scarcity during pandemic surges (White and Lo, 2020). Even when a feasible policy exists, early stopping can be preferable: for two actions with reward–cost pairs $(1, 1)$ and $(0.1, 0.5)$ and budget $T/2$, the first action exhausts the budget early but earns substantially more reward.

Our Model. This motivates studying *Constrained MDPs with Hard Stopping*, in which the agent aims to maximize cumulative reward until either the end of T episodes or the total cost exceeds

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

the budget. Importantly, *we do not assume the existence of a feasible policy, as well as null action or voluntary skipping*. We analyze regret against the optimal expected credited reward of a full-information strategy under the same physical stopping rule, while previous work in CBwK used a stationary policy safe throughout whole T episodes as benchmark (Guo and Liu, 2025).

Our Results. We propose *Aggressive Online Gradient Descent (AOGD)*, which updates the Lagrange multiplier associated with the budget constraint using aggressive projected gradient descent. Gradient descent on the Lagrangian is a natural and classic approach. However, naïve gradient descent, to some extent, wastes too much attention on non-optimal conservative policies, especially when the feasible margin is small, leading to great fluctuation around the extreme point. We discover that it is possible to reduce the penalty on cost adaptively, which allows to control the wastes of costs and also achieve a better regret. **The key novelty of our algorithm** is that before executing a policy, AOGD checks whether its optimistic predicted reward-minus-priced-cost value under the current Lagrange multiplier is nonnegative. If it is negative, the algorithm repeatedly decreases the multiplier and replans until the value becomes nonnegative. These additional updates aggressively reduce the penalty on cost, placing greater emphasis on reward in policy selection. Under the hard-stopping scenario, we show that this aggressive selection would have a better regret guarantee:

We prove that with high probablity, the regret of AOGD in terms of optimal expected credited reward of a full-information strategy is bounded by

$$\mathrm{OPT}_{\mathrm{HS}} - \mathsf{Rew}_T \le \widetilde{O}\left(H^2 S\sqrt{AT} + vHS\sqrt{\frac{AT}{b}} + H^{3/2}\sqrt{\frac{T}{b}} + \frac{vH^3S^2A}{b} + \frac{H^2}{b} \right).$$

with high probability, where S and A are the sizes of the state and action spaces, $b = B/T$ is the per-episode budget, and $v \le H$ is the optimal per-episode value of the LP benchmark. We complement the algorithmic result with a hard instance on which every algorithm has regret at least $\Omega\big(\sqrt{\frac{H^4SAT}{b}}\big)$, matching our upper bound up to a $\sqrt{S}$ factor.

Interestingly, our algorithm and its regret upper bound also improve or match the state-of-the-art of CMDP and CBwK in the dependence on b, and T. This is the first algorithm that could achieve sublinear regret as long as $B = \tilde{\omega}(1)$, while all previous work requires at least $B = \tilde{\omega}(\sqrt{T})$ and other assumptions. See Table 1 for a summary.

Further, we extend our algorithm to non-stationary setting and show its robustness. With well chosen sliding window size w, we could also show that with fixed bounded total variation of environment (reward, cost and transition kernels), with high probability, the regret could be bounded as

$$\widetilde{O}_{H,S,A}\Big(b^{-2/3}T^{2/3} + \frac{\sqrt{T}}{b}\Big)$$

when $b = o(1)$. Here constants depending on H, S, A are hidden for easier reading. Previous work in non-stationary CBwK (Slivkins et al., 2023) requires the algorithm to continue even after the hard stopping until the time horizon to get a regret bound. Our paper is the first to offer a regret guarantee for non-stationary hard-stopping scenario, both in CBwK and CMDP. Details are in Section 7.

2 Related work

CBwK. Contextual bandits and bandits with knapsacks have been studied extensively. We only focus on the literature most related to ours. CBwK was introduced by Agrawal et al. (2016). Oracle-based methods include SquareCB (Foster and Rakhlin, 2020) and the packing framework of Slivkins et al. (2023). However, both of their methods require the null action or skipping episode mechanism, which cannot be applied directly under fairness concern or non-stationary environment.

Future work focuses on more strict setting, e.g. replacing the null action assumption with feasibility margin. For sublinear budgets, Han et al. (2023) designed a two-stage algorithm and obtained $\widetilde{O}((1+v/(\delta b))\sqrt{T})$ regret when $B = \Omega(T^{3/4})$, where δb is the feasible margin; Chzhen et al. (2023) and Guo and Liu (2025) relaxed this requirement to $B = \Omega(\sqrt{T})$. These guarantee use feasibility assumptions or a known margin; Guo and Liu (2025) also gives a weaker bound $\tilde{\mathcal{O}}(\sqrt{T} + \frac{v}{\sqrt{b}}T^{3/4})$

Table 1: Comparison with related work. Here $\delta b \leq b$ is the feasibility margin. The regret for CMDP row hides b from the $\tilde{O}$ notation for easier comparison, but the actual requirement for B to get a sublinear regret is listed at the last column. Note that all other algorithms in CMDP could only bound the violation of budget with high probability, while our algorithm stops after the first budget-crossing episode and discards that episode's reward. For CBwK, LagrangeCBwLC requires the existence of null action, and PGD requires a feasibility margin δ. AOGD makes no assumption on problem instance.

	Algorithm	Regret	Requirement for Budget B to get sublinear regret
CMDP	OptDual-CMDP (Efroni et al., 2020)	$\tilde{\mathcal{O}}(\sqrt{H^4S^3AT})$	$\tilde{\omega}(\sqrt{T})$
	CONRL(Brantley et al., 2020)	$\tilde{\mathcal{O}}(H^{2.5}S\sqrt{AT})$	$\tilde{\omega}(\sqrt{T})$
	OptPess-LP (Liu et al., 2021)	$\tilde{\mathcal{O}}(\sqrt{H^6S^3AT})$	$\tilde{\omega}(\sqrt{T})$
	Triple-Q(Wei et al., 2022b)	$\tilde{\mathcal{O}}(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}T^{\frac{4}{5}})$	$\tilde{\omega}(T^{4/5})$
	AOGD	$\tilde{\mathcal{O}}\left(\sqrt{H^4S^2AT}\right)$	$\tilde{\omega}(1)$
CBwK	LagrangeCBwLC(Slivkins et al., 2023)	$\tilde{\mathcal{O}}(\frac{1}{b}\sqrt{T})$	$\tilde{\omega}(\sqrt{T})$
	PGD(Chzhen et al., 2023)	$\tilde{\mathcal{O}}(\sqrt{T}+\frac{v}{\delta b}\sqrt{T})$	$\tilde{\omega}(\sqrt{T})$
	AUPD(Guo and Liu, 2025)	$\tilde{\mathcal{O}}(\sqrt{T}+\frac{v}{\sqrt{b}}T^{3/4})$	$\tilde{\omega}(\sqrt{T})$
	AOGD	$\tilde{\mathcal{O}}\left(\sqrt{T}+\frac{v}{\sqrt{b}}\sqrt{T}\right)$	$\tilde{\omega}(1)$
AOGD (Our algorithm)		$\tilde{\mathcal{O}}\left(H^2S\sqrt{AT}+vHS\sqrt{\frac{AT}{b}}\right.$ $\left.+H^{3/2}\sqrt{\frac{T}{b}}+\frac{vH^3S^2A}{b}+\frac{H^2}{b}\right)$	$\tilde{\omega}(1)$

without strict feasibility. This is the first paper towards no strict feasibility assumption and achieves sublinear regret result.

Another line of literature assumes that the interaction could continue even when the budget is exhausted (Bernasconi et al., 2024a;b; Kumar and Kleinberg, 2022). Kumar and Kleinberg (2022) considers replenishable resource utilization under stochastic bandits with knapsacks setting. Note that they also make an explicit assumption on null action to make the budgets replenished. Bernasconi et al. (2024b) extended this setting into adversarial scenarios and achieved "best-of-both-worlds" regret guarantees. Almost the same time Bernasconi et al. (2024a) studied adversarial bandits with general constraints, where constraint violations are allowed and no hard stopping is imposed.

CMDPs. Model-based episodic CMDP algorithms were developed in Brantley et al. (2020); Efroni et al. (2020); Singh et al. (2020); later work obtained near-optimal regret with zero or bounded violation (Liu et al., 2021; Bura et al., 2022). Wei et al. Wei et al. (2022b) gave the first model-free sublinear-regret guarantee for tabular CMDPs. Extensions cover linear CMDPs (Ghosh et al., 2022; Ding et al., 2021), infinite-horizon problems (Chen et al., 2022; Wei et al., 2022a), and best-policy identification (Zhou et al., 2023). However, all work mentioned above could only offer a high-probability bound for constraint violation. In our hard stopping scenario, it is hard to have a good analysis on their regret since hard stopping requires the constraint to be obeyed almost surely.

3 PRELIMINARIES

Consider a stationary finite-horizon MDP with S states, A actions, horizon H, and at most T episodes. Its initial-state law is μ and its stage-h transition kernel is P_h. At each step, the observed reward and cost lie in $[0, 1]$. Given the past and the current state-action pair, their means are $r_h(s, a)$ and $c_h(s, a)$, and the next state has law $P_h(\cdot \mid s, a)$. Reward, cost, and next-state noise within a step need not be independent. Each episode starts from the same conditional law μ. The model is unknown.

Let Π be the set of Markov policies $\pi = (\pi_1, \ldots, \pi_H)$, where $\pi_h : \mathcal{S} \to \mathcal{A}$. Policies may depend on the stage, although the MDP is unchanged across episodes. Their expected reward and cost are

$$R(\pi) = \mathbb{E}_{\mu,P,\pi} \sum_{h=1}^{H} r_h(s_h, a_h), \qquad C(\pi) = \mathbb{E}_{\mu,P,\pi} \sum_{h=1}^{H} c_h(s_h, a_h). \tag{1}$$

Both lie in $[0, H]$. A policy is selected before the episode's initial state is drawn, using only previous observations.

The total budget is $B = Tb$, where $0 < b \leq H$. Every episode that is started executes all H steps with an ordinary policy. If its cost makes cumulative spending exceed B, its reward is discarded and interaction ends. We **DO NOT** assume null action or voluntary skipping. Write G_k^r, G_k^c for the realized episode totals and N_{valid} for the number of episodes whose rewards are retained. The learner's credited reward is $\mathsf{Rew}_T = \sum_{k=1}^{N_{\text{valid}}} G_k^r$.

We compare this reward with the stopping LP. For $B' \in [0, HT]$, define

$$\mathrm{LP}_T(B') = \max_{\omega \geq 0} \left\{ \sum_{\pi \in \Pi} \omega_\pi R(\pi) : \sum_{\pi} \omega_\pi \leq T, \quad \sum_{\pi} \omega_\pi C(\pi) \leq B' \right\}. \tag{2}$$

Here ω_π is the expected number of episodes allocated to policy π. Let $v = \mathrm{LP}_T(B)/T$ and define

$$\mathrm{Reg}_{\mathrm{LP}} = Tv - \mathsf{Rew}_T, \qquad \mathrm{Reg}_{\mathrm{HS}} = \mathrm{OPT}_{\mathrm{HS}} - \mathsf{Rew}_T, \tag{3}$$

where $\mathrm{OPT}_{\mathrm{HS}}$ is the optimal expected credited reward of a full-information strategy under the same physical stopping rule. No ordinary policy is assumed feasible at expected cost b. Lemma B.1 shows that $\mathrm{OPT}_{\mathrm{HS}} \leq Tv + Hv/b$, so an LP-regret bound also controls regret against the hard-stopping oracle.

We use natural logarithms, $x_+ = \max\{x, 0\}$, $x \vee y = \max\{x, y\}$, and $[n] = \{1, \ldots, n\}$. For a finite set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes its probability simplex. The notation $\widetilde{O}$ hides logarithmic factors in $S, A, H, T, 1/\delta$, not polynomial dependence on b.

4 Algorithm

To address the budget constraint, the algorithm considers the Lagrange $R(\pi) - \lambda C(\pi)$. At the beginning of each episode k, the algorithm chooses a policy with a large value of $R(\pi) - \lambda_k C(\pi)$, through a planner $\mathrm{Plan}_k(\lambda_k)$ using estimated confidence sets and a chosen λ_k. This favors policies that may be rewarding or inexpensive; the multiplier λ_k is the cost price determining how much the learner is willing to spend. Details about the multiplier and planner design are introduced below.

Lagrange multiplier λ_k. We maintain a nonnegative queue Z and set the cost price λ_k to Z/U, where $U > 0$ is fixed. After an episode whose reward is retained, the queue increases by the estimated cost excess $\underline{C}_k(\pi_k) - b$, with $\underline{C}_k(\pi_k)$ the lower estimation of cost, and is projected onto $[0, \infty)$. This is online gradient descent (OGD) in the multiplier, with step size $1/U$.

As for this natural algorithm, the challenge is that the Lagrangian multiplier is hard to bound. When the feasible margin is too small, it would lead to great fluctuations around the extreme point. Guo and Liu (2025) used the Lyapunov method for analysis under the CBwK setting. Their fix is to bound the expectation of the cumulative Lagrangian multiplier instead, which yields a relatively weak bound and thus leads to a weak final regret upper bound.

To address the challenge, our algorithm introduces one additional key step before execution in the ordinary OGD. Denote the estimated value of $\max_\pi R(\pi) - \lambda_k C(\pi)$ as $J_k(\lambda_k)$. If $J_k(Z/U) < 0$, decrease Z by b and maximize again, using the same estimates. This is to force the multiplier to be small enough to reduce the penalty on cost, while emphasizing more on reward. The reason is that although aggressive policy chosen may lead to budget exhausted, the regret can somehow be balanced through hard stopping mechanism; On the other hand, a conservative policy may cost less, but the lost reward cannot be gained through more episodes. Overall an aggressive policy is more favored than a conservative policy in this setting. The loop terminates because $J_k(0)$ is nonnegative.

Joint backward planning Plan_k. The key idea for the joint planner is to use backward DP to find an optimal policy that maximizes $R(\pi) - \lambda_k C(\pi)$. Let $r_{k,h}^+(s, a)$ be the clipped Hoeffding upper bound on immediate reward and $c_{k,h}^-(s, a)$ the Bernstein lower endpoint for immediate cost. The Bernstein lower bound estimation for cost function is necessary because it helps bound the estimation error of cost function by $\sqrt{B}$ instead of $\sqrt{T}$, and thus achieves better result in regret. Let $\mathcal{P}_{k,h}^{\mathrm{J}}(s, a)$

Algorithm 1 Aggressive Online Gradient Descent (AOGD)

Require: $T, H, B = Tb$, confidence $\delta \in (0,1)$, prescribed $U > 0$

Set $Z \leftarrow 0$ and cumulative realized cost $C_{\text{used}} \leftarrow 0$
for $k = 1, \dots, T$ **do**
  Construct the clipped Hoeffding upper bound $r^+_{k,h}(s,a)$ for immediate reward, and the Bernstein lower bound $c^-_{k,h}(s,a)$ for immediate cost
  Construct the confidence sets of transition kernels $\mathcal{P}^{\mathrm{J}}_{k,h}(s,a)$ and the initial occupancy measure $\mathcal{M}^{\mathrm{J}}_k$ from past trajectories
  $(J, \pi) \leftarrow \mathrm{Plan}_k(Z/U, r^+_{k,h}, c^-_{k,h}, \mathcal{P}^{\mathrm{J}}_{k,h}, \mathcal{M}^{\mathrm{J}}_k)$
  while $J < 0$ **do**
    $Z \leftarrow [Z - b]_+$ ▷ internal computation only
    $(J, \pi) \leftarrow \mathrm{Plan}_k(Z/U, r^+_{k,h}, c^-_{k,h}, \mathcal{P}^{\mathrm{J}}_{k,h}, \mathcal{M}^{\mathrm{J}}_k)$
  end while
  Set $Z_k \leftarrow Z$, $\lambda_k \leftarrow Z_k/U$
  Set $\pi_k \leftarrow \pi$; evaluate $\underline{C}_k(\pi_k)$ by backward recursion
  Execute π_k for all H steps; observe its complete trajectory, the cost G^c_k, and the reward G^r_k
  $C_{\text{used}} \leftarrow C_{\text{used}} + G^c_k$
  if $C_{\text{used}} > B$ **then**
    Discard G^r_k and **stop** ▷ no crossing-episode queue update
  end if
  Credit G^r_k; set $Z \leftarrow [Z_k + \underline{C}_k(\pi_k) - b]_+$
  Store this trajectory
end for

Algorithm 2 Joint backward planner Plan_k

Require: price $\lambda \geq 0$, and $r^+_{k,h}, c^-_{k,h}, \mathcal{P}^{\mathrm{J}}_{k,h}, \mathcal{M}^{\mathrm{J}}_k$

Set $W^\lambda_{k,H+1}(s) \leftarrow 0$ for every $s \in \mathcal{S}$
for $h = H, H-1, \dots, 1$ **do**
  for all $s \in \mathcal{S}$ **do**
    for all $a \in \mathcal{A}$ **do**
      Choose $p^\lambda_{k,h}(\cdot \mid s,a) \in \arg\max_{p \in \mathcal{P}^{\mathrm{J}}_{k,h}(s,a)} p^\top W^\lambda_{k,h+1}$
      $Q^\lambda_{k,h}(s,a) \leftarrow r^+_{k,h}(s,a) - \lambda c^-_{k,h}(s,a) + (p^\lambda_{k,h}(\cdot \mid s,a))^\top W^\lambda_{k,h+1}$
    end for
    $\pi^\lambda_{k,h}(s) \leftarrow \arg\max_{a \in \mathcal{A}} Q^\lambda_{k,h}(s,a)$
    $W^\lambda_{k,h}(s) \leftarrow Q^\lambda_{k,h}(s, \pi^\lambda_{k,h}(s))$
  end for
end for
Choose $\mu^\lambda_k \in \arg\max_{\mu' \in \mathcal{M}^{\mathrm{J}}_k} (\mu')^\top W^\lambda_{k,1}$
$J_k(\lambda) \leftarrow (\mu^\lambda_k)^\top W^\lambda_{k,1}$
return $(J_k(\lambda), \pi^\lambda_k)$

and $\mathcal{M}^{\mathrm{J}}_k$ be the intersections of the Bernstein and Hoeffding confidence sets for the transition row and initial law, respectively. Their exact definitions are in Appendix A.1. Starting from $W^\lambda_{k,H+1} = 0$, compute

$$W^\lambda_{k,h}(s) = \max_a \left\{ r^+_{k,h}(s,a) - \lambda c^-_{k,h}(s,a) + \max_{p \in \mathcal{P}^{\mathrm{J}}_{k,h}(s,a)} p^\top W^\lambda_{k,h+1} \right\},$$
$$J_k(\lambda) = \max_{\mu' \in \mathcal{M}^{\mathrm{J}}_k} (\mu')^\top W^\lambda_{k,1}. \tag{4}$$

Here W is the statewise Lagrangian value and J its initial-state value. The maximizing actions define π^λ_k. Joint planning uses one transition model for both reward and cost, so it need not maximize their separately optimized difference. Lemma A.1 gives the comparison needed for the proof. The queue still uses $\underline{C}_k(\pi_k)$, since the cost under the jointly selected model need not itself be a lower bound.

