TRAPEZE: MINIMAX-OPTIMAL EPISODE-WISE SAFE EXPLORATION IN TABULAR CMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online learning in finite-horizon tabular constrained Markov decision processes, where every adaptively deployed randomized policy must have conditional expected episode cost at most a prescribed budget. The learner is given a baseline policy with known cost value and safety slack $\kappa > 0$. We introduce TRAPEZE, which separates model-based policy construction from deployment certification. At each replan, scalar optimistic planning over a finite grid of dual multipliers and a deficit-routing rule produce a proposed episode-level distribution over policies. While this proposal remains frozen, TRAPEZE directly certifies its cost from subsequent whole-trajectory returns and deploys the largest certified episode-level mixture of the proposal and the baseline. A profile-based Bernstein analysis controls the adaptively selected continuation values without coordinate-wise transition confidence. For K episodes, with probability at least $1-\delta$, every deployed policy is feasible and the regret is $\widetilde{O}((1+H/\kappa)(\sqrt{SAH^3K}+SAH^2))$. We also prove the matching lower bound $\Omega(\min\{HK,(1+H/\kappa)\sqrt{SAH^3K}\})$ under the same baseline information and episode-wise safety requirement. These results establish minimax optimality up to logarithmic and lower-order terms and show that the inverse safety slack multiplies the full tabular exploration difficulty. Safety concerns conditional expected episode cost, not the cost of every realized trajectory.

1 INTRODUCTION

Constrained Markov decision processes (CMDPs) formalize sequential learning when reward is valuable only subject to a resource or risk budget (Altman, 1999). In the online setting, the transition law, rewards, and costs are initially unknown, so data must be gathered by deploying policies. A known safe baseline represents an incumbent policy that can always be used, while the learning objective asks whether one can improve on it without abandoning a valid cost guarantee during exploration. This is a basic safe-exploration problem: learning and constraint enforcement must occur on the same trajectories. This requirement matters when each deployment has consequences and later conservative behavior cannot repair an unsafe decision. The central question is how the baseline slack κ, which finances exploration, changes the minimax learning rate.

Episode-wise safety is stronger than an aggregate violation criterion. Regret accumulates across episodes, but safety must be decided before each one. Under signed cumulative violation, an over-budget episode can be offset by a later conservative episode (Efroni et al., 2020; Müller et al., 2024). Feasibility of only a returned policy likewise says nothing about the policies deployed while learning (Liu et al., 2025). We instead require one high-probability event on which, conditional on all preceding observations, every adaptively deployed randomized policy has expected episode cost at most b (Liu et al., 2021; Bura et al., 2022; Yu et al., 2025). Realized trajectory costs may exceed b, but unsafe and conservative deployments cannot cancel.

The tabular setting isolates the statistical price of safety. Without constraints, time-inhomogeneous tabular MDPs have minimax regret $\widetilde{\Theta}(\min\{\sqrt{SAH^3K}, HK\})$ (Domingues et al., 2021; Zhang et al., 2024). Thus $\sqrt{SAH^3K}$ is the unconstrained benchmark once the trivial HK cap is inactive. Under episode-wise safety, OptPess-LP, DOPE, and DOPE+ progressively improve the dimension dependence, culminating in the DOPE+ rate $\widetilde{O}(\kappa^{-1}H^{5/2}S\sqrt{AK})$ (Liu et al., 2021; Bura et al., 2022;

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.

Yu et al., 2025). At $\kappa = \Theta(H)$, the three bounds exceed this benchmark by factors $S\sqrt{H}$, $\sqrt{SH}$, and $\sqrt{S}$, respectively.

No matching lower bound was known under the same safety and baseline assumptions, so it was unclear whether the remaining $\sqrt{S}$ factor or the inverse-slack dependence was unavoidable.

Under the weaker requirement that only the returned policy be feasible, Liu et al. (2025) obtains strict-output sample complexity $\widetilde{O}(SAH^5/(\varepsilon^2\zeta^2))$ and regret scale $\widetilde{O}(\zeta^{-1}\sqrt{SAH^5K})$, where ζ is the global Slater margin. The deployed policies may be unsafe, so cancellation can improve both regret and signed violation. Those results therefore do not resolve episode-wise safe exploration.

The role of the safety slack is easiest to see operationally. Before TRAPEZE has certified a newly proposed policy distribution, it may use that distribution with probability only on the order of κ/H and use the baseline otherwise. Collecting one trajectory under the proposal can therefore require H/κ physical episodes in expectation. Our lower bound shows that this loss is not an artifact of the algorithm. In the hard instances, choosing the wrong continuation makes the exploratory branch more costly. To remain safe, the learner must enter that branch less often, which directly reduces reward. Less baseline slack makes the same identification error more expensive.

The difficulty is structural, not merely a loose bonus. Solving an empirical CMDP couples the selected continuation policy to the same transition block that appears in its feasibility estimate. The resulting term $(\widehat{P}_h - P_h)V_{h+1}^{\pi(\widehat{P})}$ lacks the reverse-stage measurability required by profile concentration. Controlling it uniformly by coordinatewise transition confidence incurs the familiar $\sqrt{S}$ loss. Moreover, policy-dependent Bernstein objectives do not generally obey a Bellman principle. Scalar planning avoids this failure, but recovering separate reward and cost coordinates can amplify error by the dual range $O(H/\kappa)$. Count doubling freezes the data, but alone resolves neither dependence, dual amplification, nor deployment safety.

TRAPEZE separates these roles. At each replan, disjoint doubling blocks form a frozen empirical mosaic. For each fixed multiplier λ, scalar optimistic planning preserves Bellman structure, while a plain evaluation supplies model-based estimates of reward and safety utility for the selected policy. Deficit routing assigns the scalar optimism gap to the coordinate with the larger dual coefficient and prevents an additional factor proportional to λ in the coordinate errors. A cached dual recursion then returns an episode-level policy distribution ρ.

Fresh whole-trajectory costs collected while ρ is frozen directly certify its deployment. Mixing it with the known baseline enforces safety without a correlated model-based cost estimate. Weighted profiles control the adaptive continuation values with which each frozen block is reused, and predictable thinning transfers guarantees indexed by executions of ρ to physical episodes. Thus the same trajectories may later improve the model without compromising the current certificate.

Contributions.

- We give the first minimax-optimal regret guarantee for episode-wise safe tabular CMDP exploration with a known strictly safe baseline. With high probability, every deployed policy is feasible and strong regret is at most $\widetilde{O}((1 + H/\kappa)(\sqrt{SAH^3K} + SAH^2))$ (theorem 2).
- The algorithm and proof combine fixed-multiplier scalar planning, deficit routing, weighted profile concentration, direct trajectory certification, and predictable thinning. Together these tools remove the Bellman and coordinatewise $\sqrt{S}$ obstacles while avoiding an additional dual-range factor in coordinate estimation.
- We prove the lower bound $\Omega(\min\{HK, (1 + H/\kappa)\sqrt{SAH^3K}\})$ over the full slack range (theorem 9), matching the upper rate up to logarithmic and lower-order terms. Its safety-gated direct sum shows that inverse slack multiplies the full tabular exploration difficulty rather than adding a separate overhead.