5 REGRET UPPER BOUND GUARANTEES

Learning costs matters because excess spending can remove future episodes. The analysis uses the budget twice: first to control uncertainty in total cost, and then to control the accumulated squared queue updates. These two steps give the square-root dependence on $1/b$ in the leading cost term.

Theorem 5.1 (Stationary regret upper bound)**.** *For any $\delta \in (0,1)$, run Algorithm 1 with*

$$U = b\sqrt{2Tb/H + 2[1+\log(8/\delta)]}. \tag{5}$$

With probability at least $1-\delta$, both $\mathrm{Reg}_{\mathrm{LP}}$ and $\mathrm{Reg}_{\mathrm{HS}}$ are at most

$$\widetilde{O}\left(H^2 S\sqrt{AT} + vHS\sqrt{\frac{AT}{b}} + H^{3/2}\sqrt{\frac{T}{b}} + \frac{vH^3S^2A}{b} + \frac{H^2}{b}\right). \tag{6}$$

The learner need not know v.

The first term is reward learning. The second and fourth are the square-root and additive parts of cost learning. The third and fifth cover queue control, observation noise, and the final budget crossing. The factor v/b behind the cost terms has a simple interpretation: an extra unit of spending can displace about $1/b$ episodes, each worth v to the LP. The exact finite-sample bound is Theorem B.14.

5.1 PROOF SKETCH

Let N_{valid} be the number of episodes whose rewards are retained, and let N_{attempt} include the crossing episode, if one occurs. N_{attempt} either equals to $N_{\mathrm{valid}}+1$ or T. We start from the regret decomposition

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{LP}} &= v(T-N_{\mathrm{valid}}) + \sum_{k=1}^{N_{\mathrm{valid}}}(v - R(\pi_k)) + \sum_{k=1}^{N_{\mathrm{valid}}}(R(\pi_k) - G_k^r) \\
&\le \underbrace{\sum_{k=1}^{N_{\mathrm{valid}}}(\overline{R}_k(\pi_k) - R(\pi_k))}_{\text{reward estimation}} + \underbrace{\sum_{k=1}^{N_{\mathrm{valid}}}(R(\pi_k) - G_k^r)}_{\text{reward noise}} \\
&\quad + \underbrace{v(T-N_{\mathrm{valid}}) - \sum_{k=1}^{N_{\mathrm{valid}}}\lambda_k(\underline{C}_k(\pi_k) - b)}_{\text{stopping and dual updates}}.
\end{aligned} \tag{7}$$

Here $\overline{R}_k(\pi) \in [0,H]$ is the auxiliary Hoeffding upper bound on the expected reward of policy π, defined by (A.9)–(A.10) in Appendix A.3. The inequality is the accepted-policy certificate in Lemma B.11: joint optimism and the nonnegative-score test give $v - R(\pi_k) \le \overline{R}_k(\pi_k) - R(\pi_k) - \lambda_k(\underline{C}_k(\pi_k) - b)$. We bound the three terms below. The stopping and dual terms must be handled together; this is where cost estimation enters the regret bound.

1. Reward estimation (Lemma B.4). The cumulative error of the auxiliary reward envelope satisfies

$$\sum_{k=1}^{N_{\mathrm{valid}}}(\overline{R}_k(\pi_k) - R(\pi_k)) \le E_{\mathrm{H}} = \widetilde{O}(H^2 S\sqrt{AT}).$$

The proof propagates one-step Bellman errors along the true dynamics, then sums the confidence widths using visit counts. The transition ℓ_1 confidence bound holds for every bounded downstream vector, so it also covers the history-dependent policy selected by the joint planner.

2. Observation noise (Lemmas B.9 and B.10). Reward noise and the cost noise needed below satisfy

$$\sum_{k=1}^{N_{\mathrm{valid}}}(R(\pi_k) - G_k^r) \le \overline{M}_R = \widetilde{O}(H\sqrt{T}+H), \qquad \sum_{k=1}^{N_{\mathrm{valid}}}(G_k^c - C(\pi_k)) \le \overline{M}_C = \widetilde{O}(\sqrt{HTb}+H).$$

Concentration is applied to all attempted episodes, since whether an episode is attempted is known before it starts. Removing a crossing increment afterward costs at most H. For cost, the sum

of conditional variances is at most $H\sum_{k=1}^{N_{\text{attempt}}} C(\pi_k)$; the budget lemma controls this sum by $\widetilde{O}(Tb+H)$. This gives the budget-sensitive cost-noise bound.

3. Stopping and dual updates (Lemmas B.11 and B.12). Write $g_k = \underline{C}_k(\pi_k) - b$. Let E_C bound the cumulative cost-underestimation error over attempted episodes, and define

$$Q_{\text{sq}} = \sum_{k=1}^{N_{\text{valid}}} g_k^2 + N_{\text{int}} b^2,$$

where N_{int} is the number of internal decrements actually performed. Let Z_{end} denote the final actual queue. If the budget is crossed, the discarded episode costs at most H, so the retained episodes have total realized cost greater than $Tb - H$. Combining this observation with the cost-error and noise bounds, and then using the queue recursion, gives the following (which leads to the key improvement compared with ordinary OGD):

$$\begin{aligned} b(T - N_{\text{valid}}) &\le H + E_C + \overline{M}_C + \sum_{k=1}^{N_{\text{valid}}} g_k \\ &\le H + E_C + \overline{M}_C + Z_{\text{end}} + bN_{\text{int}}. \end{aligned}$$

If no crossing occurs, then $N_{\text{valid}} = T$, and the same inequality holds trivially.

Multiplying this bound by v/b and applying the exact queue-potential inequality B.12 induced from projected OGD yields

$$\begin{aligned} & v(T - N_{\text{valid}}) - \sum_{k=1}^{N_{\text{valid}}} \lambda_k g_k \\ & \le \frac{v}{b}(H + E_C + \overline{M}_C) + \frac{v}{b} Z_{\text{end}} + vN_{\text{int}} - \frac{Z_{\text{end}}^2}{2U} + \frac{Q_{\text{sq}}}{2U} - b\sum_{j=1}^{N_{\text{int}}} \lambda_j^{\text{st}} \\ & \le \frac{v}{b}(H + E_C + \overline{M}_C) + \frac{Q_{\text{sq}}}{2U} + \frac{Uv^2}{2b^2}. \end{aligned} \tag{8}$$

Here comes the key effect of the aggressive decrements design: Indeed, the negative-score certificate gives $\lambda_j^{\text{st}} \ge v/b$ before every internal decrement, and hence $vN_{\text{int}} \le b\sum_j \lambda_j^{\text{st}}$; the two internal-update terms therefore cancel. The remaining terminal-queue terms satisfy

$$\frac{v}{b} Z_{\text{end}} - \frac{Z_{\text{end}}^2}{2U} = \frac{Uv^2}{2b^2} - \frac{(Z_{\text{end}} - Uv/b)^2}{2U} \le \frac{Uv^2}{2b^2}.$$

Thus the aggressive decrements remove the cost of the internal updates, while the retained negative square controls the final queue.

It remains to bound the two quantities E_C and Q_{sq} in (8):

3.1 Cost estimation (Theorem B.8 and Lemma B.9). Bernstein confidence gives

$$E_C = \widetilde{O}\Big(HS\sqrt{A(Tb+H)} + H^3S^2A\Big).$$

The key fact is that the expected sum of conditional variances of the true cost-to-go is at most $HC(\pi)$ in an episode. The proof first obtains a bound in terms of $\sum_{k=1}^{N_{\text{attempt}}} C(\pi_k)$; the hard budget then replaces that sum by $\widetilde{O}(Tb+H)$. To use this variance fact for the robust lower values, we charge their difference from the true values to the same estimation error and absorb the resulting square-root term. Primitive confidence sets make the argument valid for the selected policy. Multiplication by v/b gives the second and fourth terms of (6).

3.2 Squared queue updates (Lemma B.13). With the same N_{int} as above,

$$Q_{\text{sq}} = \sum_{k=1}^{N_{\text{valid}}} (\underline{C}_k(\pi_k) - b)^2 + N_{\text{int}} b^2 \le \overline{Q}_{\text{sq}}, \qquad \overline{Q}_{\text{sq}} = 2HTb + 2H^2[1 + \log(8/\delta)].$$

The proof uses $\underline{C}_k(\pi_k)^2 \le H\underline{C}_k(\pi_k)$ and charges internal queue decreases to earlier positive increases. Thus squared updates are controlled by spending, not just by the number of episodes. For the choice $U = (b/H)\sqrt{\overline{Q}_{\mathrm{sq}}}$ in (5), $v \le H$ gives

$$\frac{\overline{Q}_{\mathrm{sq}}}{2U} + \frac{Uv^2}{2b^2} \le \frac{H}{b}\sqrt{\overline{Q}_{\mathrm{sq}}} = \widetilde{O}\Big(H^{3/2}\sqrt{T/b} + H^2/b\Big).$$

Substituting these bounds into (7) and (8) proves the LP-regret rate. Lemma B.1 adds at most Hv/b for $\mathrm{Reg}_{\mathrm{HS}}$, which is absorbed by the last term of (6). All bounds hold on a common event of probability at least $1-\delta$; Appendix B gives their full proofs.

6 Regret Lower Bound Guarantees

The dependence on the budget is unavoidable even when rewards are known: the learner can lose reward simply because identifying inexpensive actions uses budget and makes the interaction stop earlier.

Theorem 6.1 (Expected hard-stopping regret lower bound)**.** *There are universal constants $c_0, C_0 > 0$ such that, whenever $H \ge 4$, $S, A \ge 2$, $0 < b \le H/4$, and $T \ge C_0 H^2 S(A-1)/b$, every learner has a stationary instance with $v = 2H/3$ on which*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{HS}}] = \mathrm{OPT}_{\mathrm{HS}} - \mathbb{E}[\mathsf{Rew}_T] \ge c_0 H^2 \sqrt{\frac{S(A-1)T}{b}}. \tag{9}$$

The instance follows the same hard-stopping protocol and has no null action or skipping option.

Construction and proof idea. Intuitively, the instance turns a small identification error into early stopping. A wrong action in the first half slightly raises the chance of entering a persistent bad state, which then consumes cost throughout the second half. These amplified costs exhaust the budget even for an informed policy; since rewards are identical, the resulting lost episodes are the sole source of regret.

Set $m = \lfloor H/2 \rfloor$, $S_0 = \lfloor S/2 \rfloor$, and use the first m steps for decisions. At each step–context pair (h, x), one hidden action triggers the bad state with probability p, while every other action uses $p + g$. Every step still gives reward one. A deterministic base cost calibrates the informed policy's expected episode cost to $3b/2$, giving $v = 2H/3$.

Set $g = \Theta(\sqrt{S(A-1)b/(H^2T)})$. A testing and stopping argument over the mS_0 independent identification problems forces $\Omega(mT)$ wrong actions while the process is safe. Cost amplification then gives total excess spending $\Omega(\sqrt{m(H-m)S(A-1)Tb})$. Relative to the informed policy, this removes order $1/b$ credited episodes per unit of excess cost; each removed episode is worth H. Using $m(H-m) = \Theta(H^2)$ yields the stated $\Omega(H^2\sqrt{S(A-1)T/b})$ regret. The comparison uses the same physical stopping rule for both policies, so it is directly against $\mathrm{OPT}_{\mathrm{HS}}$. Appendix C gives the construction, explicit constants, and full proof.

7 Tracking a changing environment

A sliding window lets the same estimates track a changing model. Recent data reduce bias from past environments, while a longer window reduces sampling error. The resulting tradeoff is between $T/\sqrt{w}$ and accumulated drift over a window of w episodes.

Suppose the model sequence is fixed before interaction, the initial law μ is unchanged, and rewards, costs, and transitions may vary between episodes. The learner knows nonnegative bounds d_j^r, d_j^c, d_j^P on each change:

$$\|r_{j+1} - r_j\|_\infty \le d_j^r, \qquad \|c_{j+1} - c_j\|_\infty \le d_j^c, \qquad \max_{h,s,a} \|P_{j+1,h}(\cdot \mid s,a) - P_{j,h}(\cdot \mid s,a)\|_1 \le d_j^P.$$

The scalar norms range over all stages, states, and actions. Write $\mathcal{V}_g = \sum_{j=1}^{T-1} d_j^g$ for $g \in \{r, c, P\}$ and $\mathcal{V} = \mathcal{V}_r + \mathcal{V}_c + \mathcal{V}_P$ for total variation. Known local bounds, not just their total, are used to construct confidence sets.

Let $R_k(\pi), C_k(\pi)$ be the expected reward and cost in episode k, extended linearly to mixtures of policies. The dynamic benchmark is

$$v_k^* = \max_{q \in \Delta(\Pi):\, C_k(q) \le b} R_k(q), \qquad \mathrm{Reg}_{\mathrm{dyn}}(T) = \sum_{k=1}^{T} v_k^* - \mathsf{Rew}_T.$$

Assume the feasible set is nonempty in every episode. Unlike the stationary LP, each comparator mixture has total mass one on ordinary policies.

Estimate reward, cost, and transitions from the last w episodes, and enlarge the confidence sets by the known drift within the window. The initial-law estimate still uses all previous initial states. Here we use a queue cap in place of the stationary sign test: decrease Z by b while $Z > UH/b$, then run $\mathrm{Plan}_k^w(Z/U)$ using the drift-enlarged confidence sets. The superscript w denotes the window-based estimates. This cap gives the certificate needed to control the rewards of different future optimal policies. All physical execution and stopping rules are unchanged. Algorithm 3 specifies the full procedure.

Cost and transition drift can remove future episodes, so their contribution is weighted by $1/b$. Define the budget-weighted variation

$$\mathcal{V}_b = \mathcal{V}_r + \frac{\mathcal{V}_c + \mathcal{V}_P}{b}.$$

It is known from the supplied envelopes. The next theorem keeps every dependence on b explicit; b may vary with T.

Theorem 7.1 (Budget-explicit dynamic regret)**.** *Under these assumptions, let $0 < b \le H$, $\delta \in (0,1)$, and choose U as in* (5)*. If $\mathcal{V}_b > 0$, use*

$$w = \left\lceil \min\left\{ T, \max\left\{ 1, \left(\frac{(1 + b^{-1/2})T}{\mathcal{V}_b} \right)^{2/3}, \sqrt{\frac{T}{b\mathcal{V}_b}} \right\} \right\} \right\rceil; \tag{10}$$

if $\mathcal{V}_b = 0$, use $w = T$. With probability at least $1 - \delta$,

$$\mathrm{Reg}_{\mathrm{dyn}}(T) \le \widetilde{O}_{H,S,A}\Big((1 + b^{-1/2})\sqrt{T} + \frac{1}{b} + (1 + b^{-1/2})^{2/3} T^{2/3} \mathcal{V}_b^{1/3} + \sqrt{\frac{T\mathcal{V}_b}{b}} \Big). \tag{11}$$

The hidden constants may depend on H, S, A, but not on b, T, or the variation budgets.

The two window scales balance the square-root and additive sampling errors against drift. Both are needed when b is small. Each model change enters at most w windows; the cost-variance and queue arguments are unchanged. Appendix D gives the full dependence on H, S, A, the proof of (11), and its small-budget specialization.

8 Conclusion

We studied constrained MDPs with hard stopping, where the learner cannot take a null action or skip an episode. In the stationary setting, AOGD provides regret guarantees without requiring a policy that is feasible over the full horizon. The analysis uses cumulative spending to control both cost-estimation error and squared dual updates, yielding a $b^{-1/2}$ dependence in the leading regret terms. An expected regret lower bound shows that this budget dependence is unavoidable in the worst case, even when rewards are known.

We also developed a sliding-window variant for changing rewards, costs, and transitions. Under known local drift bounds and a feasible per-episode comparator, it gives a dynamic regret bound with explicit dependence on the budget and environmental variation. The bound makes clear how a small budget amplifies the effect of cost and transition changes on future reward.

9 AI USE STATEMENT

In this work, we have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation. Generating synthetic data sets, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, interpreting results are not applicable to this work. Additionally, we used generative AI tools for editing the research paper to improve readability. We have reviewed all AI-assisted work. The AI edition of paper is well checked by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 4–18. PMLR, 2016.

Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization, 2024a.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Bandits with replenishable knapsacks: The best of both worlds. In *The Twelfth International Conference on Learning Representations*, 2024b.

Kiante Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in neural information processing systems*, 35:1047–1059, 2022.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270. PMLR, 2022.

Evgenii Chzhen, Christophe Giraud, Zhen Li, and Gilles Stoltz. Small total-cost constraints in contextual bandits with knapsacks, with application to fairness. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3304–3312. PMLR, 2021.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3199–3210. PMLR, 2020.

Emmanouil Fountoulakis, Themistoklis Charalambous, Anthony Ephremides, and Nikolaos Pappas. Scheduling policies for AoI minimization with timely throughput constraints. *IEEE Transactions on Communications*, 71(7):3905–3917, 2023. doi: 10.1109/TCOMM.2023.3274143. URL `https://doi.org/10.1109/TCOMM.2023.3274143`.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315, 2022.

Hengquan Guo and Xin Liu. On stochastic contextual bandits with knapsacks in small budget regime. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FCMpUOZkxi.

Yuxuan Han, Jialin Zeng, Yang Wang, Yang Xiang, and Jiheng Zhang. Optimal contextual bandits with knapsacks under realizability via regression oracles. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5011–5035. PMLR, 2023.

Raunak Kumar and Robert Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. *Advances in Neural Information Processing Systems*, 35:19248–19259, 2022.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4633–4656. PMLR, 2023.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3868–3876, 2022a.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pages 3274–3307. PMLR, 2022b.

Douglas B. White and Bernard Lo. A framework for rationing ventilators and critical care beds during the covid-19 pandemic. *JAMA*, 323(18):1773–1774, 2020. doi: 10.1001/jama.2020.5046.

Shuai Xiao, Le Guo, Zaifan Jiang, Lei Lv, Yuanbo Chen, Jun Zhu, and Shuang Yang. Model-based constrained mdp for budget allocation in sequential incentive marketing. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 971–980, 2019.

Bing Yao, Yun Chen, and Hui Yang. Constrained Markov decision process modeling for optimal sensing of cardiac events in mobile health. *IEEE Transactions on Automation Science and Engineering*, 19(2):1017–1029, 2022. doi: 10.1109/TASE.2021.3052483. URL https://doi.org/10.1109/TASE.2021.3052483.

Zihan Zhou, Honghao Wei, and Lei Ying. Model-free, regret-optimal best policy identification in online cmdps. *arXiv preprint arXiv:2309.15395*, 2023.

A CONFIDENCE CONSTRUCTION AND BACKWARD INDUCTION

The algorithm constructs confidence sets, maximizes reward minus priced cost by joint backward induction, and then evaluates the accepted policy's cost lower bound. Sections A.1 and A.2 give these computations in that order. The separate reward recursion in Section A.3 is an auxiliary bound used only in the proof; the algorithm does not run it.

Fix $\delta \in (0,1)$ and set $\varepsilon = \delta/8$. Before each reached episode k, all $k-1$ preceding episodes were valid and their full trajectories are available. Let $N_{k,h}(s,a)$ count visits to (h,s,a) in these trajectories. Form the empirical means $\widehat{r}_{k,h}, \widehat{c}_{k,h}$ and empirical transition rows $\widehat{P}_{k,h}$. The empirical initial law $\widehat{\mu}_k$ uses exactly $k-1$ initial-state observations. At zero count choose any scalar in $[0,1]$ or any probability law, as appropriate. These objects are held fixed during the internal loop preceding episode k.

For $g \in \{r,c\}$, let $V_h^{g,\pi}(s)$ denote the true expected sum of g from stage h through H, starting at s and following π. Its terminal value is zero. The planning value W combines reward and priced cost. A bar denotes the analysis-only reward upper envelope; a tilde denotes the cost lower value used for the queue update.

A.1 CONFIDENCE SETS USED BY THE PLANNER

Hoeffding radii. The scalar radius controls immediate reward. The transition and initial-law radii restrict the plausible models used by the joint recursion. Define

$$\ell_{\mathrm{H}} = \log \frac{32SAHT}{\varepsilon}, \quad \eta_P = S\log 2 + \log \frac{16SAHT}{\varepsilon}, \quad \eta_\mu = S\log 2 + \log \frac{16T}{\varepsilon}. \tag{A.1}$$

For $n \geq 0$, define

$$\begin{aligned} \beta(n) &= \begin{cases} 1, & n = 0, \\ \min\{1, \sqrt{\ell_{\mathrm{H}}/(2n)}\}, & n \geq 1, \end{cases} \\ \beta^P(n) &= \begin{cases} 2, & n = 0, \\ \min\{2, \sqrt{2\eta_P/n}\}, & n \geq 1, \end{cases} \\ \alpha_\mu(n) &= \begin{cases} 2, & n = 0, \\ \min\{2, \sqrt{2\eta_\mu/n}\}, & n \geq 1. \end{cases} \end{aligned} \tag{A.2}$$

Bernstein sets. The scalar interval CI_n and probability set Prob_n control cost uncertainty. They are also used in joint planning, so every planning model is admissible for the selected-policy cost evaluation. Let

$$\ell_{\mathrm{B}} = \log \frac{12HS^2AT}{\varepsilon}. \tag{A.3}$$

For $n \geq 1$, define the inverted scalar interval and categorical set

$$\mathsf{CI}_n(\widehat{z}) = \left\{ q \in [0,1] : |\widehat{z} - q| \leq \sqrt{\frac{2q\ell_{\mathrm{B}}}{n}} + \frac{2\ell_{\mathrm{B}}}{3n} \right\}, \tag{A.4}$$

$$\mathsf{Prob}_n(\widehat{p}) = \{ q \in \Delta(\mathcal{S}) : q(s') \in \mathsf{CI}_n(\widehat{p}(s')) \text{ for all } s' \}. \tag{A.5}$$

At zero count set $\mathsf{CI}_0 = [0,1]$ and $\mathsf{Prob}_0 = \Delta(\mathcal{S})$. All these sets are nonempty and compact: they contain the corresponding empirical scalar or probability law.

For each state-action-stage row, denote the plausible immediate costs by $\mathcal{C}_{k,h}(s,a)$ and the plausible transition laws by $\mathcal{P}_{k,h}(s,a)$:

$$\mathcal{C}_{k,h}(s,a) = \mathsf{CI}_{N_{k,h}(s,a)}(\widehat{c}_{k,h}(s,a)), \qquad \mathcal{P}_{k,h}(s,a) = \mathsf{Prob}_{N_{k,h}(s,a)}(\widehat{P}_{k,h}(\cdot \mid s,a)).$$

Shared planning model. For $n = N_{k,h}(s,a)$, combine these radii and Bernstein sets to define

$$\begin{aligned} r^+_{k,h}(s,a) &= \min\{1, \widehat{r}_{k,h}(s,a) + \beta(n)\}, & c^-_{k,h}(s,a) &= \min \mathcal{C}_{k,h}(s,a), \\ \mathcal{P}^{\mathrm{J}}_{k,h}(s,a) &= \{p \in \mathcal{P}_{k,h}(s,a) : \|p - \widehat{P}_{k,h}(\cdot \mid s,a)\|_1 \leq \beta^P(n)\}, \\ \mathcal{M}^{\mathrm{J}}_k &= \{\mu' \in \mathsf{Prob}_{k-1}(\widehat{\mu}_k) : \|\mu' - \widehat{\mu}_k\|_1 \leq \alpha_\mu(k-1)\}. \end{aligned} \tag{A.6}$$

All sets are compact and contain their empirical centers, even when confidence fails. On the primitive confidence events they contain the true model. The Bernstein restriction connects planning to the cost lower bound; the Hoeffding restriction connects it to the auxiliary reward bound in Section A.3.

A.2 Joint backward induction and the queue-cost estimate

Policy selection. For a fixed price λ, set $W^{\lambda}_{k,H+1} = 0$ and use (4), from $h = H$ down to 1, with the sets in (A.6). At each state choose a maximizing action, with the same fixed tie-breaking rule as Algorithm 1. The maximizing actions give π^{λ}_k, and maximizing the initial expectation gives $J_k(\lambda)$.

This is a single reward-minus-priced-cost recursion. In particular, the continuation term is $p^{\top} W^{\lambda}_{k,h+1}$, not a separate reward upper value. No extra Hoeffding value bonus is added to this recursion. If M^* is a model attaining its row and initial-law maxima, then

$$J_k(\lambda) = R_{M^*}(\pi^{\lambda}_k) - \lambda C_{M^*}(\pi^{\lambda}_k),$$

where R_{M^*}, C_{M^*} are expected episode totals under that same model. Lemma A.1 proves this identity. The values W and J can be negative; they are not clipped to $[0, H]$.

Cost used in the queue update. After stabilization, evaluate the accepted policy with the following fixed-policy recursion. For analysis we define it for arbitrary π, but the algorithm runs it only for the policy it is about to execute. Set $\widetilde{V}^{\pi}_{k,H+1} = 0$ and compute

$$\widetilde{Q}^{\pi}_{k,h}(s,a) = \min_{\substack{q_c \in \mathcal{C}_{k,h}(s,a) \\ q_P \in \mathcal{P}_{k,h}(s,a)}} \{q_c + q_P^{\top} \widetilde{V}^{\pi}_{k,h+1}\},$$

$$\widetilde{V}^{\pi}_{k,h}(s) = \widetilde{Q}^{\pi}_{k,h}(s, \pi_h(s)), \tag{A.7}$$

$$\underline{C}_k(\pi) = \min_{q_\mu \in \mathsf{Prob}_{k-1}(\widehat{\mu}_k)} q_\mu^{\top} \widetilde{V}^{\pi}_{k,1}. \tag{A.8}$$

Each minimum is attained. Rowwise rectangularity means that independently chosen minimizing rows form a single candidate model for this policy; thus (A.8) is exactly the minimum policy cost over the rectangular model set. The values in (A.7) lie in $[0, (H - h + 1)]$, so $\underline{C}_k(\pi) \in [0, H]$. Equation (A.8) supplies $\underline{C}_k(\pi_k)$ in Algorithm 1. It is a separate minimization, not the cost under the model maximizing the joint objective. It uses the Bernstein set $\mathcal{P}_{k,h}$, rather than the smaller planning set $\mathcal{P}^{\mathrm{J}}_{k,h} \subseteq \mathcal{P}_{k,h}$. Thus every planning model is feasible for this cost minimization, as required by the comparison in Lemma A.1.

A.3 Auxiliary reward bound for the proof

The following construction is not the planner's reward computation. It provides a common upper envelope for the reward of any policy under any plausible planning model. This lets us bound the reward error of the policy returned by the joint recursion without changing its selection. Lemma A.1 gives the comparison, and Lemma B.4 bounds the accumulated envelope error.

At stage h, $H - h + 1$ steps remain, including the current step. Define

$$B^{\mathrm{H}}_{k,h}(s,a) = \beta(N_{k,h}(s,a)) + \frac{(H-h)}{2} \beta^P(N_{k,h}(s,a)).$$

For every $\pi \in \Pi$, define $\overline{V}^{r,\pi}_{k,H+1} = 0$ and, only for the analysis, define the backward envelope:

$$\overline{Q}^{r,\pi}_{k,h}(s,a) = \min\{(H-h+1), \widehat{r}_{k,h}(s,a) + \widehat{P}_{k,h} \overline{V}^{r,\pi}_{k,h+1}(s,a) + B^{\mathrm{H}}_{k,h}(s,a)\},$$
$$\overline{V}^{r,\pi}_{k,h}(s) = \overline{Q}^{r,\pi}_{k,h}(s, \pi_h(s)). \tag{A.9}$$

Here $Pf(s,a) = \sum_{s'} P(s' \mid s,a) f(s')$. The corresponding policy upper bound is

$$\overline{R}_k(\pi) = \min\{H, \widehat{\mu}_k \overline{V}^{r,\pi}_{k,1} + (H/2)\alpha_\mu(k-1)\}. \tag{A.10}$$

Its value always lies in $[0, H]$. The clipping here applies only to the auxiliary reward envelope, not to the signed planning values W.

A.4 Link between planning and the error bounds

Lemma A.1 (Joint planning comparison)**.** *On the Hoeffding and Bernstein primitive confidence events, simultaneously for every reached k and every $\lambda \geq 0$,*

$$\max_{\pi \in \Pi} \{R(\pi) - \lambda C(\pi)\} \leq J_k(\lambda) \leq \overline{R}_k(\pi_k^\lambda) - \lambda \underline{C}_k(\pi_k^\lambda). \tag{A.11}$$

The second inequality holds for every history. Also $J_k(0) \geq 0$.

Proof. Consider models formed by independently choosing an initial law in $\mathcal{M}_k^{\mathrm{J}}$ and, at every (h, s, a), a transition row in $\mathcal{P}_{k,h}^{\mathrm{J}}(s, a)$, a reward mean in $[0, r_{k,h}^+(s, a)]$, and a cost mean in $\mathcal{C}_{k,h}(s, a)$. For such a model M, write $R_M(\pi), C_M(\pi)$ for expected episode totals. Rectangularity means that these rowwise choices can be made simultaneously. Backward induction in (4) therefore gives

$$J_k(\lambda) = \max_{\pi, M} \{R_M(\pi) - \lambda C_M(\pi)\} = R_{M^*}(\pi_k^\lambda) - \lambda C_{M^*}(\pi_k^\lambda)$$

for one attaining model M^*. Indeed, at each stage the maximizing immediate means are r^+ and c^-, and the chosen row and action attain the optimal continuation value at every state. Compactness ensures attainment, including the initial-law maximum. True-model feasibility proves the first inequality of (A.11).

For any candidate model M and policy π, backward induction also gives $R_M(\pi) \leq \overline{R}_k(\pi)$. To see this, if its downstream reward value is bounded by $\overline{V}_{k,h+1}^{r,\pi}$, its stage-$h$ value is at most

$$\widehat{r}_{k,h}(s, a) + \beta(n) + \widehat{P}_{k,h} \overline{V}_{k,h+1}^{r,\pi}(s, a) + \frac{H - h}{2} \beta^P(n).$$

The last term follows from the ℓ_1 restriction and the downstream range $[0, H - h]$. Its value is also at most $H - h + 1$, so the clipping in (A.9) preserves domination. The same argument for the initial law, followed by clipping at H, gives (A.10). Likewise, every candidate cost mean, row, and initial law is feasible in the minima (A.7)–(A.8); induction gives $\underline{C}_k(\pi) \leq C_M(\pi)$. Apply both bounds to the same attaining model M^* and use $\lambda \geq 0$. This proves the second inequality without a new concentration argument. At $\lambda = 0$, all candidate rewards are nonnegative, so $J_k(0) \geq 0$. □

B Proof of the stationary upper bound

We prove Theorem 5.1. The finite-sample version, Theorem B.14, keeps the estimation, noise, and queue terms separate.

B.1 Stopping convention and benchmark conversion

Let τ be the first budget-crossing episode, with $\tau = T + 1$ if there is no crossing. Then

$$N_{\text{attempt}} = \min\{\tau, T\}, \qquad N_{\text{valid}} = \min\{\tau - 1, T\}.$$

Thus N_{attempt} counts all executed episodes and N_{valid} counts only those whose rewards are retained. The event that an episode is attempted is known before it starts; validity is known only after observing its cost. We use $\mathcal{F}_{k-1}$ for the pre-episode information, including the selected policy, and pad unattempted episodes with zero increments.

For reference, the deterministic queue bound used to choose U is

$$\overline{Q}_{\text{sq}} = 2HTb + 2H^2[1 + \log(8/\delta)]. \tag{B.1}$$

The notation Q_{sq} below sums the squared update inputs before projection, counting only operations the algorithm performs. The quantity $\overline{Q}_{\text{sq}}$ is its high-probability bound. The choice in (5) is exactly $U = (b/H)\sqrt{\overline{Q}_{\text{sq}}}$.

Lemma B.1 (Benchmark conversion)**.** *For $B^+ = \min\{B + H, HT\}$,*

$$\text{OPT}_{\text{HS}} \le \text{LP}_T(B^+) \le Tv + \frac{Hv}{b}. \tag{B.2}$$

Proof. Attempted cost, including a crossing episode if present, is at most B^+ on every path. Credited reward is at most total attempted reward. Conditional on the past before an attempt, the expected reward and cost of any policy can be represented by a mixture of policies in Π. Indeed, its state-action occupancy can be realized by a randomized Markov policy; pre-sampling the action at each state-stage pair then gives a distribution over Π with the same occupancy. Summing these conditional mixtures over the predictable attempt indicators and taking expectations gives feasible weights for (2) at B^+. This proves the first inequality.

The LP value $f(B')$ is nondecreasing and concave, with $f(0) \ge 0$. For $B' \ge B > 0$, concavity gives $f(B) \ge (B/B')f(B')$. Since $B^+ - B \le H$ and $f(B) = Tv$, the second inequality follows. □

B.2 Concentration tools and adaptive observation counts

We record the probability tools and the precise way they are used. Only in the proofs, write

$$\ell_{\text{noise}} = \log(1/\varepsilon), \qquad \ell_{\text{width}} = \log(8/\varepsilon).$$

These logarithms allocate failure probability to observation noise and summed evaluation widths, respectively; $\varepsilon = \delta/8$ as above. For a mean-zero random variable X with conditional range length d, the bounded-variable moment-generating-function inequality gives $\mathbb{E}[e^{\theta X} \mid \mathcal{F}] \le e^{\theta^2 d^2/8}$. Iteration and Markov's inequality give Hoeffding's bound at deterministic sample count, and Azuma's bound for bounded martingale differences.

For every deterministic variance bound $\sigma^2 \ge 0$ and exponent $\ell_{\text{tail}} > 0$, we use

$$\mathbb{P}\left\{M_n > \sqrt{2\sigma^2\ell_{\text{tail}}} + \frac{2R\ell_{\text{tail}}}{3}, \quad V_n \le \sigma^2\right\} \le e^{-\ell_{\text{tail}}}, \tag{B.3}$$

where $M_n = \sum_{i=1}^n X_i$, $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0$, $|X_i| \le R$, and $V_n = \sum_i \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$. This is the scalar Freedman inequality. Its joint-event form lets us bound variance using a separate high-probability event without conditioning on that event.

Visits are adaptive, and a nominal count may never be reached. The next lemma controls estimation failure jointly with count attainment, so it can be union-bounded over rows and counts without conditioning on the visitation pattern.

Lemma B.2 (Confidence at attained local counts)**.** *Fix a state-action-stage row and an observation in $[0,1]$ with conditional mean $p \in [0,1]$ at every visit. At any deterministic $n \geq 1$, the event that the nth visit is attained and its empirical mean violates either radius has the following probabilities:*

$$\mathbb{P}\left\{n \text{ attained},\ |\widehat{p}_n - p| > \sqrt{\ell_{\mathrm{tail}}/(2n)}\right\} \leq 2e^{-\ell_{\mathrm{tail}}}, \tag{B.4}$$

$$\mathbb{P}\left\{n \text{ attained},\ |\widehat{p}_n - p| > \sqrt{2p\ell_{\mathrm{tail}}/n} + 2\ell_{\mathrm{tail}}/(3n)\right\} \leq 2e^{-\ell_{\mathrm{tail}}}. \tag{B.5}$$

The same statements hold for initial-state observations and for each indicator coordinate of a categorical observation.

Proof. Expose each joint reward, cost, and next-state observation only after the state and action have been selected. The indicator of visiting the fixed row is then predictable. Stop accepting its observations after the first n visits and pad all remaining slots through the deterministic global horizon HT with zeros. This defines a martingale with at most n nonzero increments, each having conditional range length one. Its Hoeffding exponential is compensated by the number of accepted visits, which is at most n. On the event that count n is attained, its final sum is $n(\widehat{p}_n - p)$. The exponential bound therefore proves (B.4) without conditioning on whether that count exists.

For (B.5), each centered accepted observation has absolute value at most one and conditional variance at most p, since $X^2 \leq X$ for $X \in [0,1]$. The stopped quadratic variation is at most np. Apply (B.3) to each sign and divide by n. The initial-state case uses the deterministic episode slots instead. This construction also shows why observing contemporaneous cost or reward noise before defining the transition filtration would be inappropriate: we use the pre-joint-observation filtration throughout. □

The next inequality converts an observed nonnegative sum into a bound on its conditional expectation. We use it for reciprocal visit counts and for cumulative cost.

Lemma B.3 (A nonnegative-variable exponential bound)**.** *If $Y_i \in [0,1]$ is adapted and $p_i = \mathbb{E}[Y_i \mid \mathcal{F}_{i-1}]$, then, for $a = 1 - e^{-1}$,*

$$\mathbb{P}\left\{a\sum_{i=1}^{n} p_i > \sum_{i=1}^{n} Y_i + \ell_{\mathrm{tail}}\right\} \leq e^{-\ell_{\mathrm{tail}}}. \tag{B.6}$$

Proof. Convexity gives $e^{-y} \leq 1 - ay$ for $y \in [0,1]$. Hence $\mathbb{E}[e^{ap_i - Y_i} \mid \mathcal{F}_{i-1}] \leq e^{ap_i}(1 - ap_i) \leq 1$. Multiply these conditional inequalities and use Markov's inequality. □

B.3 Uniform reward confidence and its cumulative error

The next lemma concerns the auxiliary envelope (A.9), not the signed planning recursion (4). It proves envelope optimism by backward induction, then sums its Bellman widths using a trajectory martingale and visit counts. The conclusion applies to policies selected by the joint planner even though it never computes this envelope.