2 RELATED WORK

Episode-wise safe tabular CMDPs. OptPess-LP (Liu et al., 2021), DOPE (Bura et al., 2022), and DOPE+ (Yu et al., 2025) study exactly our problem setting. They assume a known strictly feasible

baseline and require every deployed randomized policy to satisfy the expected-cost constraint. These methods improve the leading regret successively, but their analyses use coordinatewise transition control and retain the gap shown in table 1. In contrast, TRAPEZE separates model-based policy construction from model-free deployment certification.

Weaker and adjacent guarantees. Other online CMDP work studies signed cumulative, strong, or bounded constraint violation rather than feasibility of every deployed policy (Efroni et al., 2020; Müller et al., 2024; Stradi et al., 2025; Ghosh et al., 2024; Zuo et al., 2026; Kalagarla et al., 2025). These notions permit unsafe learning episodes and are not directly comparable to zero episode-wise violation. StepMix (Li et al., 2023) protects a return aligned with the learning objective. Safe linear CMDPs (Kitamura et al., 2025) and policy-prior intervention (Wendl et al., 2026) rely on different structure or complexity measures.

Profile-based minimax learning. Disjoint doubling blocks and profile concentration were developed for minimax online RL (Zhang et al., 2024) and adapted to online CMDPs with PAC output guarantees, including strict feasibility of the returned policy (Liu et al., 2025). TRAPEZE adds fixed-multiplier planning, coordinate routing, and direct trajectory certification. Its weighted profile lemma aggregates suffix values across repeated uses of each frozen block.

Table 1: Finite-horizon tabular results in our episode-wise safety setting. Logarithms and lower-order terms are omitted. The final column gives the multiplicative gap to the lower bound at $\kappa = \Theta(H)$.

Work	constraint guarantee	safe each episode?	upper/lower rate	leading gap
OptPess-LP (Liu et al., 2021)	zero expected-cost violation	Yes	$\widetilde{O}(\kappa^{-1}H^3S^{3/2}\sqrt{AK})$	$S\sqrt{H}$
DOPE (Bura et al., 2022)	zero expected-cost violation	Yes	$\widetilde{O}(\kappa^{-1}H^3S\sqrt{AK})$	$\sqrt{SH}$
DOPE+ (Yu et al., 2025)	zero expected-cost violation	Yes	$\widetilde{O}(\kappa^{-1}H^{5/2}S\sqrt{AK})$	$\sqrt{S}$
TRAPEZE (upper)	zero expected-cost violation	Yes	$\widetilde{O}((1+H/\kappa)\sqrt{SAH^3K})$	None†
This work (lower)	zero expected-cost violation	Required	$\Omega((1+H/\kappa)\sqrt{SAH^3K})^\dagger$	Reference

All upper rates are ordinary cumulative regret. Simultaneous safety makes them strong regret. † The exact lower bound is $\Omega(\min\{HK, (1+H/\kappa)\sqrt{SAH^3K}\})$ for $0 < \kappa < H$, $S \geq 8$, $A \geq 2$, and $H \geq 8$. The upper rate can also be clipped at HK, so the two bounds match up to logarithmic and lower-order terms.

3 PROBLEM FORMULATION

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r, c, p, b)$ be a time-inhomogeneous episodic CMDP, with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, known initial distribution p, and unknown P, r, c. On each visit to $x = (s, a, h)$, the learner observes a next state S', a reward $R \in [0, 1]$, and a cost $C \in [0, 1]$. Conditional on x, this triple is drawn afresh with distribution $P_h(\cdot \mid s, a)$ for S' and means $r_h(s, a)$ and $c_h(s, a)$ for R and C. The three observations from the same visit may be statistically dependent.

For a possibly randomized, nonstationary policy π and any signal $g = (g_h)_{h=1}^H$, define

$$V_{h,g}^{\pi}(s) := \mathbb{E}_{P,\pi}\left[\sum_{t=h}^{H} g_t(S_t, A_t) \,\middle|\, S_h = s\right]. \tag{1}$$

Because the present paper allows an initial distribution rather than a fixed state, we use the root-value shorthand $V_{1,g}^{\pi} := \mathbb{E}_{S_1\sim p}[V_{1,g}^{\pi}(S_1)]$. A policy may draw a persistent latent seed once per episode. This allows the policy class to represent episode-level mixtures of deterministic nonstationary Markov policies and yields the usual randomized-policy occupancy polytope. Let $\pi^\star \in \arg\max_\pi\{V_{1,r}^{\pi} : V_{1,c}^{\pi} \leq b\}$ and $V_{1,r}^{\star} := V_{1,r}^{\pi^\star}$. We assume $b \in (0, H)$.

Assumption 1 (Known strictly safe baseline). *The learner knows a policy π_0 and its true cost value $V_{1,c}^{\pi_0} < b$. Let $\kappa = b - V_{1,c}^{\pi_0}$ and $\Delta_0 = V_{1,r}^{\star} - V_{1,r}^{\pi_0} \le H$.*

Let $\mathcal{H}_{k-1}$ be the sigma-field generated by all actions, observations, past learner randomization, and internal algorithm state through episode $k-1$. It excludes the fresh branch and policy randomization used in episode k. At the beginning of that episode, the learner maps $\mathcal{H}_{k-1}$ to a randomized deployed policy μ_k. For this adaptive choice, let $V_{1,g}^{\mu_k|\mathcal{H}_{k-1}}$ denote the root value conditional on the observed history. We suppress the conditioning in later displays when it is clear. Our safety target is the single high-probability event

$$V_{1,c}^{\mu_k|\mathcal{H}_{k-1}} \le b \quad \text{for every } k \le K. \tag{2}$$

The random seed of μ_k is part of the policy. Thus equation (2) bounds its conditional expected cost before the episode coin is drawn, not each realized trajectory cost.

Our performance criterion is the conditional pseudo-regret

$$\mathrm{Regret}(K) := \sum_{k=1}^{K} \big(V_{1,r}^{\star} - V_{1,r}^{\mu_k|\mathcal{H}_{k-1}}\big).$$

On the simultaneous-safety event in equation (2), every summand is nonnegative by optimality of $\pi^\star$. Therefore, the same bound controls strong reward regret.

4 TRAPEZE: ALGORITHM AND DATA FLOW

4.1 OPERATIONAL OVERVIEW

At initialization, TRAPEZE calls REPLAN. This offline procedure forms records from all newly completed blocks, installs them into a stitched empirical model (or *mosaic*) $\widehat{\mathcal{M}}_e$, and freezes that model. It then constructs an episode-level distribution ρ_e over deterministic policies. We call ρ_e the *candidate*. The candidate need not be safe. REPLAN also resets the segment certificate. Segment e is the interval between two calls to REPLAN. The model $\widehat{\mathcal{M}}_e$ and candidate ρ_e remain fixed throughout the segment.

Before each physical episode, total costs already observed in the segment determine a certified mixture of ρ_e and the known safe baseline. If the candidate branch is selected, one component of ρ_e is followed for the whole trajectory. This *candidate execution* supplies step observations to the local model buffers and one total cost to the segment certificate. A baseline execution changes neither collection. If a candidate execution fills a local buffer, the entire trajectory is first completed using the current candidate ρ_e. After the trajectory ends, REPLAN installs all newly completed blocks, constructs the next candidate, and starts a new segment. Partially filled buffers retain their samples.