The resulting reward-error bound is denoted by E_{H}:

$$\begin{aligned} E_{\mathrm{H}} = {} & 2H\sqrt{2\eta_\mu T} + [2H\sqrt{2\ell_{\mathrm{H}}} + H(H-1)\sqrt{2\eta_P}]\sqrt{SAT} \\ & + H(H+1)\sqrt{2T\ell_{\mathrm{width}}}. \end{aligned} \tag{B.7}$$

Its three terms come from estimating the initial law, the state-action rows, and replacing expected trajectory widths by observed widths.

Lemma B.4 (Hoeffding reward confidence and error envelope)**.** *Fix any policy-selection rule that uses only pre-episode history. Under the resulting interaction, except on an event of probability at most ε, simultaneously for every reached k and every $\pi \in \Pi$,*

$$\overline{R}_k(\pi) \geq R(\pi).$$

Along the selected sequence, the inclusive attempted-prefix error satisfies

$$\sum_{k=1}^{N_{\mathrm{attempt}}} (\overline{R}_k(\pi_k) - R(\pi_k)) \leq E_{\mathrm{H}}. \tag{B.8}$$

Proof. Apply (B.4) to the reward means, and union bound over SAH rows and $n \le T$. This costs at most $2SAHTe^{-\ell_{\mathrm{H}}} = \varepsilon/16$. For each $\sigma \in \{-1,1\}^S$, the categorical variable $\sigma(s')$ has range length two, so its one-sided Hoeffding tail at $\sqrt{2\eta_P/n}$ is at most $e^{-\eta_P}$. Since $\|\widehat{P}-P\|_1 = \max_\sigma \sigma^\top(\widehat{P}-P)$, the row/count/sign union costs at most $2^S SAHTe^{-\eta_P} = \varepsilon/16$. The initial-law counterpart costs $2^S Te^{-\eta_\mu} = \varepsilon/16$. Thus, except with probability at most $\varepsilon/4$, all primitive reward and ℓ_1 errors are bounded by (A.2). Clipping radii and zero-count radii are valid because their corresponding maximum errors are one and two.

On this event, for *every* $f : \mathcal{S} \to [0, D_f]$,

$$|(\widehat{P} - P)f| \le \frac{D_f}{2}\|\widehat{P} - P\|_1. \tag{B.9}$$

This is a deterministic inequality, so f may depend on all historical data. Applying it to the auxiliary downstream values $\overline{V}^{r,\pi}_{k,h+1}$ in (A.9) and using backward induction proves the displayed envelope optimism.

More precisely, set

$$\omega_{k,h}(s,a) = 2\beta(N_{k,h}(s,a)) + (H-h)\beta^P(N_{k,h}(s,a)).$$

For reward, adding and subtracting $P_h\overline{V}^{r,\pi}_{k,h+1}$ shows that the nonnegative stagewise gap $d_h = \overline{V}^{r,\pi}_{k,h} - V^{r,\pi}_h$ satisfies

$$d_h(s) \le \omega_{k,h}(s,\pi_h(s)) + P_h d_{h+1}(s,\pi_h(s)).$$

Clipping at $(H-h+1)$ only reduces the gap. The initial-law correction and initial empirical error add at most $H\alpha_\mu(k-1)$. Consequently each policywise reward error is bounded by

$$H\alpha_\mu(k-1) + \mathbb{E}_{\mu,P,\pi}\sum_{h=1}^{H}\omega_{k,h}(s_h,a_h). \tag{B.10}$$

On each attempt let $W_k = \sum_h \omega_{k,h}(s_{k,h},a_{k,h})$. It lies in $[0, H(H+1)]$. Pad unattempted episodes with zeros and use the predictable attempt indicators. Azuma gives, except with probability $\varepsilon/8$,

$$\sum_{k=1}^{N_{\mathrm{attempt}}}\mathbb{E}[W_k \mid \mathcal{F}_{k-1}] \le \sum_{k=1}^{N_{\mathrm{attempt}}} W_k + H(H+1)\sqrt{2T\ell_{\mathrm{width}}}.$$

The pre-episode filtration includes the chosen policy; conditional on it, the episode follows the true model and that policy.

As $\ell_{\mathrm{H}}, \eta_P, \eta_\mu \ge 2$, including at zero counts,

$$\omega_{k,h}(s,a) \le \frac{\sqrt{2\ell_{\mathrm{H}}} + (H-h)\sqrt{2\eta_P}}{\sqrt{N_{k,h}(s,a)\vee 1}}, \qquad H\alpha_\mu(k-1) \le \frac{H\sqrt{2\eta_\mu}}{\sqrt{(k-1)\vee 1}}.$$

The elementary count bound

$$\sum_{j=1}^{n}\frac{1}{\sqrt{(j-1)\vee 1}} \le 2\sqrt{n} \qquad (n \ge 1) \tag{B.11}$$

follows by integration of $t^{-1/2}$ (and directly for $n = 1$). For the final attempted count $n_{h,s,a}$, $\sum_{s,a} n_{h,s,a} = N_{\mathrm{attempt}} \le T$, so $\sum_{s,a}\sqrt{n_{h,s,a}} \le \sqrt{SAT}$. Summing (B.11) across rows and stages, and separately for the initial counts, bounds the two nonmartingale terms by

$$[2H\sqrt{2\ell_{\mathrm{H}}} + H(H-1)\sqrt{2\eta_P}]\sqrt{SAT}, \qquad 2H\sqrt{2\eta_\mu T}.$$

Combine with (B.10). The total failure probability is at most $\varepsilon/4 + \varepsilon/8 \le \varepsilon$. No policy union bound or pointwise concentration of a selected value is used. □

B.4 Bernstein policy evaluation

Let $C_{\mathrm{tot}} = \sum_{k=1}^{N_{\mathrm{attempt}}} C(\pi_k)$ be cumulative conditional expected cost over all attempted episodes. We seek a cost-error bound proportional to $\sqrt{C_{\mathrm{tot}}}$. The proof needs true-model feasibility, a confidence-set diameter valid for all downstream values, and control of reciprocal visit counts.

B.4.1 PRIMITIVE CONFIDENCE AND A UNIFORM DIAMETER

We first place the true model in every confidence set by applying Bernstein concentration to cost observations and next-state indicators. This ensures that model minimization gives a lower value for every policy.

Lemma B.5 (Bernstein primitive event)**.** *With probability at least $1 - \varepsilon/2$, the true cost means, transition rows, and initial law belong to all their sets* (A.4)–(A.5) *at every attained count.*

Proof. Use (B.5) for costs and for each transition and initial-state indicator. Union bounding over $n = 1, \dots, T$ and at most $HSA + HS^2A + S \le 3HS^2A$ coordinates costs at most

$$2T(HSA + HS^2A + S)e^{-\ell_{\mathrm{B}}} \le 6HS^2ATe^{-\ell_{\mathrm{B}}} = \varepsilon/2.$$

At zero count the sets contain every feasible scalar or law. □

Feasibility gives the sign of the error; the next lemma gives its size. Centering a downstream vector at its mean exposes its variance. The diameter bound is deterministic, so the vector may depend on the data.

Lemma B.6 (Diameter valid for every downstream vector)**.** *Suppose $p, q \in \mathsf{CI}_n(\widehat{z})$ and let $\bar{n} = n \vee 1$. Then*

$$|p - q| \le 4\sqrt{\frac{2p\ell_{\mathrm{B}}}{\bar{n}}} + \frac{14\ell_{\mathrm{B}}}{3\bar{n}}. \tag{B.12}$$

If probability vectors p, q belong to the same categorical set, then simultaneously for every $f \in \mathbb{R}^S$, with $\mathrm{Var}_p(f) = \sum_i p_i(f_i - p^\top f)^2$,

$$|(p - q)^\top f| \le 4\sqrt{\frac{2S\ell_{\mathrm{B}}\,\mathrm{Var}_p(f)}{\bar{n}}} + \frac{14S\ell_{\mathrm{B}}\,\mathrm{span}(f)}{3\bar{n}}, \qquad \mathrm{span}(f) = \max_i f_i - \min_i f_i. \tag{B.13}$$

Proof. For $n \ge 1$, write $d = |p - q|$ and $a_n = \ell_{\mathrm{B}}/n$. The two defining interval inequalities and $\sqrt{q} \le \sqrt{p} + \sqrt{d}$ give, in either ordering of p and q,

$$d \le 2\sqrt{2a_np} + \sqrt{2a_nd} + \frac{4a_n}{3} \le 2\sqrt{2a_np} + \frac{d}{2} + \frac{7a_n}{3}.$$

Rearranging proves (B.12) with the designated reference coordinate p. At $n = 0$ it follows from $d \le 1$ and $\ell_{\mathrm{B}} > 1$.

For categorical laws, use $\sum_i (p_i - q_i) = 0$ to center f at $p^\top f$. Sum (B.12) coordinatewise and apply

$$\sum_i \sqrt{p_i}|f_i - p^\top f| \le \sqrt{S\,\mathrm{Var}_p(f)}, \qquad \sum_i |f_i - p^\top f| \le S\,\mathrm{span}(f).$$

These are deterministic inequalities, which establishes simultaneity over all vectors f, including data-dependent ones. □

B.4.2 RECIPROCAL COUNTS ALONG THE SELECTED TRAJECTORIES

The diameter bounds contain reciprocal sample counts. The next lemma controls their conditional-occupancy-weighted sums by transferring pathwise harmonic-count bounds through Lemma B.3.

For each reached episode, write

$$\rho_{k,h}(s,a) = \Pr_{\mu,P,\pi_k}(s_h = s, a_h = a \mid \mathcal{F}_{k-1}),$$

where $\mathcal{F}_{k-1}$ denotes its pre-episode history. In all sums below, quantities for unattempted episodes are zero-padded. Define

$$\Xi = \sum_{k=1}^{N_{\mathrm{attempt}}} \left[\frac{1}{(k-1) \vee 1} + \sum_{h,s,a} \frac{\rho_{k,h}(s,a)}{N_{k,h}(s,a) \vee 1} \right]. \tag{B.14}$$

One deterministic bound covers both the initial and transition counts:

$$\Lambda = \frac{HSA(2 + \log T) + H\log(2/\varepsilon)}{1 - e^{-1}} + 2 + \log T. \tag{B.15}$$

Lemma B.7 (Reciprocal-visitation control)**.** *With probability at least* $1-\varepsilon/2$*,* $\Xi \le \Lambda$*.*

Proof. Apply Lemma B.3, with exponent $\log(2/\varepsilon)$, to

$$Y_k = \frac{\mathbf{1}\{k \le N_{\mathrm{attempt}}\}}{H} \sum_{h=1}^{H} \frac{1}{N_{k,h}(s_{k,h}, a_{k,h}) \vee 1} \in [0,1],$$

setting $Y_k = 0$ when there is no attempt. Its conditional mean is the transition-count summand in (B.14) divided by H. A row visited $n \ge 1$ times contributes

$$\sum_{j=1}^{n} \frac{1}{(j-1) \vee 1} = 1 + \sum_{j=1}^{n-1} \frac{1}{j} \le 2 + \log T;$$

an unvisited row contributes zero. Summing over HSA rows and multiplying the exponential bound by H gives

$$\sum_{k=1}^{N_{\mathrm{attempt}}} \sum_{h,s,a} \frac{\rho_{k,h}(s,a)}{N_{k,h}(s,a) \vee 1} \le \frac{HSA(2+\log T) + H\log(2/\varepsilon)}{1-e^{-1}}.$$

The initial count runs through $0, \dots, N_{\mathrm{attempt}} - 1$, so its contribution is at most $2 + \log T$ pathwise. Adding the two bounds proves the claim. Internal updates do not change any count. □

B.4.3 Policywise pessimism and an exact gap recursion

We now sum the nonnegative local Bellman gaps. Their diameters involve the robust lower values, but total variance is naturally bounded for the true cost values. Comparing these two variances charges the difference to the same estimation error being bounded. Absorbing that error's square-root term gives the following result.

The next bound uses two coefficients: α_{B} multiplies the square root of cumulative cost, and β_{B} is the additive remainder. They are

$$\alpha_{\mathrm{B}} = 28\sqrt{SH\ell_{\mathrm{B}}\Lambda}, \tag{B.16}$$

$$\beta_{\mathrm{B}} = 84H^2 S\ell_{\mathrm{B}}\Lambda. \tag{B.17}$$

Theorem B.8 (First-order error of Bernstein cost evaluation)**.** *Fix any policy-selection rule that uses only pre-episode history. Under the resulting interaction, with probability at least* $1-\varepsilon$*, simultaneously for every reached episode and every policy,* $0 \le \underline{C}_k(\pi) \le C(\pi)$*. On the same event, the selected sequence satisfies*

$$\Gamma_{\mathrm{B}}^{+} := \sum_{k=1}^{N_{\mathrm{attempt}}} (C(\pi_k) - \underline{C}_k(\pi_k)) \le \alpha_{\mathrm{B}}\sqrt{C_{\mathrm{tot}}} + \beta_{\mathrm{B}}, \qquad C_{\mathrm{tot}} = \sum_{k=1}^{N_{\mathrm{attempt}}} C(\pi_k). \tag{B.18}$$

Proof. Intersect the events of Lemmas B.5 and B.7. On the primitive event, the true model is feasible in every rectangular set. Backward induction therefore gives $0 \le \widetilde{V}_{k,h}^{\pi} \le V_h^{c,\pi}$ pointwise, and minimizing over the initial-law set gives uniform policywise pessimism.

Fix one pre-episode history and its selected ordinary policy π_k. For this proof abbreviate its true cost value by V_h, its Bernstein lower value by $\widetilde{V}_h$, and put $\Delta_h = V_h - \widetilde{V}_h$. Let $\widetilde{Q}_h$ be its robust Bellman value. Define the nonnegative local and initial gaps

$$e_h(s,a) = c_h(s,a) + P_h\widetilde{V}_{h+1}(s,a) - \widetilde{Q}_h(s,a),$$
$$e_0 = \mu^{\top}\widetilde{V}_1 - \min_{q_\mu \in \mathsf{Prob}_{k-1}(\widehat{\mu}_k)} q_\mu^{\top}\widetilde{V}_1.$$

Their nonnegativity follows because the true row or initial law is feasible. The following gap recursion is exact:

$$\Delta_h(s) = e_h(s, \pi_{k,h}(s)) + P_h\Delta_{h+1}(s, \pi_{k,h}(s)). \tag{B.19}$$

Writing $e_k = C(\pi_k) - \underline{C}_k(\pi_k)$ and $\zeta_{k,h} = \mathbb{E}_{\mu,P,\pi_k}\Delta_h(s_h)$, with $\zeta_{k,H+1} = 0$, it follows that

$$e_k = e_0 + \sum_{h=1}^{H} \mathbb{E}e_h(s_h, a_h), \qquad \zeta_{k,h} = \mathbb{E}e_h + \zeta_{k,h+1}, \qquad \sum_{h=1}^{H} \zeta_{k,h} \le He_k. \tag{B.20}$$

In the final inequality, $\zeta_{k,1} \ge \cdots \ge \zeta_{k,H} \ge 0$ and $\zeta_{k,1} \le e_k$.

Total variance under the true model. The bound $\sum_h \zeta_{k,h} \le H e_k$ ensures that the discrepancies between true and lower values across all stages can be charged to the episode's total error. We next use this fact to transfer a variance bound for the true values to the lower values appearing in the diameter.

Conditional on this pre-episode history, define the mean-cost path functional

$$F_k = \sum_{h=1}^{H} c_h(s_{k,h}, \pi_{k,h}(s_{k,h})) \in [0, H].$$

This is not the noisy observed cost G_k^c. Bellman telescoping gives

$$F_k - C(\pi_k) = V_1(s_{k,1}) - \mu^\top V_1 + \sum_{h=1}^{H-1} [V_{h+1}(s_{k,h+1}) - P_h V_{h+1}(s_{k,h}, a_{k,h})]. \tag{B.21}$$

These terms are orthogonal martingale differences in the state-path filtration, because π_k is fixed conditional on the pre-episode history. Thus

$$\mathrm{Var}_\mu(V_1) + \sum_{h=1}^{H-1} \mathbb{E}\,\mathrm{Var}_{P_h}(V_{h+1}) = \mathrm{Var}(F_k) \le \mathbb{E}F_k^2 \le HC(\pi_k). \tag{B.22}$$

Within-step reward/cost noise correlations do not enter this identity.

For any one of the initial or transition laws, using $0 \le \Delta \le H$,

$$\mathrm{Var}(\widetilde{V}) \le \mathbb{E}[(\widetilde{V} - \mathbb{E}V)^2] \le 2\,\mathrm{Var}(V) + 2\mathbb{E}\Delta^2 \le 2\,\mathrm{Var}(V) + 2H\mathbb{E}\Delta. \tag{B.23}$$

Return now to the episode indices and define

$$\Omega_{\mathrm{val}} = \sum_{k=1}^{N_{\mathrm{attempt}}} \left[\mathrm{Var}_\mu(\widetilde{V}_{k,1}^{\pi_k}) + \sum_{h=1}^{H-1} \mathbb{E}_{\mu,P,\pi_k}\,\mathrm{Var}_{P_h}(\widetilde{V}_{k,h+1}^{\pi_k}) \right].$$

The initial and downstream Δ terms in (B.23) sum to $\sum_h \zeta_{k,h}$ for each episode. Combining (B.20)–(B.23) therefore yields

$$\Omega_{\mathrm{val}} \le 2HC_{\mathrm{tot}} + 2H^2\Gamma_{\mathrm{B}}^+. \tag{B.24}$$

All of these are deterministic statements about the conditional true model at each history. They do not apply a new concentration inequality after conditioning on the confidence event.

Sum the confidence diameters. Equation (B.24) expresses the total downstream variance in terms of cumulative cost and cumulative estimation error. We now sum the local diameters using the reciprocal potentials. The resulting square root of the estimation error will be absorbed algebraically.

For a minimizing scalar and row in (A.7),

$$e_h = (c_h - \widetilde{c}_h) + (P_h - \widetilde{P}_h)^\top \widetilde{V}_{h+1}.$$

Bound each term above by its absolute value and apply Lemma B.6; apply its categorical bound also to e_0. Weighted Cauchy–Schwarz and (B.14) give

$$\sum_{k=1}^{N_{\mathrm{attempt}}} \sum_{h,s,a} \rho_{k,h}(s,a) \sqrt{\frac{c_h(s,a)}{N_{k,h}(s,a) \vee 1}} \le \sqrt{C_{\mathrm{tot}}\Lambda}.$$

For the categorical terms, the explicit initial-plus-transition sum obeys

$$\begin{aligned} &\sum_{k=1}^{N_{\mathrm{attempt}}} \sqrt{\frac{\mathrm{Var}_\mu(\widetilde{V}_{k,1}^{\pi_k})}{(k-1) \vee 1}} \\ &\quad + \sum_{k=1}^{N_{\mathrm{attempt}}} \sum_{h=1}^{H-1} \sum_{s,a} \rho_{k,h}(s,a) \sqrt{\frac{\mathrm{Var}_{P_h(\cdot|s,a)}(\widetilde{V}_{k,h+1}^{\pi_k})}{N_{k,h}(s,a) \vee 1}} \\ &\le \sqrt{\Omega_{\mathrm{val}}\Xi} \le \sqrt{\Omega_{\mathrm{val}}\Lambda}. \end{aligned}$$

The first inequality may include the terminal-stage reciprocal counts on its right-hand side; this only enlarges the bound. Since every relevant span is at most H, the corresponding linear terms sum to at most $(14\ell_{\mathrm{B}}/3)(1+SH)\Lambda$. Using (B.20), we obtain

$$\Gamma_{\mathrm{B}}^{+} \le 4\sqrt{2\ell_{\mathrm{B}}\Lambda C_{\mathrm{tot}}} + 4\sqrt{2S\ell_{\mathrm{B}}\Lambda\Omega_{\mathrm{val}}} + \frac{14\ell_{\mathrm{B}}}{3}(1+SH)\Lambda. \tag{B.25}$$

Substitute (B.24) and use $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$:

$$\Gamma_{\mathrm{B}}^{+} \le (4\sqrt{2} + 8\sqrt{SH})\sqrt{\ell_{\mathrm{B}}\Lambda C_{\mathrm{tot}}} + 8H\sqrt{S\ell_{\mathrm{B}}\Lambda\Gamma_{\mathrm{B}}^{+}} + \frac{14\ell_{\mathrm{B}}}{3}(1+SH)\Lambda.$$

The middle term is at most $\Gamma_{\mathrm{B}}^{+}/2 + 32H^2S\ell_{\mathrm{B}}\Lambda$ by $2xy \le x^2 + y^2$. Move $\Gamma_{\mathrm{B}}^{+}/2$ to the left and multiply by two. Since $H, S \ge 1$,

$$8\sqrt{2} + 16\sqrt{SH} \le 28\sqrt{SH}, \qquad 64H^2S + \frac{28}{3}(1+SH) \le 84H^2S.$$

This proves (B.18) with (B.16)–(B.17). The two events cost at most $\varepsilon/2$ each. □

B.5 Hard-budget control of conditional cost and realized noise

The confidence analysis uses expected costs, while the stopping rule bounds observed costs. The next lemma connects them; its bound also controls cost noise and squared queue increments.

To distinguish an actual quantity from its bound, write

$$\overline{C}_{\mathrm{tot}} = \frac{Tb + H(1+\ell_{\mathrm{noise}})}{1-e^{-1}}. \tag{B.26}$$

Thus C_{tot} is random, whereas $\overline{C}_{\mathrm{tot}}$ depends only on the prescribed problem parameters and confidence level.

Lemma B.9 (Cumulative expected cost)**.** *With failure probability at most ε,*

$$C_{\mathrm{tot}} = \sum_{k=1}^{N_{\mathrm{attempt}}} C(\pi_k) \le \overline{C}_{\mathrm{tot}}. \tag{B.27}$$

Proof. Use $a = 1 - e^{-1}$ as in Lemma B.3, and apply that lemma with

$$Y_k = \mathbf{1}\{k \le N_{\mathrm{attempt}}\}\frac{G_k^c}{H}, \qquad p_k = \mathbf{1}\{k \le N_{\mathrm{attempt}}\}\frac{C(\pi_k)}{H}.$$

Use the padded horizon T and exponent ℓ_{noise}. Since total attempted cost is at most $Tb + H$, the lemma gives

$$aC_{\mathrm{tot}} \le \sum_{k=1}^{N_{\mathrm{attempt}}} G_k^c + H\ell_{\mathrm{noise}} \le Tb + H(1+\ell_{\mathrm{noise}}).$$

Dividing by a proves (B.27). □

Combining Bernstein cost evaluation with the budget bound gives the cost-error bound

$$E_C = \alpha_{\mathrm{B}}\sqrt{\overline{C}_{\mathrm{tot}}} + \beta_{\mathrm{B}}. \tag{B.28}$$

We next pass from expected episode values to observed totals. The martingales include the crossing attempt; removing its one increment afterward costs at most H.

Write $\overline{M}_R$ and $\overline{M}_C$ for the following nonnegative bounds on accumulated reward noise and cost noise:

$$\overline{M}_R = H + H\sqrt{2T\ell_{\mathrm{noise}}},$$
$$\overline{M}_C = H + \sqrt{2H\overline{C}_{\mathrm{tot}}\ell_{\mathrm{noise}}} + \frac{2H\ell_{\mathrm{noise}}}{3}. \tag{B.29}$$

Lemma B.10 (Realized reward and cost deviations)**.** *Outside additional events with total failure probability at most* 2ε*, on the budget event* (B.27)*,*

$$\sum_{k=1}^{N_{\mathrm{valid}}} (R(\pi_k) - G_k^r) \le \overline{M}_R, \qquad \sum_{k=1}^{N_{\mathrm{valid}}} (G_k^c - C(\pi_k)) \le \overline{M}_C. \tag{B.30}$$

Proof. Pad the inclusive attempted-prefix reward and cost differences by zeros. They are martingale differences of absolute value at most H, because the attempt indicator and policy are predictable. One-sided Azuma at failure level ε bounds the inclusive reward sum by $H\sqrt{2T\ell_{\mathrm{noise}}}$.

For the inclusive cost sum, let V_C^+ be its predictable quadratic variation. Since $0 \le G_k^c \le H$,

$$V_C^+ \le \sum_{k=1}^{N_{\mathrm{attempt}}} \mathbb{E}[(G_k^c)^2 \mid \mathcal{F}_{k-1}] \le H \sum_{k=1}^{N_{\mathrm{attempt}}} C(\pi_k) = HC_{\mathrm{tot}}.$$

By (B.3), an event of failure at most ε is

$$\left\{ \sum_{k=1}^{N_{\mathrm{attempt}}} (G_k^c - C(\pi_k)) > \sqrt{2H\overline{C}_{\mathrm{tot}}\ell_{\mathrm{noise}}} + \frac{2H\ell_{\mathrm{noise}}}{3}, \quad V_C^+ \le H\overline{C}_{\mathrm{tot}} \right\}. \tag{B.31}$$

On the budget event, $V_C^+ \le H\overline{C}_{\mathrm{tot}}$. Thus outside (B.31) the Bernstein–Freedman threshold bounds this sum. We have used a joint-event bound, not conditioned the martingale inequality on a future budget event.

If there is a crossing, removing its one increment increases either noise sum by at most H, proving (B.30). The valid-episode indicator is never treated as predictable. □

B.6 Comparator, queue, and stopping time

We now connect the estimation bounds to regret. On the confidence event, this part is a deterministic calculation with the policy scores, queue updates, and stopping rule.

Work on the uniform confidence events of Lemma B.4 and Theorem B.8. Optimal LP weights give a subprobability vector $q_\pi = \omega_\pi^*/T$ with

$$q_\pi \ge 0, \quad \sum_\pi q_\pi \le 1, \quad \sum_\pi q_\pi R(\pi) = v, \quad \sum_\pi q_\pi C(\pi) \le b. \tag{B.32}$$

B.6.1 Selection and internal-decrement certificates

The LP mixture has mass at most one, so the score's sign matters. A nonnegative maximum gives the policy comparison; a negative maximum certifies a multiplier large enough to cancel an internal decrement.

Lemma B.11 (The two signs of the maximization test)**.** *Each accepted physical policy satisfies*

$$v - R(\pi_k) \le \overline{R}_k(\pi_k) - R(\pi_k) - \lambda_k(\underline{C}_k(\pi_k) - b). \tag{B.33}$$

Every internal decrement, at its pre-decrement multiplier λ_j^{st}*, satisfies*

$$\lambda_j^{\mathrm{st}} \ge v/b. \tag{B.34}$$

Proof. For any $\lambda \ge 0$, (B.32) and Lemma A.1 imply, with $m = \sum_\pi q_\pi \le 1$,

$$v - \lambda b \le \sum_\pi q_\pi (R(\pi) - \lambda C(\pi)) \le m J_k(\lambda).$$

At acceptance, $J_k(\lambda_k) \ge 0$, so $mJ_k(\lambda_k) \le J_k(\lambda_k) \le \overline{R}_k(\pi_k) - \lambda_k \underline{C}_k(\pi_k)$. This gives (B.33). Before an internal decrement, $J_k(\lambda) < 0$, hence $v - \lambda b \le mJ_k(\lambda) \le 0$. This gives (B.34) without introducing a null response. □

B.6.2 Actual updates only and the terminal negative square

Let N_{int} be the number of internal decrements actually performed, including those preceding the crossing attempt if one occurs. For $k \le N_{\mathrm{valid}}$ put $g_k = \underline{C}_k(\pi_k) - b$, and define

$$Q_{\mathrm{sq}} = \sum_{k=1}^{N_{\mathrm{valid}}} g_k^2 + N_{\mathrm{int}} b^2. \tag{B.35}$$

Let Z_{end} be the actual final queue: the stabilized queue just before the crossing execution if there is a crossing, or the queue after the last valid physical update otherwise. No queue update is added for a crossing episode, including for purposes of analysis.

We telescope the actual projected updates and retain the terminal negative square: it will absorb the positive terminal-queue term created by the stopping calculation.

Lemma B.12 (Exact potential inequality)**.** *The actual chronological updates satisfy*

$$-\sum_{k=1}^{N_{\mathrm{valid}}} \lambda_k g_k + b \sum_{j=1}^{N_{\mathrm{int}}} \lambda_j^{\mathrm{st}} \le -\frac{Z_{\mathrm{end}}^2}{2U} + \frac{Q_{\mathrm{sq}}}{2U}. \tag{B.36}$$

Proof. For $Z \ge 0$ and an increment y, $[Z+y]_+^2 \le (Z+y)^2$, and therefore

$$2Z(-y) \le Z^2 - [Z+y]_+^2 + y^2.$$

Apply this to each actual valid physical update $y = g_k$ and each actual internal update $y = -b$, divide by $2U$, and telescope from the initial queue zero. □

B.6.3 Budget-sensitive sum of squared increments

To control squared increments, use $\underline{C}_k(\pi_k)^2 \le H\underline{C}_k(\pi_k)$ and bound internal decreases by preceding positive queue increases. This leaves $(H-b)\sum_{k=1}^{N_{\mathrm{valid}}} \underline{C}_k(\pi_k) + 2Tb^2$, which the budget event controls.

Lemma B.13 (Control of the OGD quadratic term)**.** *On the confidence and budget events,* $Q_{\mathrm{sq}} \le \overline{Q}_{\mathrm{sq}}$.

Proof. Write $S_x = \sum_{k=1}^{N_{\mathrm{valid}}} \underline{C}_k(\pi_k)$. An internal decrement is full if its queue decrease equals b, and partial otherwise. Each internal block has at most one partial decrement: it reaches zero, where the maximization test is nonnegative. There are at most N_{attempt} blocks. Total internal queue decrease cannot exceed the total positive physical increase from the zero initial queue. Hence

$$N_{\mathrm{int}} b^2 \le b \sum_{k=1}^{N_{\mathrm{valid}}} (\underline{C}_k(\pi_k) - b)_+ + N_{\mathrm{attempt}} b^2 \le bS_x + N_{\mathrm{attempt}} b^2. \tag{B.37}$$

For $0 \le \underline{C}_k(\pi_k) \le H$, $\underline{C}_k(\pi_k)^2 \le H\underline{C}_k(\pi_k)$ gives $(\underline{C}_k(\pi_k)-b)^2 \le (H-2b)\underline{C}_k(\pi_k) + b^2$. Adding this to (B.37) yields

$$Q_{\mathrm{sq}} \le (H-b)S_x + (N_{\mathrm{valid}} + N_{\mathrm{attempt}})b^2 \le (H-b)S_x + 2Tb^2. \tag{B.38}$$

On confidence, $S_x \le C_{\mathrm{tot}} \le \overline{C}_{\mathrm{tot}}$. Since $H - b \ge 0$,

$$\begin{aligned} Q_{\mathrm{sq}} &\le a^{-1}(H-b)(Tb + H(1+\ell_{\mathrm{noise}})) + 2Tb^2 \\ &= Tb[a^{-1}H + (2-a^{-1})b] + a^{-1}H(H-b)(1+\ell_{\mathrm{noise}}) \\ &\le 2HTb + 2H^2(1+\ell_{\mathrm{noise}}), \end{aligned}$$

using $a = 1 - e^{-1}$, $1 < a^{-1} < 2$, and $b \le H$. This is exactly $\overline{Q}_{\mathrm{sq}}$. □

B.7 PROOF OF THE UPPER BOUND

All ingredients now have a concrete meaning: E_{H} and E_C bound evaluation errors; $\overline{M}_R$ and $\overline{M}_C$ bound observation noise; and $\overline{Q}_{\mathrm{sq}}$ bounds squared queue increments. The following finite-sample certificate keeps these contributions separate. Its proof is the final assembly step.

Theorem B.14 (Explicit realized regret)**.** *Under the setting in Section 3, for each prescribed $U > 0$, Algorithm 1 satisfies, with probability at least $1 - \delta$,*

$$\mathrm{Reg}_{\mathrm{LP}} \le E_{\mathrm{H}} + \overline{M}_R + \frac{v}{b}(E_C + \overline{M}_C) + \frac{\overline{Q}_{\mathrm{sq}}}{2U} + \frac{Uv^2}{2b^2} + \frac{Hv}{b}, \tag{B.39}$$

$$\mathrm{Reg}_{\mathrm{HS}} \le E_{\mathrm{H}} + \overline{M}_R + \frac{v}{b}(E_C + \overline{M}_C) + \frac{\overline{Q}_{\mathrm{sq}}}{2U} + \frac{Uv^2}{2b^2} + \frac{2Hv}{b}. \tag{B.40}$$

Moreover, $\mathrm{Reg}_{\mathrm{LP}} \le Tv$ *deterministically.*

Proof of Theorem B.14. Intersect the Hoeffding event of Lemma B.4, the Bernstein event of Theorem B.8, the budget event of Lemma B.9, and the two noise events in Lemma B.10. Their total failure probability is at most $5\varepsilon = 5\delta/8 < \delta$. Lemma A.1 holds on these same events. The algorithm chooses each policy from pre-episode information, so both evaluation-error bounds apply. Their summands are nonnegative; dropping a crossing attempt and adding the noise bounds therefore gives

$$\begin{aligned}
\sum_{k=1}^{N_{\mathrm{valid}}} (\overline{R}_k(\pi_k) - G_k^r) &\le E_{\mathrm{H}} + \overline{M}_R, \\
\sum_{k=1}^{N_{\mathrm{valid}}} (G_k^c - \underline{C}_k(\pi_k)) &\le E_C + \overline{M}_C.
\end{aligned} \tag{B.41}$$

The reward lost after stopping. Since $[Z + y]_+ \ge Z + y$, the actual queue increments satisfy

$$Z_{\mathrm{end}} \ge \sum_{k=1}^{N_{\mathrm{valid}}} \underline{C}_k(\pi_k) - bN_{\mathrm{valid}} - bN_{\mathrm{int}}. \tag{B.42}$$

On crossing, the retained episodes have total cost greater than $Tb - H$. Combining this fact with (B.41) and (B.42) gives

$$b(T - N_{\mathrm{valid}}) \le H + E_C + \overline{M}_C + Z_{\mathrm{end}} + bN_{\mathrm{int}}. \tag{B.43}$$

Without crossing, the left side is zero and the right side is nonnegative, so the same bound holds.

Cancel the internal updates and complete the square. The selected-policy inequality (B.33) and (B.41) imply

$$\mathrm{Reg}_{\mathrm{LP}} \le v(T - N_{\mathrm{valid}}) + E_{\mathrm{H}} + \overline{M}_R - \sum_{k=1}^{N_{\mathrm{valid}}} \lambda_k g_k.$$

Use (B.43) with the fixed factor v/b. By (B.34), $vN_{\mathrm{int}} \le b\sum_j \lambda_j^{\mathrm{st}}$; the potential bound (B.36) consequently gives

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{LP}} \le E_{\mathrm{H}} + \overline{M}_R + \frac{v}{b}(H + E_C + \overline{M}_C) \\
+ \frac{v}{b} Z_{\mathrm{end}} - \frac{Z_{\mathrm{end}}^2}{2U} + \frac{Q_{\mathrm{sq}}}{2U}.
\end{aligned}$$

Only actual queue operations enter this cancellation. Internal updates receive no reward, and $v(T - N_{\mathrm{valid}})$ counts only missing or discarded physical rewards. Finally,

$$\frac{v}{b} Z_{\mathrm{end}} - \frac{Z_{\mathrm{end}}^2}{2U} = \frac{Uv^2}{2b^2} - \frac{(Z_{\mathrm{end}} - Uv/b)^2}{2U} \le \frac{Uv^2}{2b^2}.$$

Together with $Q_{\mathrm{sq}} \le \overline{Q}_{\mathrm{sq}}$ from Lemma B.13, this proves (B.39). Benchmark conversion adds at most Hv/b and yields (B.40). The deterministic bound follows from $\mathsf{Rew}_T \ge 0$. □

Proof of Theorem 5.1. The definitions imply

$$E_{\mathrm{H}} = \widetilde{O}(H^2 S\sqrt{AT}), \qquad \alpha_{\mathrm{B}} = \widetilde{O}(HS\sqrt{A}), \qquad \beta_{\mathrm{B}} = \widetilde{O}(H^3 S^2 A),$$
$$\overline{C}_{\mathrm{tot}} = \widetilde{O}(Tb + H), \qquad \overline{M}_R = \widetilde{O}(H\sqrt{T} + H), \qquad \overline{M}_C = \widetilde{O}(\sqrt{HTb} + H).$$

Thus the cost-estimation term satisfies

$$\begin{aligned}\frac{v}{b} E_C &\le \widetilde{O}\left(\frac{v}{b}[HS\sqrt{A(Tb+H)} + H^3 S^2 A]\right) \\ &\le \widetilde{O}\left(vHS\sqrt{AT/b} + \frac{vH^3 S^2 A}{b}\right).\end{aligned}$$

The term $HS\sqrt{AH}$ is absorbed by $H^3 S^2 A$, since $H, S, A \ge 1$. Also, $v \le H$ gives

$$\frac{v}{b}\overline{M}_C + \frac{2Hv}{b} \le \widetilde{O}\left(H^{3/2}\sqrt{T/b} + H^2/b\right).$$

For (5),

$$\frac{\overline{Q}_{\mathrm{sq}}}{2U} + \frac{Uv^2}{2b^2} = \frac{H\sqrt{\overline{Q}_{\mathrm{sq}}}}{2b} + \frac{v^2\sqrt{\overline{Q}_{\mathrm{sq}}}}{2Hb} \le \frac{H}{b}\sqrt{\overline{Q}_{\mathrm{sq}}} \le \sqrt{2}H^{3/2}\sqrt{T/b} + \frac{\sqrt{2}H^2\sqrt{1+\ell_{\mathrm{noise}}}}{b}.$$

Finally, $\overline{M}_R$ is absorbed by the displayed Hoeffding reward term. Substitution into (B.40) proves (6). □

C Proof of the lower bound

Why learning can exhaust the budget early. All actions in our construction earn the same reward, so the learner only needs to discover which actions consume less cost. A wrong action slightly increases the chance of entering a bad state. Once entered, that state incurs cost throughout the final part of the episode. Small transition errors therefore accumulate into a substantial budget loss, and each episode lost to early stopping forfeits reward H.