Completed blocks improve future candidates, while total costs collected under the fixed ρ_e certify its current deployment without concentrating a correlated model-based quantity such as $\widehat{P}\widehat{V}$.

4.2 CANDIDATE-ONLY BUFFERS AND FROZEN MODEL UPDATES

We use three online indices: physical episode k, candidate execution n, and frozen-model segment e. Within a replan, $t \in [T]$ indexes offline dual updates. Within segment e, the local counter j records completed candidate executions and hence certificate samples. The index t is reset at each replan, whereas j is reset when the resulting segment begins. Let $N \le K$ and M be the realized numbers of candidate executions and segments.

Each cell $x = (s, a, h)$ has one active buffer. Its successive capacity targets are the disjoint doubling blocks (Zhang et al., 2024; Liu et al., 2025)

$$\mathcal{B}_{x,\ell} = \{2^\ell, \dots, 2^{\ell+1} - 1\}, \qquad |\mathcal{B}_{x,\ell}| = 2^\ell, \tag{3}$$

where $(S'_{x,i}, R_{x,i}, C_{x,i})$ is the observation from the ith candidate visit to x. When the active buffer reaches its target $q = (x, \ell)$, it becomes a completed block and determines the record

$$\widehat{P}_q(s') = \frac{1}{n_q} \sum_{i \in \mathcal{B}_{x,\ell}} \mathbf{1}\{S'_{x,i} = s'\}, \qquad (\widehat{r}_q, \widehat{c}_q) = \frac{1}{n_q} \sum_{i \in \mathcal{B}_{x,\ell}} (R_{x,i}, C_{x,i}), \qquad \widehat{d}_q = 1 - \widehat{c}_q, \tag{4}$$

with $n_q = 2^\ell$. At the following replan, the record becomes active and an empty buffer of size $2n_q$ is opened. We call q installed until its successor fills, so it is charged at most $2n_q$ times. The mosaic $\widehat{\mathcal{M}}_e$ contains the latest installed record at each cell. Let $q_e(x)$ denote its installed block, with $q_e(x) = \perp$ before any record is installed. We suppress e below and write $(\widehat{P}_h(\cdot \mid s,a), \widehat{r}_h(s,a), \widehat{c}_h(s,a))$ for the installed record at $x = (s,a,h)$. For $q_e(x) = \perp$, choose a fixed $s_{\mathrm{ref}} \in \mathcal{S}$, use $(\delta_{s_{\mathrm{ref}}}, 0, 0)$, and set the optimistic action value to the remaining horizon. Records are installed only after a trajectory ends, and partial buffers persist. The realized number M of segments satisfies

$$M \le M_K, \qquad M_K := 1 + SAH\lceil \log_2(K+1) \rceil. \tag{5}$$

The width of a size-n_q record decreases with n_q, while its charge lasts for at most $2n_q$ visits. Lemma 6 sums these shrinking errors. Doubling also gives low switching, but by itself no safety certificate.

Formal data separation. Let $\mathcal{D}_e^-$ contain the completed blocks available when segment e begins. The mosaic $\widehat{\mathcal{M}}_e$, its optimistic and plain values, the dual path, and ρ_e are functions of $\mathcal{D}_e^-$ and remain frozen while fresh total trajectory costs update the certificate. Step observations enter active buffers but cannot change the candidate they certify. For the ith candidate execution in segment e, define $Y_{e,i} := \sum_{h=1}^H C_{e,i,h}$ and let $\mathcal{F}_{e,i}^-$ denote the history before its latent component is drawn. Then $\mathbb{E}[Y_{e,i} \mid \mathcal{F}_{e,i}^-] = V_{1,c}^{\rho_e}$. This identity holds under adaptive stopping without a global sample split.

4.3 Replanning by scalar optimization and deficit routing

Let $d_h = 1 - c_h$ be safety utility, so $V_{1,d}^\pi = H - V_{1,c}^\pi$. For parameters $\lambda_{\max}, T$, set $\varepsilon_1 = \lambda_{\max}/T$ and $\Lambda = \{0, \varepsilon_1, \dots, T\varepsilon_1 = \lambda_{\max}\}$. All argmax and rounding ties are deterministic. Throughout this subsection, the planner, values, routed surrogates, and dual path depend on the current segment e. We suppress that index.

Fixed-multiplier scalar planning. A direct empirical constrained optimizer makes the selected suffix depend on the same transition block through the global budget, so $(\widehat{P}_h - P_h)V_{h+1}^{\pi(\widehat{P})}$ lacks reverse-stage measurability. Fixing λ restores ordinary backward dynamic programming. Routing later recovers separate reward and cost surrogates. For the frozen model and each $\lambda \in \Lambda$, define

$$f_{\lambda,h} = \frac{r_h + \lambda d_h}{1+\lambda} \in [0,1], \qquad \widehat{f}_{\lambda,q} = \frac{\widehat{r}_q + \lambda \widehat{d}_q}{1+\lambda}. \tag{6}$$

We then run an ordinary empirical-Bernstein optimistic dynamic program of the monotonic value-propagation (MVP) form. Let $V_{h,f_\lambda}^\star(s) = \max_\pi V_{h,f_\lambda}^\pi(s)$. This is the unconstrained scalar-MDP optimum for the fixed multiplier λ, not the CMDP-optimal policy $\pi^\star$. For the exact empirical-Bernstein bonus B_q in equation (31), start from $\overline{V}_{H+1,f_\lambda} = 0$ and, for an installed record $q = q_e(s,a,h)$, compute

$$\overline{Q}_{h,f_\lambda}(s,a) = \min\{H - h + 1, \widehat{f}_{\lambda,q} + \langle \widehat{P}_q, \overline{V}_{h+1,f_\lambda} \rangle + B_q(\overline{V}_{h+1,f_\lambda})\}. \tag{7}$$

Choose $a_{\lambda,h}^\star(s) \in \arg\max_a \overline{Q}_{h,f_\lambda}(s,a)$ and set $\overline{V}_{h,f_\lambda}(s) = \overline{Q}_{h,f_\lambda}(s, a_{\lambda,h}^\star(s))$ and $\pi_{\lambda,h}(\cdot \mid s) = \delta_{a_{\lambda,h}^\star(s)}$. For an unobserved cell set $\overline{Q} = H - h + 1$. Ties are deterministic. The root value is averaged over p, and $\overline{L}_\lambda = (1+\lambda)\overline{V}_{1,f_\lambda}$.

Separately, plain-evaluate the selected policy in the same frozen model, with no bonus: for $g \in \{r, d\}$,

$$\widehat{V}_{h,g}^{\pi_\lambda} = \pi_{\lambda,h}\big(\widehat{g}_h + \widehat{P}_h \widehat{V}_{h+1,g}^{\pi_\lambda}\big), \qquad \widehat{V}_{H+1,g}^{\pi_\lambda} = 0. \tag{8}$$

Average over p to obtain $\widehat{V}_{1,r}^{\pi_\lambda}, \widehat{V}_{1,d}^{\pi_\lambda} \in [0,H]$. They need not be coordinatewise optimistic.