The proof has three steps. First, a testing argument forces many wrong actions during actual attempts. Second, the bad state converts these mistakes into excess cost. Third, we compare the learner's number of credited episodes with that of an ordinary fully informed policy under the same stopping rule. This proves an expected regret lower bound against $\mathrm{OPT}_{\mathrm{HS}}$; no high-probability lower bound is claimed.

We prove Theorem 6.1 with $c_0 = 1/200000000$ and $C_0 = 2^{48}$. Let $m = \lfloor H/2 \rfloor$ be the number of decision steps and $S_0 = \lfloor S/2 \rfloor$ the number of contexts. Since $H^2/8 \le m(H-m) \le H^2/4$ and $S/4 \le S_0 \le S/2$, the theorem's assumption $T \ge C_0 H^2 S(A-1)/b$ implies

$$T \ge \max\left\{48S_0, \frac{65536\, m(H-m)S_0(A-1)}{b}\right\}, \tag{C.1}$$

and, using $H \ge 4$, $S \ge 2$, and $A - 1 \ge 1$,

$$\frac{\sqrt{m(H-m)S_0(A-1)Tb}}{7864320} \ge \frac{\sqrt{C_0/32}\, H^2 S(A-1)}{7864320} \ge 2H. \tag{C.2}$$

All actions below are ordinary; there is no null action or option to skip an episode. Rewards from a budget-crossing episode are discarded.

C.1 Learning the low-cost action requires mistakes

Each decision step and context hides its own favorable action. Observations at other pairs cannot identify that action, even though an adaptive learner may use the entire history when choosing its next action. The following lemma formalizes this by comparing complete adaptive observation laws.

Lemma C.1 (One favorable action per context)**.** *Consider T episodes, each with m decision steps. At each step the context is sampled independently and uniformly from $[S_0]$. For every pair $(h, x) \in [m] \times [S_0]$, one hidden action has Bernoulli signal mean p; the other $A-1$ actions have mean $p+g$. Each signal uses fresh Bernoulli randomness with the stated mean conditional on the past, current context, and selected action. Suppose $A \ge 2$, $T \ge 16S_0$, $0 < p \le 1/2$, $0 < g \le 1 - p$, and*

$$g^2 \le \frac{S_0(A-1)p}{16384T}. \tag{C.3}$$

For every adaptive learner, some joint assignment of the hidden actions forces an expected total of at least $mT/80$ wrong actions.

Proof. Choose every hidden action independently and uniformly from $[A]$. Fix a pair (h, x) and condition on the hidden actions at all other pairs. Its number of visits has distribution $\mathrm{Binomial}(T, 1/S_0)$. For $n = \lfloor T/(2S_0) \rfloor$, Chernoff's inequality gives

$$\mathbb{P}(\text{at least } n \text{ visits}) \ge 1 - e^{-T/(8S_0)} \ge \frac{1}{5}, \qquad n \ge \frac{T}{4S_0}.$$

Condition on any complete context schedule with at least n visits. This conditioning does not favor any hidden-action assignment, since contexts are exogenous. Consider the transcript through the first n visits, including all intervening observations at other pairs.

Let $\mathbb{P}_0$ be its law when all actions at the fixed pair have mean $p+g$, and let $\mathbb{P}_i$ be its law when action i instead has mean p. Write N_i for the number of times action i is selected on these n visits. The learner's conditional action rule is the same under both laws, and observations at other pairs have identical conditional laws. The KL chain rule therefore gives

$$D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_i) = \mathbb{E}_0[N_i]\, \mathrm{kl}(p+g, p) \le \frac{2g^2}{p}\mathbb{E}_0[N_i],$$

where kl is the divergence between Bernoulli laws and $\mathrm{kl}(p+g,p) \le g^2/[p(1-p)]$. Since $0 \le N_i \le n$, Pinsker's inequality implies

$$\mathbb{E}_i[N_i] \le \mathbb{E}_0[N_i] + n\sqrt{\frac{g^2}{p}\mathbb{E}_0[N_i]}.$$

Average over i and use $\sum_i \mathbb{E}_0[N_i] = n$ and Cauchy–Schwarz:

$$\frac{1}{A}\sum_{i=1}^{A}\mathbb{E}_i[N_i] \le \frac{n}{A} + n\sqrt{\frac{ng^2}{Ap}} \le \frac{n}{2} + \frac{n}{128}.$$

The last inequality follows from (C.3) and $n \le T/(2S_0)$. Thus the expected number of wrong actions, averaged over the hidden favorable action at this pair, is at least $n/4$. After averaging over context schedules and the other hidden actions, this pair contributes at least $T/(80S_0)$ expected mistakes. Summing over all mS_0 pairs gives $mT/80$ under the single product prior. Consequently, at least one deterministic joint assignment has that many expected mistakes. □

C.2 A BAD STATE AMPLIFIES THE COST OF MISTAKES

A mistake only matters before the bad state has been reached. The next lemma relates such mistakes to the probability of entering that state. Once this happens, each of the last $H-m$ steps costs one.

Lemma C.2 (Cost amplification)**.** *Let $1 \le m < H$, $p, g > 0$, $p+g \le 1$, $mp \le 1/8$, and $mg \le 1/2$. Contexts are exogenous and independent of the trigger randomness; actions may depend on past observations. During m decision steps, a safe flag changes permanently to a bad flag with probability p after a correct action and $p+g$ after a wrong action. These triggers use fresh randomness. A bad flag costs one at each of the last $H-m$ steps. If M_{safe} counts wrong actions taken while the flag is still safe, the excess expected final-part cost over the all-correct policy lies between*

$$\frac{(H-m)g}{4}\mathbb{E}[M_{\mathrm{safe}}] \quad \textit{and} \quad m(H-m)g.$$

Proof. Condition on the history before the episode, its contexts, and the learner's random seed. Follow the hypothetical within-episode path on which no trigger occurs, and let q be the number of wrong actions on that path. This path is only a way to calculate a probability, not a learner action. The increase in the probability of a bad flag is exactly

$$\Delta = (1-p)^m\left[1 - \left(1 - \frac{g}{1-p}\right)^q\right].$$

Here $p \le 1/8$, $(1-p)^m \ge 7/8$, and $qg/(1-p) \le 2mg \le 1$. Using $1 - e^{-u} \ge u/2$ for $u \in [0,1]$ gives $\Delta \ge gq/4$. Every wrong action counted while safe lies on the no-trigger path, so its conditional expected count is at most q. This proves the lower bound after multiplying by $H-m$ and averaging. For the upper bound, telescoping the difference of the two products of survival probabilities gives $\Delta \le gq \le mg$. □

C.3 CONSTRUCTION AND COMPARISON WITH THE HARD-STOPPING ORACLE

Proof of Theorem 6.1. **The MDP.** Use $2S_0 \le S$ reachable states (z,x), where $z \in \{0,1\}$ is the bad flag and $x \in [S_0]$ is the context. Extra states are unreachable. Initially the flag is safe and the context is uniform. At each of the first m steps, sample the next context independently and uniformly and generate a Bernoulli trigger. At pair (h,x), its mean is p for one hidden action $a_h^*(x)$ and $p+g$ otherwise, with

$$p = \frac{b}{512m(H-m)}, \qquad g = \frac{1}{4096}\sqrt{\frac{S_0(A-1)b}{m(H-m)T}}.$$

The next flag is the maximum of its current value and the trigger. During the final $H-m$ steps, the state stays fixed.

Every step gives deterministic reward one. Each decision step has the same deterministic cost

$$c_{\mathrm{base}} = \frac{3b/2 - (H-m)[1-(1-p)^m]}{m},$$

and each final-part step has cost z. Since $(H-m)[1-(1-p)^m] \le (H-m)mp = b/512$,

$$(3/2 - 1/512)b \le mc_{\text{base}} \le 3b/2, \qquad 0 \le c_{\text{base}} \le \frac{3b}{2m} \le 1.$$

The last inequality uses $b \le H/4$ and $3H \le 8\lfloor H/2 \rfloor$ for $H \ge 4$. Condition (C.1) gives $g \le p/2048$, while $p \le 1/2048$, so the transition probabilities are valid. This is a stationary MDP across episodes; only its stage-dependent transition rows differ within an episode.

The policy selecting every hidden favorable action has expected cost $3b/2$ and reward H. Every other policy has expected cost at least $3b/2$, by Lemma C.2. Hence the LP allocates mass $2T/3$ to a favorable policy and has value

$$\mathrm{LP}_T(B) = \frac{2TH}{3}, \qquad v = \frac{2H}{3}.$$

This LP calculation identifies v but is not used to transfer the lower bound. We compare directly with a policy admissible for $\mathrm{OPT}_{\mathrm{HS}}$.

Every episode costs at least $mc_{\text{base}} > b$, on every path. Thus every learner and the favorable policy cross $B = Tb$ by episode T. They run ordinary policies until that crossing, with no null action or skipping, and discard the crossing reward. In particular, $N_{\text{valid}} = N_{\text{attempt}} - 1$ almost surely.

Learning forces mistakes before physical stopping. Set $n = \lfloor T/2 \rfloor$. Condition (C.1) gives $n \ge T/3 \ge 16S_0$. For analysis only, couple the learner to an n-episode statistical experiment. It agrees with the real interaction up to hard stopping; afterward, generate independent MDP randomness using any fixed ordinary action rule. Also reveal the latent triggers at decision steps whose flag is already bad. The procedure ignores these extra signals when reproducing the real learner. The auxiliary episodes are never executed by the learner and receive no credited reward.

All observed flags and costs are determined by these signals and actions, so they provide no additional hidden-assignment information. Lemma C.1, applied with horizon n, is applicable because

$$g^2 = \frac{S_0(A-1)p}{32768T} \le \frac{S_0(A-1)p}{16384n}.$$

Fix a hidden assignment for which the full experiment has at least $mn/80$ expected wrong actions. We next remove mistakes after a bad flag and after physical stopping.

Uniformly over histories and hidden assignments, an episode has a trigger with probability at most

$$m(p+g) \le \frac{b}{512(H-m)}\left(1 + \frac{1}{2048}\right) < \frac{1}{1000}, \tag{C.4}$$

where $b/(H-m) \le 1/2$. Thus at most $mn/1000$ expected mistakes occur after the flag has become bad.

Let Y count episodes with a bad final flag among the n auxiliary episodes. Their total cost is $mnc_{\text{base}} + (H-m)Y$, and $\mathbb{E}[Y] \le nm(p+g)$. Because $mnc_{\text{base}} \le 3Tb/4$, a crossing before attempt n requires $(H-m)Y > Tb/4$. Markov's inequality gives

$$\mathbb{P}(N_{\text{attempt}} < n) \le \frac{4(H-m)\mathbb{E}[Y]}{Tb} \le \frac{1 + 1/2048}{256} < \frac{1}{250}. \tag{C.5}$$

At most mn mistakes can occur after physical stopping within this experiment. Consequently, the expected number of safe mistakes during actual attempts among its first n episodes is at least

$$mn\left(\frac{1}{80} - \frac{1}{1000} - \frac{1}{250}\right) \ge \frac{mn}{160}.$$

Let E_{attempt} be the expected sum, over all actual attempts, of their conditional mean costs in excess of $3b/2$. These excesses are nonnegative. The parameters satisfy $mp \le 1/8$ and $mg \le 1/2$, so Lemma C.2 applies conditionally before each attempt. Restricting to the actual attempts among the first n episodes therefore gives

$$E_{\text{attempt}} \ge \frac{m(H-m)ng}{640} \ge \frac{m(H-m)Tg}{1920} = \frac{1}{7864320}\sqrt{m(H-m)S_0(A-1)Tb}. \tag{C.6}$$

The auxiliary post-stopping observations were used only for the testing argument; (C.6) counts excess cost incurred during the real interaction.

Comparison under the same stopping rule. Let N^*_{attempt} count attempted episodes when the all-favorable policy is executed until the same budget crossing. Its conditional mean cost per attempt is $3b/2$. Attempt indicators are predictable and both attempt counts are bounded by T, so summing conditional expectations gives

$$\frac{3b}{2}\mathbb{E}[N^*_{\mathrm{attempt}}] = \mathbb{E}\left[\sum_{k=1}^{N^*_{\mathrm{attempt}}} G_k^{c,*}\right] \geq Tb,$$
$$\frac{3b}{2}\mathbb{E}[N_{\mathrm{attempt}}] + E_{\mathrm{attempt}} = \mathbb{E}\left[\sum_{k=1}^{N_{\mathrm{attempt}}} G_k^{c}\right] \leq Tb + H.$$

Here $G_k^{c,*}$ is the realized cost in episode k of the favorable-policy interaction. The first inequality uses its guaranteed crossing; the second uses the at-most-H crossing overshoot.

The favorable policy is admissible for the hard-stopping oracle; we do not need to establish that it is optimal. Both interactions discard exactly their crossing episode, and every other episode earns H. Therefore

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_{\mathrm{HS}}] &= \mathrm{OPT}_{\mathrm{HS}} - \mathbb{E}[\mathsf{Rew}_T] \\ &\geq H\big(\mathbb{E}[N^*_{\mathrm{attempt}}] - \mathbb{E}[N_{\mathrm{attempt}}]\big) \\ &\geq \frac{2H}{3b}(E_{\mathrm{attempt}} - H) \geq \frac{H}{3b}E_{\mathrm{attempt}}.\end{aligned}$$

The last step follows from (C.2) and (C.6), which imply $E_{\mathrm{attempt}} \geq 2H$. Finally, $m(H-m) \geq H^2/8$ and $S_0 \geq S/4$ yield

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{HS}}] \geq \frac{H^2}{23592960\sqrt{32}}\sqrt{\frac{S(A-1)T}{b}} \geq \frac{H^2}{200000000}\sqrt{\frac{S(A-1)T}{b}}.$$

This proves the claimed lower bound directly against $\mathrm{OPT}_{\mathrm{HS}}$. □

D NON-STATIONARY CONSTRUCTION AND PROOF

Algorithm and proof idea. A sliding window discards stale data; known drift margins account for the difference between that window and the current environment. A longer window reduces sampling error but increases drift error. The proof sums these two contributions separately and chooses the window length to balance them.

There is one important control difference from the stationary algorithm. The dynamic benchmark uses full probability mass on ordinary policies, so policy comparison does not require a nonnegative score. Instead, an internal queue cap gives a certificate strong enough to control the reward of different future optimal policies. No internal update is an episode, and the physical hard-stopping rule is unchanged.

D.1 SETTING, VARIATION, AND DYNAMIC BENCHMARK

Episode k has model $\mathcal{M}_k = (\mathcal{S}, \mathcal{A}, H, \mu, P_k, r_k, c_k)$. The complete model sequence is fixed in advance. The initial law μ remains fixed, while reward means, cost means, and transition kernels may change between episodes. Observations satisfy the conditional laws in Section 3, with the corresponding episode index.

For $j \in [T-1]$, assume known deterministic envelopes $d_j^r, d_j^c, d_j^P \geq 0$ satisfying

$$\begin{aligned} \max_{h,s,a} |r_{j+1,h}(s,a) - r_{j,h}(s,a)| &\leq d_j^r, \\ \max_{h,s,a} |c_{j+1,h}(s,a) - c_{j,h}(s,a)| &\leq d_j^c, \\ \max_{h,s,a} \|P_{j+1,h}(\cdot \mid s,a) - P_{j,h}(\cdot \mid s,a)\|_1 &\leq d_j^P. \end{aligned} \tag{D.1}$$

Write $\mathcal{V}_g = \sum_{j=1}^{T-1} d_j^g$ for $g \in \{r, c, P\}$. Only these envelopes, not the unknown model differences, are given to the learner.

For $\pi \in \Pi$, let $R_k(\pi)$ and $C_k(\pi)$ be its expected reward and cost in $\mathcal{M}_k$. Extend them linearly to distributions over Π. Define

$$\mathcal{Q}_k(b) = \{q \in \Delta(\Pi) : C_k(q) \leq b\}, \qquad v_k^* = \max_{q \in \mathcal{Q}_k(b)} R_k(q). \tag{D.2}$$

Assume $\mathcal{Q}_k(b) \neq \varnothing$ for every episode. This comparator has total mass one on ordinary policies and the same expected per-episode budget $b = B/T$. The dynamic regret is

$$\mathrm{Reg}_{\mathrm{dyn}}(T) = \sum_{k=1}^{T} v_k^* - \mathsf{Rew}_T. \tag{D.3}$$

The learner follows the same physical hard-stopping protocol. The comparator in (D.2) is an expected-budget benchmark; its reward is not subject to the learner's realized stopping time.

D.2 QUEUE-CAP ALGORITHM

Choose a window length $w \in [T]$, measured in past episodes. The superscript w identifies sliding-window quantities: $\overline{R}_k^w$ is the analysis-only reward envelope and $\underline{C}_k^w$ is the Bernstein cost lower bound used for the queue. The joint planner and these quantities are defined below.

The cap condition $Z > UH/b$ ensures that each internal decrement has $\lambda b > H$, dominating every future comparator value. This certificate controls reward lost after stopping. The full-mass dynamic comparator permits policy comparison even at a negative score.

No internal operation increments the episode index or collects a sample. Every attempt executes an ordinary policy for all H steps, and the crossing episode has no queue update. Exact policy maximization and the fixed tie-breaking rule are as in Algorithm 1.

Algorithm 3 Sliding-window confidence with queue-cap OGD

Require: $T, H, B = Tb, \delta \in (0,1)$, window w, $U > 0$, and envelopes (D.1)

Set $Z \leftarrow 0$ and cumulative realized cost $C_{\text{used}} \leftarrow 0$
for $k = 1, \dots, T$ **do**
  Construct the window confidence sets in Section D.3
  while $Z > UH/b$ **do**
    $Z \leftarrow [Z - b]_+$ ▷ internal; no episode is consumed
  end while
  $Z_k \leftarrow Z$ and $\lambda_k \leftarrow Z_k/U$
  $(J_k^w(\lambda_k), \pi_k) \leftarrow \mathrm{Plan}_k^w(\lambda_k)$
  Evaluate $\underline{C}_k(\pi_k) \leftarrow \underline{C}_k^w(\pi_k)$ by backward recursion
  Execute π_k for all H steps and observe the trajectory
  $C_{\text{used}} \leftarrow C_{\text{used}} + G_k^c$
  if $C_{\text{used}} > B$ **then**
    Discard G_k^r and **stop**
  end if
  Credit G_k^r; set $Z \leftarrow [Z_k + \underline{C}_k(\pi_k) - b]_+$
  Store the trajectory for subsequent confidence evaluations
end for

D.3 Sliding-window confidence sets

A longer window reduces sampling error but includes older models. We add a known drift margin to each confidence bound. Since the initial law is fixed, its estimate uses all preceding initial states.

Fix $w \in [T]$ and $\varepsilon = \delta/8$. Before episode k, use the window

$$\mathcal{W}_{k,w} = \{(k-w) \vee 1, \dots, k-1\}, \qquad N_{k,h}^w(s,a) = \sum_{j \in \mathcal{W}_{k,w}} \mathbf{1}\{s_{j,h} = s, a_{j,h} = a\}.$$

An interval is empty if its lower endpoint exceeds its upper endpoint. Let $\widehat{r}_{k,h}^w, \widehat{c}_{k,h}^w, \widehat{P}_{k,h}^w$ be the corresponding sample averages, with the same zero-count conventions as before. The initial-law estimate $\widehat{\mu}_k$ continues to use all $k-1$ preceding initial states. Define the known local margins

$$\nu_{k,w}^g = \sum_{j=(k-w)\vee 1}^{k-1} d_j^g, \qquad g \in \{r, c, P\}. \tag{D.4}$$

Every increment enters at most w margins, so

$$\sum_{k=1}^{T} \nu_{k,w}^g \le w \mathcal{V}_g. \tag{D.5}$$

For the Hoeffding confidence radii, use

$$\ell_{\mathrm{H,w}} = \log \frac{32SAHT^2}{\varepsilon}, \qquad \eta_{P,w} = S \log 2 + \log \frac{16SAHT^2}{\varepsilon}.$$

Define β_w, β_w^P by (A.2), replacing $\ell_{\mathrm{H}}, \eta_P$ by $\ell_{\mathrm{H,w}}, \eta_{P,w}$. The initial-state radius α_μ and η_μ are unchanged. For the auxiliary reward envelope only, put

$$B_{k,h}^{r,w}(s,a) = \beta_w(n) + \nu_{k,w}^r + \frac{(H-h)}{2}\big(\beta_w^P(n) + \nu_{k,w}^P\big), \quad n = N_{k,h}^w(s,a). \tag{D.6}$$

Define $\overline{R}_k^w$ by (A.9) with the window model and $B^{r,w}$, followed by the initial-law correction in (A.10). As in Section A.3, this is an analysis-only envelope; the algorithm instead runs the joint recursion below.

On the confidence event, the base Bernstein sets cover the window-averaged model. Enlarging them by the drift margins covers the current model. The superscript w distinguishes these base sets from the all-history sets in Section A.1. Set

$$\ell_{\mathrm{B,w}} = \log \frac{12HS^2AT^3}{\varepsilon}. \tag{D.7}$$

For $n \geq 1$, define the base sets

$$\mathsf{Cl}_n^w(\widehat{z}) = \left\{ q \in [0,1] : |\widehat{z} - q| \leq \sqrt{\frac{2q\ell_{\mathrm{B,w}}}{n}} + \frac{3\ell_{\mathrm{B,w}}}{n} \right\},$$
$$\mathsf{Prob}_n^w(\widehat{p}) = \Delta(\mathcal{S}) \cap \prod_{s'} \mathsf{Cl}_n^w(\widehat{p}(s')). \tag{D.8}$$

Set $\mathsf{Cl}_0^w = [0,1]$ and $\mathsf{Prob}_0^w = \Delta(\mathcal{S})$. The cost and transition confidence sets are their drift enlargements:

$$\mathcal{C}_{k,h}^w(s,a) = \left\{ q \in [0,1] : \exists q_0 \in \mathsf{Cl}_n^w(\widehat{c}_{k,h}^w(s,a)),\ |q - q_0| \leq \nu_{k,w}^c \right\},$$
$$\mathcal{P}_{k,h}^w(s,a) = \left\{ q \in \Delta(\mathcal{S}) : \exists q_0 \in \mathsf{Prob}_n^w(\widehat{P}_{k,h}^w(\cdot \mid s,a)),\ \|q - q_0\|_1 \leq \nu_{k,w}^P \right\}, \tag{D.9}$$

where $n = N_{k,h}^w(s,a)$. Use $\mathsf{Prob}_{k-1}^w(\widehat{\mu}_k)$ for the initial law, without enlargement. These sets are nonempty and compact, including outside the confidence event.

For each policy, define the robust recursion (A.7) using $\mathcal{C}_{k,h}^w, \mathcal{P}_{k,h}^w$ and denote its state values by $\widetilde{V}_{k,h}^{\pi,w}$. Algorithm 3 evaluates it only for the policy selected by joint planning. Its cost lower bound minimizes the initial value over the initial-law confidence set:

$$\underline{C}_k^w(\pi) = \min_{p \in \mathsf{Prob}_{k-1}^w(\widehat{\mu}_k)} p^\top \widetilde{V}_{k,1}^{\pi,w}. \tag{D.10}$$

Joint planning. Use (4) with the window quantities

$$r_{k,h}^{+,w}(s,a) = \min\{1, \widehat{r}_{k,h}^w(s,a) + \beta_w(n) + \nu_{k,w}^r\},$$
$$c_{k,h}^{-,w}(s,a) = \min \mathcal{C}_{k,h}^w(s,a),$$
$$\mathcal{P}_{k,h}^{\mathrm{J},w}(s,a) = \{p \in \mathcal{P}_{k,h}^w(s,a) : \|p - \widehat{P}_{k,h}^w(\cdot \mid s,a)\|_1 \leq \beta_w^P(n) + \nu_{k,w}^P\},$$
$$\mathcal{M}_k^{\mathrm{J},w} = \{\mu' \in \mathsf{Prob}_{k-1}^w(\widehat{\mu}_k) : \|\mu' - \widehat{\mu}_k\|_1 \leq \alpha_\mu(k-1)\}, \tag{D.11}$$

where $n = N_{k,h}^w(s,a)$, and denote the output by $(J_k^w(\lambda), \pi_k^{\lambda,w}) = \mathrm{Plan}_k^w(\lambda)$. On the confidence events of Lemma D.2,

$$\max_{\pi \in \Pi} \{R_k(\pi) - \lambda C_k(\pi)\} \leq J_k^w(\lambda) \leq \overline{R}_k^w(\pi_k^{\lambda,w}) - \lambda \underline{C}_k^w(\pi_k^{\lambda,w}). \tag{D.12}$$

Indeed, the current model belongs to both intersected sets. The proof of Lemma A.1 applies with $\beta(n)$ replaced by $\beta_w(n) + \nu_{k,w}^r$ and $\beta^P(n)$ by $\beta_w^P(n) + \nu_{k,w}^P$; the resulting reward correction is exactly (D.6). Candidate costs are still bounded below by (D.10). This gives (D.12) deterministically on the existing events, with no additional failure probability.

D.4 Dynamic regret bound

The error has a sampling part that decreases with the window length and a drift part that increases with it. Cost errors are weighted by H/b, since a cost discrepancy can remove episodes worth at most H each. The bound below makes the window tradeoff explicit.

The next result retains the dependence on the state and action spaces, the horizon, and the budget. We then choose the window to obtain the main-body rate, keeping b explicit even when it varies with T.

Theorem D.1 (Dynamic regret for a fixed window)**.** *Assume* (D.1) *and that $\mathcal{Q}_k(b)$ is nonempty in every episode. Choose U as in* (5)*. For any prescribed $w \in [T]$, with probability at least $1 - \delta$, Algorithm 3 satisfies*

$$\mathrm{Reg}_{\mathrm{dyn}}(T) \leq \widetilde{O}\left(\frac{H^2 S\sqrt{A}\,T}{\sqrt{w}}(1 + b^{-1/2}) + \frac{H^4 S^2 A T}{bw} + H^{3/2}\sqrt{T/b} + \frac{H^2}{b} \right)$$
$$+ 2Hw\mathcal{V}_r + \frac{8H^2 w}{b}\mathcal{V}_c + w\left[\frac{H(H-1)}{2} + \frac{3H^2(H-1)}{b} \right]\mathcal{V}_P. \tag{D.13}$$

D.5 SLIDING-WINDOW CONFIDENCE

Window observations estimate a random average of past conditional means. We first control the centered observations, then use the drift margin to reach the current model. Deterministic variance bins justify the Bernstein bound.

Lemma D.2 (Current-model feasibility)**.** *With failure probability at most $\varepsilon/4$, the window Hoeffding estimates satisfy, simultaneously at every reached episode and row,*

$$|\widehat{r}^{w}_{k,h}(s,a) - r_{k,h}(s,a)| \le \beta_w(N^{w}_{k,h}(s,a)) + \nu^{r}_{k,w},$$
$$\|\widehat{P}^{w}_{k,h}(\cdot \mid s,a) - P_{k,h}(\cdot \mid s,a)\|_1 \le \beta^{P}_{w}(N^{w}_{k,h}(s,a)) + \nu^{P}_{k,w}. \tag{D.14}$$

The all-history initial-law bound in Lemma B.4 also holds. With failure probability at most $\varepsilon/2$, the current cost means, transition rows, and initial law belong to their Bernstein sets (D.9) *and* $\mathsf{Prob}^{w}_{k-1}(\widehat{\mu}_k)$.

Proof. Fix the deterministic episode window $\mathcal{W}_{k,w}$ and a row. List the observations selected in this window. The conditional mean of a selected scalar observation is its mean in the episode in which it was collected. For a positive attained count n, denote their average conditional mean by $\bar{p}$ and their empirical mean by $\widehat{p}$. The selected observations are not assumed identically distributed.

For Hoeffding concentration, cap the window-row martingale after n selected observations and pad it to the global deterministic horizon HT. The attained-count argument in Lemma B.2 bounds $|\widehat{p} - \bar{p}|$ by $\sqrt{\ell_{\mathrm{H,w}}/(2n)}$. The union now includes $k \le T$ and $n \le T$, so the reward means cost $2SAHT^2 e^{-\ell_{\mathrm{H,w}}} = \varepsilon/16$. The sign-vector argument for the averaged conditional transition row costs $2^S SAHT^2 e^{-\eta_{P,w}} = \varepsilon/16$, and the initial-law event costs $\varepsilon/16$. Every historical mean or row in the window is within $\nu^{r}_{k,w}$ or $\nu^{P}_{k,w}$ of the current one by (D.1). Averaging and the triangle inequality give (D.14). Clipped radii and zero counts are handled as in Lemma B.4.

For Bernstein concentration, form the centered window-row martingale for a cost observation or a transition indicator. Its quadratic variation is at most A_n, the sum of conditional means of its accepted observations, with $0 \le A_n \le n$. This quantity is random. For each deterministic integer $j \in \{1, \dots, n\}$, apply (B.3) to each sign with $\sigma^2 = j$ and $\ell_{\mathrm{tail}} = \ell_{\mathrm{B,w}}$. Outside the union of these bad events, use $j = \max\{1, \lceil A_n \rceil\} \le n$ to obtain

$$|M| \le \sqrt{2(A_n + 1)\ell_{\mathrm{B,w}}} + \frac{2\ell_{\mathrm{B,w}}}{3} \le \sqrt{2A_n\ell_{\mathrm{B,w}}} + 3\ell_{\mathrm{B,w}}.$$

The last inequality uses $\ell_{\mathrm{B,w}} > 1$. On attainment of count n, $M = n(\widehat{p} - \bar{p})$ and $A_n = n\bar{p}$. Thus the random average $\bar{p}$ belongs to $\mathsf{CI}^{w}_{n}(\widehat{p})$. The same argument applies coordinatewise to the averaged conditional transition row, placing that row in $\mathsf{Prob}^{w}_{n}(\widehat{P}^{w})$. Neither the inequality nor its union bound conditions on $\bar{p}$ or on attainment of a random count.

Union over the windows, counts, variance bins, coordinates, and two signs. The failure probability is at most

$$2T^3(HSA + HS^2A + S)e^{-\ell_{\mathrm{B,w}}} \le 6HS^2AT^3 e^{-\ell_{\mathrm{B,w}}} = \varepsilon/2.$$

This also covers initial-state averages, for which there is no drift. The averaged scalar mean is within $\nu^{c}_{k,w}$ of the current mean, and the averaged conditional transition row is within $\nu^{P}_{k,w}$ in ℓ_1 distance of the current row. The enlargements in (D.9) therefore contain the current model. At zero count this is deterministic. □

The next lemma separates sampling uncertainty from drift by comparing each enlarged-set candidate with a representative in its base set. Its bound holds for every downstream vector, including one computed from the window data.

Lemma D.3 (Diameters of enlarged Bernstein sets)**.** *Let $\bar{n} = n \vee 1$. If p, q belong to the same scalar base set enlarged by a radius $\nu \ge 0$, then*

$$|p - q| \le 4\sqrt{\frac{2p\ell_{\mathrm{B,w}}}{\bar{n}}} + \frac{18\ell_{\mathrm{B,w}}}{\bar{n}} + 4\nu. \tag{D.15}$$

If probability vectors p, q belong to the same categorical base set enlarged by ℓ_1 radius ν, then, for every $f \in \mathbb{R}^S$,

$$|(p - q)^\top f| \le 4\sqrt{\frac{2S\ell_{\mathrm{B,w}}\operatorname{Var}_p(f)}{\bar{n}}} + \frac{16S\ell_{\mathrm{B,w}}\operatorname{span}(f)}{\bar{n}} + 3\nu\operatorname{span}(f). \tag{D.16}$$

Proof. For base scalar values $p_0, q_0 \in \mathsf{CI}_n^w(\widehat{z})$, the proof of Lemma B.6, with scalar linear radius $3\ell_{\mathrm{B,w}}/n$, gives

$$|p_0 - q_0| \le 4\sqrt{\frac{2p_0\ell_{\mathrm{B,w}}}{\bar{n}}} + \frac{14\ell_{\mathrm{B,w}}}{\bar{n}}.$$

Choose base-set representatives with $|p - p_0|, |q - q_0| \le \nu$. Then $p_0 \le p + \nu$, and

$$|p - q| \le 2\nu + 4\sqrt{2(p+\nu)\ell_{\mathrm{B,w}}/\bar{n}} + 14\ell_{\mathrm{B,w}}/\bar{n}.$$

Use $4\sqrt{2\nu\ell_{\mathrm{B,w}}/\bar{n}} \le 4\ell_{\mathrm{B,w}}/\bar{n} + 2\nu$ to obtain (D.15).

For categorical representatives p_0, q_0, the corresponding base diameter is

$$|(p_0 - q_0)^\top f| \le 4\sqrt{2S\ell_{\mathrm{B,w}}\,\mathrm{Var}_{p_0}(f)/\bar{n}} + 14S\ell_{\mathrm{B,w}}\,\mathrm{span}(f)/\bar{n}.$$

The two representative discrepancies add at most $\nu\,\mathrm{span}(f)$ by (B.9). Centering at $p^\top f$ gives

$$\mathrm{Var}_{p_0}(f) \le \mathrm{Var}_p(f) + \frac{\nu}{2}\,\mathrm{span}(f)^2.$$

Indeed, the squared centered function has range at most $\mathrm{span}(f)^2$. The extra square-root contribution is bounded by

$$4\sqrt{S\ell_{\mathrm{B,w}}\nu/\bar{n}}\;\mathrm{span}(f) \le (2S\ell_{\mathrm{B,w}}/\bar{n} + 2\nu)\,\mathrm{span}(f).$$

Adding these terms proves (D.16). At $n = 0$, both claims follow from the full scalar or simplex range and $\ell_{\mathrm{B,w}} > 1$. □

D.6 Cumulative evaluation errors

Window counts may decrease. For analysis only, partition episodes into blocks of length at most w and lower-bound window counts by earlier visits in the same block. This lets us sum statistical widths; drift is summed using (D.5).

All quantities in this subsection use the policies selected by Algorithm 3 and the current true model $\mathcal{M}_k$. Define the inclusive-prefix cost

$$C_{\mathrm{tot}}^w = \sum_{k=1}^{N_{\mathrm{attempt}}} C_k(\pi_k).$$

The count bound has the same form as in the stationary proof, except that visits are counted separately in $\lceil T/w \rceil$ blocks:

$$\Lambda_w = \frac{HSA\lceil T/w \rceil(2 + \log w) + H\log(2/\varepsilon)}{1 - e^{-1}} + 2 + \log T. \tag{D.17}$$

The last two terms account for the all-history initial-state estimate. Define the Bernstein coefficients

$$\alpha_{\mathrm{B},w} = 28\sqrt{SH\ell_{\mathrm{B},w}\Lambda_w}, \qquad \beta_{\mathrm{B},w} = 132H^2S\ell_{\mathrm{B},w}\Lambda_w. \tag{D.18}$$

The statistical reward-error envelope is

$$\begin{aligned} E_{\mathrm{H,w}} = {} & 2H\sqrt{2\eta_\mu T} + [2H\sqrt{2\ell_{\mathrm{H,w}}} + H(H-1)\sqrt{2\eta_{P,w}}]\,\lceil T/w \rceil\sqrt{SAw} \\ & + H(H+1)\sqrt{2T\ell_{\mathrm{width}}}. \end{aligned} \tag{D.19}$$

Set

$$\begin{aligned} \mathcal{D}_r(w) &= 2Hw\mathcal{V}_r + \frac{H(H-1)}{2}w\mathcal{V}_P, \\ \mathcal{D}_{\mathrm{B}}(w) &= 8Hw\mathcal{V}_c + 3H(H-1)w\mathcal{V}_P, \\ E_{C,w} &= \alpha_{\mathrm{B},w}\sqrt{\overline{C}_{\mathrm{tot}}} + \beta_{\mathrm{B},w} + \mathcal{D}_{\mathrm{B}}(w). \end{aligned} \tag{D.20}$$

Here $\mathcal{D}_r(w)$ and $\mathcal{D}_{\mathrm{B}}(w)$ bound accumulated reward and cost drift, and $E_{C,w}$ is the resulting cost-error bound. The logarithms $\ell_{\mathrm{noise}}, \ell_{\mathrm{width}}$ and the budget, noise, and queue bounds $\overline{C}_{\mathrm{tot}}, \overline{M}_R, \overline{M}_C, \overline{Q}_{\mathrm{sq}}$ are unchanged from the stationary proof; they depend only on T, H, b, δ.