From scalar optimism to coordinate surrogates. The dual update needs separate reward and cost coordinates, whereas lemma 3 supplies only one inequality for their weighted sum. Let $\widehat{L}_\lambda := \widehat{V}_{1,r}^{\pi_\lambda} + \lambda \widehat{V}_{1,d}^{\pi_\lambda}$ and $\Delta_\lambda := \left[\overline{L}_\lambda - \widehat{L}_\lambda\right]_+$. Thus Δ_λ is exactly the weighted optimism missing from the plain coordinate estimates. Route it as

$$(\widetilde{v}_{r,\lambda}, \widetilde{v}_{d,\lambda}) = \begin{cases} (\widehat{V}_{1,r}^{\pi_\lambda} + \Delta_\lambda, \widehat{V}_{1,d}^{\pi_\lambda}), & \lambda \le 1, \\ (\widehat{V}_{1,r}^{\pi_\lambda}, \widehat{V}_{1,d}^{\pi_\lambda} + \Delta_\lambda/\lambda), & \lambda > 1. \end{cases} \tag{9}$$

Let $\widetilde{v}_{c,\lambda} = H - \widetilde{v}_{d,\lambda}$. By construction,

$$\begin{aligned}\widetilde{v}_{r,\lambda} + \lambda \widetilde{v}_{d,\lambda} &= \widehat{L}_\lambda + \Delta_\lambda = \max\{\widehat{L}_\lambda, \overline{L}_\lambda\} \\ &\geq \max_\pi \{V^\pi_{1,r} + \lambda V^\pi_{1,d}\}.\end{aligned} \tag{10}$$

The last inequality is lemma 3. The rule is the minimum-ℓ_1 nonnegative correction satisfying $x_r + \lambda x_d \geq \Delta_\lambda$: it routes through the larger coefficient, with ties to reward. This avoids $\lambda_{\max}$ amplification. Since $d = 1 - c$, utility routing lowers the cost surrogate. These are algebraic surrogates, not confidence bounds or safety certificates.

Within the frozen segment, choose a planning budget $b_{\text{plan}} \in [0, H]$ and run the rounded projected multiplier update

$$\lambda_{t+1} = \mathcal{R}_\Lambda\left(\Pi_{[0,\lambda_{\max}]}[\lambda_t + \eta(\widetilde{v}_{c,\lambda_t} - b_{\text{plan}})]\right), \tag{11}$$

where $\lambda_1 = 0$ and $\mathcal{R}_\Lambda$ rounds to the nearest grid point, using the fixed tie rule and clipping to the endpoints. Every grid-point tuple $(\pi_\lambda, \overline{V}_{1,f_\lambda}, \widehat{V}^{\pi_\lambda}_{1,r}, \widehat{V}^{\pi_\lambda}_{1,d}, \widetilde{v}_{r,\lambda}, \widetilde{v}_{c,\lambda})$ is computed once per replan and cached, so the T dual steps collect no data. They return the latent policy mixture $\rho_e = T^{-1}\sum_{t=1}^T \delta_{\pi_{\lambda_t}}$: at the beginning of a candidate execution, sample one component and follow its policy for the whole trajectory. For later analysis, write $p_e(u) = 1/T$ and $\pi_{e,u} = \pi_{\lambda_u}$, retaining repeated grid points as distinct components. Values are affine in this latent distribution, not in pointwise averages of action probabilities. The main algorithm uses $b_{\text{plan}} = b$. It never estimates an empirical-CMDP multiplier: the baseline slack implies $\lambda^\star_b \leq H/\kappa$ by lemma 11, so $\lambda_{\max} = 2H/\kappa$ contains a valid dual comparator.

Algorithm 1 summarizes candidate construction, certification, and the exact order of model updates. The trajectory certificate used in its physical-episode loop is defined in the next subsection.

Algorithm 1: TRAPEZE

Input: K, b, δ, known $(\pi_0, V^{\pi_0}_{1,c})$, and $(\lambda_{\max}, \Lambda, T, \eta)$ from theorem 7
Data: Persistent local buffers and segment cost lists $\mathcal{Y}_e$
Procedure `Replan` $(e, \widehat{\mathcal{M}}, \mathcal{Q}; b)$ **:**
  foreach *newly completed block* $q \in \mathcal{Q}$ **do**
    form its record by equation (4); replace that cell's mosaic record and open its empty doubled successor buffer
  Freeze the updated mosaic as $\widehat{\mathcal{M}}^+$
  foreach $\lambda \in \Lambda$ **do**
    form $\widehat{f}_\lambda$ and compute $(\pi_\lambda, \overline{L}_\lambda)$ by equations (6) and (7); plain-evaluate by equation (8), then compute Δ_λ and $(\widetilde{v}_{r,\lambda}, \widetilde{v}_{d,\lambda})$ by equation (9), and set $\widetilde{v}_{c,\lambda} \leftarrow H - \widetilde{v}_{d,\lambda}$
  Set $b_{\text{plan}} \leftarrow b$ and $\lambda_1 \leftarrow 0$, then run the T cached updates in equation (11)
  Cache $\pi_{e,u} \leftarrow \pi_{\lambda_u}$ for $u \in [T]$ and set $\rho_e \leftarrow T^{-1}\sum_{u=1}^T \delta_{\pi_{e,u}}$
  Reset $\mathcal{Y}_e \leftarrow \varnothing$ and $j \leftarrow 0$; return $(\widehat{\mathcal{M}}^+, \rho_e)$
Initialize $\widehat{\mathcal{M}}^{\text{def}}$ as the placeholder mosaic specified above and create size-one local buffers; set $e \leftarrow 1, n \leftarrow 0$, and $(\widehat{\mathcal{M}}_1, \rho_1) \leftarrow$ `Replan`$(1, \widehat{\mathcal{M}}^{\text{def}}, \varnothing; b)$
for $k = 1, \ldots, K$ **do**
  Compute $(U^c_{e,j}, \alpha_{e,j})$ by equations (12) and (13) and draw $Z_k \sim \text{Ber}(\alpha_{e,j})$
  if $Z_k = 0$ **then**
    execute π_0; leave $n, j, \mathcal{Y}_e$, and all buffers unchanged
  else
    Draw $u \sim \text{Unif}([T])$, execute $\pi_{e,u}$ for the episode, and observe $(S_h, A_h, R_h, C_h, S_{h+1})_{h=1}^H$
    Set $n \leftarrow n+1$, $j \leftarrow j+1$, and $Y_{e,j} \leftarrow \sum_h C_h$; append $Y_{e,j}$ to $\mathcal{Y}_e$ and every step tuple to its cell buffer
    Let $\mathcal{Q}$ be the set of blocks completed by this trajectory
    if $\mathcal{Q} \neq \varnothing$ **then**
      set $e \leftarrow e+1$ and $(\widehat{\mathcal{M}}_e, \rho_e) \leftarrow$ `Replan`$(e, \widehat{\mathcal{M}}_{e-1}, \mathcal{Q}; b)$; retain all unfinished buffers

4.4 TRAJECTORY CERTIFICATION AND SAFE DEPLOYMENT

In segment e, recall that ρ_e is frozen and $Y_{e,i} = \sum_{h=1}^H C_{e,i,h} \in [0, H]$ is the ith candidate-execution cost. Here j is the number completed in the current segment, and n_e denotes its final value when

segment e ends. Before the next branch coin, use the empirical mean, radius, and cost certificate

$$\begin{aligned} \overline{Y}_{e,j} &:= j^{-1}\textstyle\sum_{i=1}^{j} Y_{e,i}, \qquad \beta_{e,j} := H\sqrt{2\log(4(e+1)^2(j+1)^2/\delta)/j} \quad (j \geq 1), \\ \overline{Y}_{e,0} &:= H, \quad \beta_{e,0} := 0, \qquad U^c_{e,j}(\rho_e) := \min\{H, \overline{Y}_{e,j} + \beta_{e,j}\}. \end{aligned} \tag{12}$$

The deployment weight and policy have the equivalent one-line form

$$\alpha_{e,j} := \frac{\kappa}{\max\{\kappa, U^c_{e,j}(\rho_e) - V^{\pi_0}_{1,c}\}}, \qquad \mu_k := \alpha_{e,j}\rho_e + (1-\alpha_{e,j})\pi_0. \tag{13}$$

The denominator lies in $[\kappa, H]$, so $\kappa/H \leq \alpha_{e,j} \leq 1$. Whenever $V^{\rho_e}_{1,c} \leq U^c_{e,j}(\rho_e)$, this choice is safe because the affinity of policy values gives $V^{\mu_k|\mathcal{H}_{k-1}}_{1,c} \leq \alpha_{e,j}U^c_{e,j} + (1-\alpha_{e,j})(b-\kappa) \leq b$. A new segment starts conservatively with $U^c_{e,0} = H$, but always has $\alpha_{e,j} \geq \kappa/H$. It may end after one candidate execution. Section 5.4 permits arbitrary adaptive lengths. Algorithm 1 makes the boundary order explicit: the triggering trajectory remains in the old segment, and records are installed only afterward, before replanning and resetting the certificate.

5 Analysis and Minimax Upper Bound

5.1 Main Guarantee

We first state the end-to-end result and then explain its three supporting estimates.

Theorem 2 (Episode-wise safe minimax regret)**.** *Under assumption 1, for all $K \geq 1$ and $\delta \in (0,1)$,* TRAPEZE*, with confidence parameter $\delta/3$ in each of its candidate-learner, certificate, and thinning events, satisfies with probability at least $1-\delta$,*

$$\begin{aligned} V^{\mu_k|\mathcal{H}_{k-1}}_{1,c} &\leq b \quad \textit{for every } k \leq K, \\ \mathrm{Regret}(K) &\leq \widetilde{O}\left[\left(1+\frac{H}{\kappa}\right)\left(\sqrt{SAH^3K} + SAH^2\right)\right]. \end{aligned} \tag{14}$$

Every reward gap is then nonnegative, so the same rate controls strong reward regret. The bound may be clipped at HK and, in the tabular regime of theorem 9, is minimax optimal for every $0 < \kappa < H$, up to logarithms and lower-order terms. The regret is sublinear in HK once $K \gtrsim SAH^3/\kappa^2$.

5.2 Scalar Optimism and Coordinate Routing

The optimistic planner needs to control only the scalarized signal, not reward and cost separately.

Lemma 3 (Gridwise scalar optimism)**.** *Suppose $|\Lambda| \leq K^2+1$. With probability at least $1-\delta$, simultaneously for every frozen segment and $\lambda \in \Lambda$,*

$$\overline{L}_\lambda \geq \max_\pi\{V^\pi_{1,r} + \lambda V^\pi_{1,d}\}, \qquad \overline{V}_{1,f_\lambda} \geq V^{\pi_\lambda}_{1,f_\lambda}.$$

The proof concentrates at a fixed true optimal suffix and uses pathwise monotonicity for backward induction. It never concentrates directly on the correlated $\overline{V}$. The complete argument appears in lemma 14 and appendix C.2.

On the event in lemma 3, for a component π_λ write $\xi^r_\lambda = \widehat{V}^{\pi_\lambda}_{1,r} - V^{\pi_\lambda}_{1,r}$, $\xi^d_\lambda = \widehat{V}^{\pi_\lambda}_{1,d} - V^{\pi_\lambda}_{1,d}$, and $z_\lambda = \overline{V}_{1,f_\lambda} - V^{\pi_\lambda}_{1,f_\lambda} \geq 0$.

Lemma 4 (Coordinate routing)**.** *On the event in lemma 3, for every $\lambda \geq 0$,*

$$\left[\widetilde{v}_{r,\lambda} - V^{\pi_\lambda}_{1,r}\right]_+ + \left[\widetilde{v}_{d,\lambda} - V^{\pi_\lambda}_{1,d}\right]_+ \leq 2(|\xi^r_\lambda| + |\xi^d_\lambda| + z_\lambda).$$

Thus neither coordinate error contains a factor λ.

The reason is the elementary ratio $(1+\lambda)/\max\{1,\lambda\} \leq 2$: routing the scalar deficit through its larger coefficient prevents the statistical error from being multiplied by the dual range $\lambda_{\max}$. The complete case analysis is in appendix B.

5.3 WEIGHTED PROFILES AND CANDIDATE STRONG METRICS

For a local block $q = (s, a, h, \ell)$, write $P_q := P_h(\cdot \mid s, a)$. Let $m_q \le 2n_q$ count later candidate visits while that block is installed, and let ω_q be the empirical distribution of the admissible continuation-value vectors used on those visits. The fixed-profile construction in lemma 17 makes each such suffix measurable before block q is revealed.

Lemma 5 (Weighted convexified profile concentration)**.** *Suppose $|\Lambda| \le K^2 + 1$. With probability at least $1 - \delta$, simultaneously for all admissible profiles, $0 \le m_q \le 2n_q$, and empirical distributions ω_q supported on the corresponding family $\mathcal{F}_q^{\mathcal{I}}$ from lemma 17,*

$$\sum_q m_q \mathbb{E}_{\omega_q} \langle \widehat{P}_q - P_q, W \rangle \le c \left\{ \sqrt{\Gamma_K \sum_q m_q \mathbb{E}_{\omega_q} \operatorname{Var}_{P_q}(W)} + H\Gamma_K \right\}, \tag{15}$$

where $\Gamma_K = O(SAH \log^2(SAHK/\delta))$. The negative part obeys the same bound. The precise admissible families are in appendix C.3.

The proof first establishes one event for every deterministic blockwise suffix selection. Only after this simultaneous event holds do we average it under the realized ω_q. This retains the variance of coherent component policies. An independent maximum at every block would instead create a non-policy mosaic and break the total-variance recursion.

Let $e(n)$ be the segment active at candidate execution n, and abbreviate $\rho_n := \rho_{e(n)}$, $p_n := p_{e(n)}$, and $\pi_{n,u} := \pi_{e(n),u}$. For component u, let $\xi^r_{n,u}, \xi^d_{n,u}, z_{n,u}$ be the three errors above. The component actually executed is $J_n \sim p_n$.