Lemma D.4 (Sliding-window cumulative bounds)**.** *The following statements hold on events of failure probability at most ε each:*

$$\overline{R}_k^w(\pi) \ge R_k(\pi) \quad \textit{for every reached } k \textit{ and } \pi,$$

$$\sum_{k=1}^{N_{\mathrm{attempt}}} (\overline{R}_k^w(\pi_k) - R_k(\pi_k)) \le E_{\mathrm{H,w}} + \mathcal{D}_r(w); \tag{D.21}$$

$$0 \le \underline{C}_k^w(\pi) \le C_k(\pi) \quad \textit{for every reached } k \textit{ and } \pi,$$

$$\Gamma^+_{\mathrm{B},w} := \sum_{k=1}^{N_{\mathrm{attempt}}} (C_k(\pi_k) - \underline{C}_k^w(\pi_k)) \le \alpha_{\mathrm{B},w}\sqrt{C^w_{\mathrm{tot}}} + \beta_{\mathrm{B},w} + \mathcal{D}_{\mathrm{B}}(w). \tag{D.22}$$

The first event controls all of (D.21)*, and the second controls* (D.22)*.*

Proof. On the Hoeffding event of Lemma D.2, backward induction for the auxiliary envelope, using the current true kernel, proves $\overline{R}_k^w(\pi) \ge R_k(\pi)$. The same one-step error calculation as in (B.10) gives the statistical width

$$\omega^w_{k,h}(s,a) = 2\beta_w(N^w_{k,h}(s,a)) + (H-h)\beta^P_w(N^w_{k,h}(s,a))$$

plus the per-episode drift term $2H\nu^r_{k,w} + H(H-1)\nu^P_{k,w}/2$. The initial-law contribution remains $H\alpha_\mu(k-1)$.

Partition $[T]$ into $\lceil T/w \rceil$ consecutive blocks of length at most w. At a visit in a block, the window count is at least the number of earlier visits to the same row in that block. Thus (B.11), applied blockwise, gives

$$\sum_{k=1}^{N_{\mathrm{attempt}}} \sum_h \omega^w_{k,h}(s_{k,h},a_{k,h}) \le [2H\sqrt{2\ell_{\mathrm{H,w}}} + H(H-1)\sqrt{2\eta_{P,w}}]\,\lceil T/w\rceil\sqrt{SAw}.$$

The predictable-attempt Azuma argument for these widths costs $\varepsilon/8$ and contributes $H(H+1)\sqrt{2T\ell_{\mathrm{width}}}$. The initial count sum contributes $2H\sqrt{2\eta_\mu T}$. Summing the drift terms with (D.5) proves (D.21), with failure at most $\varepsilon/4 + \varepsilon/8 \le \varepsilon$.

For the Bernstein cost bound, define the current occupancies by

$$\rho^w_{k,h}(s,a) = \Pr_{\mu,P_k,\pi_k}(s_h = s, a_h = a \mid \mathcal{F}_{k-1}).$$

Combine the transition and initial-state reciprocal counts into

$$\Xi_w = \sum_{k=1}^{N_{\mathrm{attempt}}} \left[\frac{1}{(k-1)\vee 1} + \sum_{h,s,a} \frac{\rho^w_{k,h}(s,a)}{N^w_{k,h}(s,a)\vee 1} \right].$$

At each attempted episode, let $Y_k = H^{-1}\sum_h [N^w_{k,h}(s_{k,h},a_{k,h})\vee 1]^{-1}$, and pad Y_k by zero after stopping. Then $0 \le Y_k \le 1$, and $H\mathbb{E}[Y_k \mid \mathcal{F}_{k-1}]$ is the transition part of the corresponding summand in Ξ_w. The attempt indicator is predictable. For a row visited $n \ge 1$ times in one block, the realized reciprocal sum is at most $1 + \sum_{j=1}^{n-1} 1/j \le 2 + \log w$. Apply Lemma B.3 once to Y_k, with exponent $\log(2/\varepsilon)$. Together with the pathwise initial-count bound $\sum_{k=1}^{N_{\mathrm{attempt}}} [(k-1)\vee 1]^{-1} \le 2 + \log T$, this gives

$$\Xi_w \le \Lambda_w$$

except with probability $\varepsilon/2$.

Intersect this event with the Bernstein event of Lemma D.2. Feasibility of the current true model gives $0 \le \widetilde{V}^{\pi,w}_{k,h} \le V^{c,\pi}_{k,h}$ and policywise pessimism. Define the nonnegative local Bellman gaps using $c_{k,h}, P_{k,h}$ as in (B.19). The exact gap recursion (B.20) holds within each episode.

For the current-model mean-cost path $F_k = \sum_h c_{k,h}(s_{k,h}, \pi_{k,h}(s_{k,h}))$, the conditional telescoping and variance calculation (B.21)–(B.22) remain valid: the model is fixed throughout that episode. Define the total robust downstream variance by

$$\Omega_{\mathrm{val},w} = \sum_{k=1}^{N_{\mathrm{attempt}}} \left[\mathrm{Var}_\mu(\widetilde{V}^{\pi_k,w}_{k,1}) + \sum_{h=1}^{H-1} \mathbb{E}_{\mu,P_k,\pi_k} \mathrm{Var}_{P_{k,h}}(\widetilde{V}^{\pi_k,w}_{k,h+1}) \right].$$

The value-gap comparison (B.23) therefore gives

$$\Omega_{\mathrm{val},w} \le 2HC^w_{\mathrm{tot}} + 2H^2\Gamma^+_{\mathrm{B},w}. \tag{D.23}$$

Apply Lemma D.3 to each local gap. Weighted Cauchy–Schwarz uses the reciprocal potentials just proved. The scalar linear term contributes $18\ell_{\mathrm{B,w}}\Lambda_w$, and the categorical linear terms contribute at most $16SH\ell_{\mathrm{B,w}}\Lambda_w$. The scalar drift contributes at most $4H\sum_k \nu^c_{k,w}$. The transition drift contributes at most $3H(H-1)\sum_k \nu^P_{k,w}/2$, because the stage-$h$ downstream span is at most $H-h$. There is no initial-law drift. Thus

$$\begin{aligned}\Gamma^+_{\mathrm{B},w} \le\ & 4\sqrt{2\ell_{\mathrm{B,w}}\Lambda_w C^w_{\mathrm{tot}}} + 4\sqrt{2S\ell_{\mathrm{B,w}}\Lambda_w\Omega_{\mathrm{val},w}} \\ & + 18\ell_{\mathrm{B,w}}\Lambda_w + 16SH\ell_{\mathrm{B,w}}\Lambda_w \\ & + 4Hw\mathcal{V}_c + \frac{3H(H-1)}{2}w\mathcal{V}_P.\end{aligned} \tag{D.24}$$

Substitute (D.23) and absorb the resulting term in $\sqrt{\Gamma^+_{\mathrm{B},w}}$ using $\kappa\sqrt{x} \le (x+\kappa^2)/2$ with $\kappa = 8H\sqrt{S\ell_{\mathrm{B,w}}\Lambda_w}$. The square-root coefficient is at most $(8\sqrt{2}+16)\sqrt{SH\ell_{\mathrm{B,w}}\Lambda_w}$, and the linear coefficient is at most $(64H^2S + 36 + 32SH)\ell_{\mathrm{B,w}}\Lambda_w$. These are bounded by (D.18); the drift terms double to $\mathcal{D}_{\mathrm{B}}(w)$. This proves (D.22), with total failure probability at most ε. □

D.7 Completing the dynamic regret proof

The confidence analysis is the only part that changes with the window. The budget and observation-noise bounds remain valid episode by episode. We first insert those nonnegative bounds, then use the queue cap to pay for comparator reward lost after stopping.

Proof of Theorem D.1. The two events in Lemma D.4 cost at most 2ε. Lemma B.9, applied to the current conditional costs $C_k(\pi_k)$, gives $C^w_{\mathrm{tot}} \le \overline{C}_{\mathrm{tot}}$ with failure at most ε. The two deviation bounds in Lemma B.10 also apply: attempts and policies are predictable, episode totals remain in $[0,H]$, and cost quadratic variation is at most HC^w_{tot}. They cost at most 2ε in addition to the budget event. After removing the nonnegative estimation gaps of the crossing attempt, these events give

$$\begin{aligned}\sum_{k=1}^{N_{\mathrm{valid}}} \big(\overline{R}^w_k(\pi_k) - G^r_k\big) &\le E_{\mathrm{H,w}} + \mathcal{D}_r(w) + \overline{M}_R, \\ \sum_{k=1}^{N_{\mathrm{valid}}} \big(G^c_k - \underline{C}_k(\pi_k)\big) &\le E_{C,w} + \overline{M}_C.\end{aligned} \tag{D.25}$$

Both right-hand sides are nonnegative.

Choose $q^*_k \in \mathcal{Q}_k(b)$ attaining v^*_k. Because q^*_k has total mass one, (D.12) gives

$$v^*_k - \lambda_k b \le \sum_{\pi} q^*_{k,\pi}(R_k(\pi) - \lambda_k C_k(\pi)) \le J^w_k(\lambda_k) \le \overline{R}^w_k(\pi_k) - \lambda_k\underline{C}_k(\pi_k).$$

Thus the regret on valid episodes is bounded by the first sum in (D.25) minus $\sum_{k=1}^{N_{\mathrm{valid}}}\lambda_k(\underline{C}_k(\pi_k)-b)$. Unlike the stationary subprobability comparator, this comparison needs no sign restriction.

Let N_{int} count the actual internal cap decrements and let Z_{end} be the actual terminal queue. Projection gives

$$Z_{\mathrm{end}} \ge \sum_{k=1}^{N_{\mathrm{valid}}} \underline{C}_k(\pi_k) - bN_{\mathrm{valid}} - bN_{\mathrm{int}}.$$

On crossing, $\sum_{k=1}^{N_{\mathrm{valid}}} G^c_k > Tb - H$, so (D.25) implies

$$b(T - N_{\mathrm{valid}}) \le H + E_{C,w} + \overline{M}_C + Z_{\mathrm{end}} + bN_{\mathrm{int}}.$$

Since every comparator reward is at most H,

$$\sum_{t=N_{\mathrm{valid}}+1}^{T} v^*_t \le \frac{H}{b}\big(H + E_{C,w} + \overline{M}_C + Z_{\mathrm{end}} + bN_{\mathrm{int}}\big). \tag{D.26}$$

When there is no crossing, the left-hand side is zero and this bound still holds because its right-hand side is nonnegative.

Immediately before each internal decrement, the cap test gives $\lambda_j^{\mathrm{cap}} b > H$. Lemma B.12 therefore yields

$$-\sum_{k=1}^{N_{\mathrm{valid}}} \lambda_k(\underline{C}_k(\pi_k) - b) \le -\frac{Z_{\mathrm{end}}^2}{2U} + \frac{Q_{\mathrm{sq}}}{2U} - b\sum_{j=1}^{N_{\mathrm{int}}} \lambda_j^{\mathrm{cap}} \le -\frac{Z_{\mathrm{end}}^2}{2U} + \frac{Q_{\mathrm{sq}}}{2U} - HN_{\mathrm{int}},$$

where $Q_{\mathrm{sq}} = \sum_{k=1}^{N_{\mathrm{valid}}} (\underline{C}_k(\pi_k) - b)^2 + N_{\mathrm{int}} b^2$ counts only update inputs that the algorithm actually uses. The square-increment argument (B.37)–(B.38) applies unchanged to the cap loop: total internal decrease cannot exceed preceding positive physical increase, and each internal block has at most one partial decrement. With $\underline{C}_k(\pi_k) \le C_k(\pi_k)$ and $C_{\mathrm{tot}}^w \le \overline{C}_{\mathrm{tot}}$, it gives $Q_{\mathrm{sq}} \le \overline{Q}_{\mathrm{sq}}$.

Combine the valid-episode comparison with (D.26). The terms HN_{int} cancel; no internal operation is counted as a regret response. Finally, $(H/b)Z_{\mathrm{end}} - Z_{\mathrm{end}}^2/(2U) \le UH^2/(2b^2)$ gives the finite-sample certificate

$$\begin{aligned}\mathrm{Reg}_{\mathrm{dyn}}(T) \le E_{\mathrm{H,w}} + \mathcal{D}_r(w) + \overline{M}_R + \frac{H}{b}(E_{C,w} + \overline{M}_C) \\ + \frac{\overline{Q}_{\mathrm{sq}}}{2U} + \frac{UH^2}{2b^2} + \frac{H^2}{b}.\end{aligned} \tag{D.27}$$

The total failure probability is at most $5\varepsilon < \delta$.

To simplify this bound, use $\lceil T/w \rceil \le 2T/w$:

$$E_{\mathrm{H,w}} = \widetilde{O}(H^2 S\sqrt{A}\,T/\sqrt{w}), \qquad \alpha_{\mathrm{B},w} = \widetilde{O}(HS\sqrt{AT/w}), \qquad \beta_{\mathrm{B},w} = \widetilde{O}(H^3 S^2 A\,T/w).$$

Because $\overline{C}_{\mathrm{tot}} = \widetilde{O}(Tb + H)$, the statistical cost term satisfies

$$\frac{H}{b}\left(\alpha_{\mathrm{B},w}\sqrt{\overline{C}_{\mathrm{tot}}} + \beta_{\mathrm{B},w}\right) \le \widetilde{O}\left(\frac{H^2 S\sqrt{A}\,T}{\sqrt{bw}} + \frac{H^4 S^2 AT}{bw}\right).$$

Here the part from $\sqrt{H}$ in $\sqrt{Tb + H}$ is absorbed by the additive Bernstein term. The drift contribution is exactly

$$\mathcal{D}_r(w) + \frac{H}{b}\mathcal{D}_{\mathrm{B}}(w) = 2Hw\mathcal{V}_r + \frac{8H^2 w}{b}\mathcal{V}_c + w\left[\frac{H(H-1)}{2} + \frac{3H^2(H-1)}{b}\right]\mathcal{V}_P.$$

The choice (5) is $U = (b/H)\sqrt{\overline{Q}_{\mathrm{sq}}}$, so the two quadratic terms in (D.27) sum to $(H/b)\sqrt{\overline{Q}_{\mathrm{sq}}}$. Together with the noise bounds, they contribute $\widetilde{O}(H^{3/2}\sqrt{T/b} + H^2/b)$; the reward-noise term is absorbed by $E_{\mathrm{H,w}}$. This proves (D.13). □

D.8 Choosing the window with the budget dependence explicit

The window must account for two sampling terms, not only $T/\sqrt{w}$. The additive Bernstein term $T/(bw)$ can matter when the budget is small. We balance each term against the drift and use the larger window.

Proof of Theorem 7.1. In this proof only, write $a_b = 1 + b^{-1/2}$ for the coefficient of the square-root sampling term. Recall the budget-weighted variation $\mathcal{V}_b = \mathcal{V}_r + (\mathcal{V}_c + \mathcal{V}_P)/b$. Equation (D.13) implies

$$\mathrm{Reg}_{\mathrm{dyn}}(T) \le \widetilde{O}_{H,S,A}\left(\frac{a_b T}{\sqrt{w}} + \frac{T}{bw} + w\mathcal{V}_b + a_b\sqrt{T} + \frac{1}{b}\right). \tag{D.28}$$

The constants here are independent of b, T, and the variation budgets. In particular, the transition-drift coefficient without $1/b$ is absorbed using $b \le H$:

$$\frac{H(H-1)}{2}\mathcal{V}_P \le \frac{H^2(H-1)}{2b}\mathcal{V}_P.$$

The queue and noise terms are covered by $a_b\sqrt{T} + 1/b$. All logarithmic factors in the fixed-window bound are bounded uniformly for $1 \le w \le T$. The selected window depends only on the known envelopes and prescribed parameters, so the fixed-window probability guarantee applies directly; no data-dependent choice or extra union bound is used.

If $\mathcal{V}_b = 0$, choose $w = T$. Substitution into (D.28) gives the first two terms of (11). Otherwise define the unrounded balancing scale

$$z = \max\left\{(a_b T/\mathcal{V}_b)^{2/3}, \sqrt{T/(b\mathcal{V}_b)}\right\}.$$

We verify the window choice, including both clipping cases.

If $1 \le z \le T$, then $z \le w \le 2z$, and

$$\frac{a_b T}{\sqrt{w}} \le a_b^{2/3} T^{2/3} \mathcal{V}_b^{1/3}, \qquad \frac{T}{bw} \le \sqrt{\frac{T\mathcal{V}_b}{b}},$$
$$w\mathcal{V}_b \le 2\max\left\{a_b^{2/3} T^{2/3} \mathcal{V}_b^{1/3}, \sqrt{\frac{T\mathcal{V}_b}{b}}\right\}.$$

These are exactly the two tracking terms in (11).

If $z > T$, at least one balancing scale exceeds T, which implies

$$T\mathcal{V}_b \le \max\{a_b\sqrt{T}, 1/b\}.$$

With $w = T$, both sampling terms and the drift are therefore bounded by the baseline $a_b\sqrt{T} + 1/b$.

If $z < 1$, then $\mathcal{V}_b > a_b T$, so $a_b^{2/3} T^{2/3} \mathcal{V}_b^{1/3} > a_b T \ge T$. The pathwise bound $\mathrm{Reg}_{\mathrm{dyn}}(T) \le HT$ proves (11) in this case. This completes all regimes, without assuming that b stays bounded away from zero as T grows. □

Small-budget form. Let $0 < b \le \min\{1, H\}$ and $\mathcal{V} = \mathcal{V}_r + \mathcal{V}_c + \mathcal{V}_P$. For the same algorithm and window choice, Theorem 7.1 gives, with probability at least $1 - \delta$,

$$\mathrm{Reg}_{\mathrm{dyn}}(T) \le \widetilde{O}_{H,S,A}\left(\sqrt{\frac{T}{b}} + \frac{1}{b} + \frac{T^{2/3}\mathcal{V}^{1/3}}{b^{2/3}} + \frac{\sqrt{T\mathcal{V}}}{b}\right), \tag{D.29}$$

and the pathwise cap HT still applies. Indeed, $1 + b^{-1/2} \le 2b^{-1/2}$ and $\mathcal{V}_b \le \mathcal{V}/b$; substituting these two inequalities into (11) proves the claim. The last term comes from the additive Bernstein correction and is retained when displaying the dependence on b.