Lemma 6 (Latent-grid width bound)**.** *With probability at least $1 - \delta$, uniformly for every stopping time $N \le K$,*

$$\sum_{n=1}^{N} \mathbb{E}_{J \sim p_n}(|\xi^r_{n,J}| + |\xi^d_{n,J}| + z_{n,J}) \le \mathsf{B}_N(\delta) := c_0 \left[\sqrt{SAH^3 N}\, L_K + SAH^2 L_K^2 \right], \tag{16}$$

where $L_K = \log(c_1 SAHK/\delta)$ and c_0, c_1 are universal constants.

Here a block of size n_q has width of order $n_q^{-1/2}$ and is charged at most $2n_q$ times. Lemma 5 sums these shrinking errors without a coordinatewise union over next states. Full Bellman and variance reductions are proved in appendix C.

Define the candidate-clock strong metrics

$$\mathrm{SRegret}(N) = \sum_{n \le N} \left[V^\star_{1,r} - V^{\rho_n}_{1,r} \right]_+, \qquad \mathrm{SCV}(N) = \sum_{n \le N} \left[V^{\rho_n}_{1,c} - b \right]_+.$$

Theorem 7 (Candidate strong metrics)**.** *Under assumption 1, choose $\lambda_{\max} = 2H/\kappa$, $T = \max\{1, \lceil K/(SAH) \rceil\}$, $\varepsilon_1 = \lambda_{\max}/T$, and $\eta = \lambda_{\max}/(3H\sqrt{T})$, and run equation (11) with $b_{\mathrm{plan}} = b$. With probability at least $1 - \delta$, for every $N \le K$,*

$$\mathrm{SRegret}(N) \le \widetilde{O}((1 + H/\kappa)\mathsf{B}_N), \tag{17}$$

$$\mathrm{SCV}(N) \le \widetilde{O}(\mathsf{B}_N). \tag{18}$$

There are at most M_K segments.

Proof sketch. Lemmas 10 and 11 convert the routed saddle error into per-candidate reward and constraint bounds. Lemma 4 removes dual amplification, while lemmas 5, 6 and 17 sum the statistical widths over every candidate prefix. The complete reductions and parameter substitution are in appendices A and C. □

5.4 CERTIFICATION AND PHYSICAL-TIME TRANSFER

The frozen empirical mosaic constructs future candidates but never certifies the current one. The only deployment bound is $U^c_{e,j}$, formed from total trajectory costs collected while ρ_e is fixed. For the pre-component history $\mathcal{F}^-_{e,i}$ formalized in appendix D, $\mathbb{E}[Y_{e,i} \mid \mathcal{F}^-_{e,i}] = V^{\rho_e}_{1,c}$. The known baseline value is not re-estimated.

Lemma 8 (Anytime trajectory certificate). *With probability at least $1-\delta$, for every created segment e and attained sample size $j \geq 1$,*

$$V_{1,c}^{\rho_e} \leq U_{e,j}^{c}(\rho_e) \leq V_{1,c}^{\rho_e} + 2\beta_{e,j}. \tag{19}$$

At $j = 0$ its excess is at most H. If the first M segments contain n_e candidate executions and $N = \sum_e n_e$, then on the same event

$$\sum_{e=1}^{M} \sum_{j=0}^{n_e-1} \left(U_{e,j}^{c}(\rho_e) - V_{1,c}^{\rho_e}\right) \leq \widetilde{O}\big(HM + H\sqrt{MN}\big). \tag{20}$$

Conditional Hoeffding bounds, summation within segments, and Cauchy's inequality across segments prove the result. Appendix D gives the adaptive filtration and stopping details. Thus lemmas 8 and 24 imply episode-wise safety without a uniform transition confidence set. Substituting equation (5) into equation (20) makes both certification terms at most $\widetilde{O}(\mathsf{B}_N)$.

If $\mathfrak{R}_N$ and $\mathfrak{C}_N$ bound the candidate-clock strong metrics and there are M segments, the predictable-thinning transfer theorem 23 yields

$$\begin{aligned}\mathrm{Regret}(K) \leq \widetilde{O}\Big(&\mathfrak{R}_N + H\sqrt{K} \\ &+ \frac{\Delta_0}{\kappa}[\mathfrak{C}_N + H\sqrt{MN} + HM + H\sqrt{K}]\Big).\end{aligned}$$

Candidate executions are charged to $\mathfrak{R}_N$. Baseline fallback is charged to $\mathfrak{C}_N$ and certificate width, producing only one inverse-slack factor. The proof is in appendix D.

Proof sketch of theorem 2. For $N \leq K$, equation (5) gives $H\sqrt{M_K N} = \widetilde{O}(\sqrt{SAH^3K})$ and $HM_K = \widetilde{O}(SAH^2)$, while $H\sqrt{K} \leq \sqrt{SAH^3K}$. Insert these and theorem 7 into theorem 23, then use $\Delta_0 \leq H$. This proves equation (14). Safety follows from lemmas 8 and 24. Stopping-time uniformity is supplied by lemma 6 and appendices C and D. □

6 Matching lower bound and minimax optimality

Theorem 9 (Matching minimax lower bound). *Fix $S \geq 8$, $A \geq 2$, $H \geq 8$, $K \geq 1$, and $0 < \kappa < H$. Consider any learner that, for every budget $b \in (0, H)$ and every instance on which it is given a baseline with known cost $b - \kappa$, guarantees*

$$\mathbb{P}_\nu\Big(V_{1,c}^{\mu_k|\mathcal{H}_{k-1}} \leq b \text{ for all } k \leq K\Big) \geq 0.9.$$

There are a budget $b \in (0, H)$ and an S-state, A-action, H-horizon CMDP with rewards and costs in $[0,1]$ for which, for a universal constant $c > 0$,

$$\mathbb{P}_\nu\left(\mathrm{Regret}(K) \geq c\min\left\{HK, \left(1 + \frac{H}{\kappa}\right)\sqrt{SAH^3K}\right\},\ V_{1,r}^{\star} - V_{1,r}^{\mu_k|\mathcal{H}_{k-1}} \geq 0\ \forall k \leq K\right) \geq 0.3.$$

Hence the same lower bound holds for strong reward regret.

Proof sketch. A product-prior KL direct sum over $\Theta(S)$ contexts and $\Theta(AH)$ arms gives the small-slack rate. A zero-cost tabular instance gives the large-slack rate. See appendix E. □

The bounds match up to logarithms and lower-order terms: both clip at HK for $K < SAH$, and $SAH^2 \leq \sqrt{SAH^3K}$ for $K \geq SAH$.

7 Conclusion

TRAPEZE separates model-based candidate learning from direct trajectory certification and attains the minimax rate. Thus inverse safety slack multiplies the full tabular exploration difficulty. The guarantee concerns conditional expected cost and a known safe baseline. Multiple constraints, unknown baseline values, and function approximation remain open.

AI USE STATEMENT

Generative AI tools were used as writing assistants and for auxiliary checks of proof presentation and consistency. The authors independently verified the citations, mathematical claims, and proofs and take responsibility for the final manuscript.

ETHICS STATEMENT

Our guarantee concerns a policy's conditional expected episode cost. It does not prevent an individual trajectory from incurring a high cost. Applications with pathwise or irreversible harms require additional safeguards and domain-specific validation beyond the theory developed here.

REPRODUCIBILITY STATEMENT

The main text specifies all assumptions, parameters, and algorithmic steps. The appendix gives complete proofs, including the concentration filtrations, profile construction, transfer argument, and lower-bound instance. This theoretical paper uses no datasets or empirical results.

REFERENCES

Eitan Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1999.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/076a93fd42aa85f5ccee921a01d77dd5-Abstract-Conference.html.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 578–598, 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020. URL https://arxiv.org/abs/2003.02189.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free RL. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1054–1062. PMLR, 2024. URL https://proceedings.mlr.press/v238/ghosh24a.html.

Krishna C. Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A safe bayesian learning algorithm for constrained MDPs with bounded constraint violation. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2782–2790, 2025. URL https://proceedings.mlr.press/v258/kalagarla25a.html.

Toshinori Kitamura, Arnob Ghosh, Tadashi Kozuno, Wataru Kumagai, Kazumi Kasaura, Kenta Hoshino, Yohei Hosoe, and Yutaka Matsuo. Provably efficient RL under episode-wise safety in constrained MDPs with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/435e8fbbfc2c6072d4f3a5cb6e56a39a-Abstract-Conference.html.

Donghao Li, Ruiquan Huang, Cong Shen, and Jing Yang. Near-optimal conservative exploration in reinforcement learning under episode-wise constraints. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19527–19564, 2023. URL https://proceedings.mlr.press/v202/li23k.html.

Chang Liu, Yunfan Li, and Lin F. Yang. Near-optimal sample complexity for online constrained MDPs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/445500190fe42e9f58f65199d2bd70d1-Abstract-Conference.html.

Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8ec2ba5e96ec1c050bc631abda80f269-Abstract.html.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36605–36653, 2024. URL https://proceedings.mlr.press/v235/muller24b.html.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2827–2835. PMLR, 2021. URL https://proceedings.mlr.press/v130/pacchiano21a.html.

Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *Journal of Machine Learning Research*, 26(170):1–57, 2025. URL https://jmlr.org/papers/v26/24-0267.html.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained MDPs via policy optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8105a80cea247fa6f8441a8069b65f2a-Abstract-Conference.html.

Manuel Wendl, Yarden As, Manish Prajapat, Anton Pollak, Stelian Coros, and Andreas Krause. Safe exploration via policy priors. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cddca48052589b6ab24f009e52f0d923-Abstract-Conference.html.

Kihyun Yu, Duksang Lee, William Overman, and Dabeen Lee. Improved regret bound for safe reinforcement learning via tighter cost pessimism and reward optimism. *Reinforcement Learning Journal*, 6:493–546, 2025. URL https://rlj.cs.umass.edu/2025/papers/Paper41.html.

Zihan Zhang, Yuxin Chen, Jason D. Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5213–5219, 2024. URL https://proceedings.mlr.press/v247/zhang24a.html.

Qian Zuo, Zhiyong Wang, and Fengxiang He. Near-constant strong violation and last-iterate convergence for online CMDPs via decaying safety margins. *arXiv preprint arXiv:2602.10917*, 2026. doi: 10.48550/arXiv.2602.10917. URL https://arxiv.org/abs/2602.10917.

A ROUTED SADDLE ANALYSIS

Throughout the appendix, $\mathcal{G}_{\mathrm{cand}}$ denotes the intersection of the candidate-learner concentration events, with failure probabilities allocated so that $\mathbb{P}(\mathcal{G}_{\mathrm{cand}}) \geq 1-\delta$.

For a segment mixture $\rho = T^{-1}\sum_t \delta_{\pi_t}$, write $\widetilde{v}_g(\rho) = T^{-1}\sum_t \widetilde{v}_{g,\lambda_t}$ and $V^\star_{1,r}(b') = \max_{\pi: V^\pi_{1,c}\leq b'} V^\pi_{1,r}$. The routed range is $\widetilde{v}_{c,\lambda} \in [-2H, H]$, so $|\widetilde{v}_{c,\lambda} - b'| \leq 3H$ for $b' \in [0, H]$.

Lemma 10 (Rounded dual saddle inequality)**.** *Suppose that ρ is generated by equation (11) with $b_{\mathrm{plan}} = b'$ and that the b'-feasible set is nonempty. Then on $\mathcal{G}_{\mathrm{cand}}$, simultaneously for every $\lambda \in [0, \lambda_{\max}]$,*

$$V^\star_{1,r}(b') - \widetilde{v}_r(\rho) + \lambda(\widetilde{v}_c(\rho) - b') \leq \varepsilon_T^{\mathrm{opt}},$$

where

$$\varepsilon_T^{\mathrm{opt}} \leq \frac{\lambda_{\max}^2}{2\eta T} + \frac{9\eta H^2}{2} + \frac{\varepsilon_1 \lambda_{\max} + \varepsilon_1^2/4}{2\eta}. \tag{21}$$

Lemma 11 (Multiplier bound from slack)**.** *If some policy has cost at most $b' - \xi$ for $\xi > 0$, finite-CMDP strong duality admits an optimal multiplier $\lambda^\star_{b'} \leq H/\xi$.*

Theorem 12 (Baseline-free strong metrics)**.** *Let $\tau = \mathsf{B}_K/K$, $b' = b + \tau$, $\lambda_{\max} = 2H/\tau$, $T = K^2$, $\varepsilon_1 = \lambda_{\max}/T$, and $\eta = \lambda_{\max}/(3H\sqrt{T})$. Run the candidate learner with planning budget $b_{\mathrm{plan}} = b'$. In the nontrivial regime $b + \tau \leq H$ and $\mathsf{B}_K < HK$, with probability at least $1-\delta$,*

$$\mathrm{SRegret}(N),\ \mathrm{SCV}(N) \leq \widetilde{O}(\mathsf{B}_N) \qquad \textit{for every } N \leq K.$$

This tuning does not use assumption 1.

Proof of lemma 10. Define

$$\mathcal{L}(\pi, \lambda) = V^\pi_{1,r} - \lambda(V^\pi_{1,c} - b') = V^\pi_{1,r} + \lambda(V^\pi_{1,d} - (H - b')).$$

Abbreviate $\pi_t := \pi_{\lambda_t}$ and $\widetilde{v}_{g,t} := \widetilde{v}_{g,\lambda_t}$ for $g \in \{r, c, d\}$. For the surrogate response at round t, equation (10) implies, for every policy π,

$$V^\pi_{1,r} + \lambda_t V^\pi_{1,d} \leq \widetilde{v}_{r,t} + \lambda_t \widetilde{v}_{d,t}. \tag{22}$$

Let $g_t = \widetilde{v}_{c,t} - b'$. Projection nonexpansiveness and a rounding error at most $\varepsilon_1/2$ give, for every $\lambda \in [0, \lambda_{\max}]$,

$$(\lambda - \lambda_t) g_t \leq \frac{(\lambda_t - \lambda)^2 - (\lambda_{t+1} - \lambda)^2}{2\eta} + \frac{\eta g_t^2}{2} + \frac{\varepsilon_1 \lambda_{\max} + \varepsilon_1^2/4}{2\eta}.$$

Summing and using $|g_t| \leq 3H$ yields equation (21). Define $\widetilde{v}_g(\rho) := T^{-1}\sum_{t=1}^T \widetilde{v}_{g,t}$ and $V^\star_{1,r}(b') := \max_{\pi: V^\pi_{1,c}\leq b'} V^\pi_{1,r}$. Let $\pi^\star_{b'}$ attain $V^\star_{1,r}(b')$. Its feasibility gives $V^{\pi^\star_{b'}}_{1,d} \geq H - b'$, so equation (22) implies

$$V^\star_{1,r}(b') - \widetilde{v}_{r,t} \leq \lambda_t(\widetilde{v}_{d,t} - V^{\pi^\star_{b'}}_{1,d}) \leq -\lambda_t g_t.$$

Consequently, for every comparator $\lambda \in [0, \lambda_{\max}]$,

$$V^\star_{1,r}(b') - \widetilde{v}_{r,t} + \lambda g_t \leq (\lambda - \lambda_t) g_t.$$

Averaging this inequality and the summed projected-gradient inequality gives, for the latent mixture ρ and every $\lambda \in [0, \lambda_{\max}]$,

$$V^\star_{1,r}(b') - \widetilde{v}_r(\rho) + \lambda(\widetilde{v}_c(\rho) - b') \leq \varepsilon_T^{\mathrm{opt}}. \tag{23}$$

This is the claimed saddle inequality. The preceding projected-gradient calculation also gives equation (21). □

Proof of lemma 11. The occupancy-measure formulation is a finite linear program, so strong duality gives

$$V^\star_{1,r}(b') = \min_{\lambda \geq 0} \max_{\pi} \{V^\pi_{1,r} - \lambda(V^\pi_{1,c} - b')\}.$$

If $V^{\pi_{\mathrm{s}}}_{1,c} \leq b' - \xi$, then at an optimal multiplier $\lambda^\star_{b'}$ the inner maximum is at least $V^{\pi_{\mathrm{s}}}_{1,r} + \lambda^\star_{b'}\xi \geq \lambda^\star_{b'}\xi$. Its value is $V^\star_{1,r}(b') \leq H$, proving $\lambda^\star_{b'} \leq H/\xi$. □

Proof of theorems 7 and 12. Index the candidate used in candidate execution n by ρ_n and define

$$\epsilon_{r,n} = \mathbb{E}_{J\sim p_n}\left[\widetilde{v}_{r,n,J} - V_{1,r}^{\pi_{n,J}}\right]_+,$$
$$\epsilon_{d,n} = \mathbb{E}_{J\sim p_n}\left[\widetilde{v}_{d,n,J} - V_{1,d}^{\pi_{n,J}}\right]_+.$$

By lemmas 4 and 6, simultaneously for every $N \le K$,

$$\sum_{n=1}^{N}(\epsilon_{r,n} + \epsilon_{d,n}) \le c\mathsf{B}_N. \tag{24}$$

Taking the comparator $\lambda = 0$ in equation (23), using $V_{1,r}^{\star}(b') \ge V_{1,r}^{\star}$, and translating the reward surrogate gives, for each n,

$$\left[V_{1,r}^{\star} - V_{1,r}^{\rho_n}\right]_+ \le \varepsilon_T^{\mathrm{opt}} + \epsilon_{r,n}. \tag{25}$$

Let $v_n = \left[V_{1,c}^{\rho_n} - b'\right]_+$. At an optimal multiplier $\lambda_{b'}^{\star}$, strong duality gives

$$V_{1,r}^{\star}(b') - V_{1,r}^{\rho_n} \ge -\lambda_{b'}^{\star} v_n. \tag{26}$$

If $v_n = 0$, the desired violation bound is immediate. If $v_n > 0$, the $\lambda = \lambda_{\max}$ instance of equation (23), the one-sided translations in the definitions of $\epsilon_{r,n}, \epsilon_{d,n}$, and equation (26) imply

$$(\lambda_{\max} - \lambda_{b'}^{\star})v_n \le \varepsilon_T^{\mathrm{opt}} + \epsilon_{r,n} + \lambda_{\max}\epsilon_{d,n}. \tag{27}$$

Thus equation (27) holds in both cases.

For either tuning, direct substitution in equation (21) gives

$$\varepsilon_T^{\mathrm{opt}} \le c\lambda_{\max} H/\sqrt{T}. \tag{28}$$

For the baseline-free tuning, the original optimum has slack τ for $b' = b + \tau$. Hence lemma 11 gives $\lambda_{b'}^{\star} \le H/\tau = \lambda_{\max}/2$. Summing equations (25) and (27), using equation (24), and converting from b' back to b gives

$$\mathrm{SRegret}(N) \le N\varepsilon_T^{\mathrm{opt}} + c\mathsf{B}_N, \tag{29}$$

$$\mathrm{SCV}(N) \le \frac{2N\varepsilon_T^{\mathrm{opt}}}{\lambda_{\max}} + \frac{2}{\lambda_{\max}}\sum_{n\le N}\epsilon_{r,n} + 2\sum_{n\le N}\epsilon_{d,n} + N\tau. \tag{30}$$

Here we used $[V_{1,c}^{\rho_n} - b]_+ \le v_n + \tau$. Since $T = K^2$, $\lambda_{\max} = 2HK/\mathsf{B}_K$, $\mathsf{B}_K \ge H\sqrt{K}$, and $(N/K)\mathsf{B}_K \le \mathsf{B}_N$,

$$N\varepsilon_T^{\mathrm{opt}} \le cH^2 N/\mathsf{B}_K \le c\mathsf{B}_N, \quad N\varepsilon_T^{\mathrm{opt}}/\lambda_{\max} \le cHN/K \le c\mathsf{B}_N, \quad N\tau \le \mathsf{B}_N.$$

Equations (29) and (30) prove theorem 12.

For the tuning in theorem 7, the baseline has slack κ, so lemma 11 gives $\lambda_b^{\star} \le H/\kappa = \lambda_{\max}/2$ and no budget relaxation is needed. If $K \ge SAH$, then $T \ge K/(SAH)$ and, for every $N \le K$,

$$N\varepsilon_T^{\mathrm{opt}} \le c\lambda_{\max}\sqrt{SAH^3 N}, \qquad N\varepsilon_T^{\mathrm{opt}}/\lambda_{\max} \le c\mathsf{B}_N.$$

If $K < SAH$, the trivial charges $\lambda_{\max}HN$ and HN are absorbed by $\lambda_{\max}SAH^2$ and SAH^2, respectively. Summing equations (25) and (27) and applying equation (24) proves equations (17) to (18) for all $N \le K$. □

B Proof of coordinate routing

Proof of lemma 4. Because

$$(1+\lambda)\overline{V}_{1,f_\lambda} = V_{1,r}^{\pi_\lambda} + \lambda V_{1,d}^{\pi_\lambda} + (1+\lambda)z_\lambda = \widehat{V}_{1,r}^{\pi_\lambda} + \lambda\widehat{V}_{1,d}^{\pi_\lambda} + (1+\lambda)z_\lambda - \xi_\lambda^r - \lambda\xi_\lambda^d,$$

we have $\Delta_\lambda = \left[(1+\lambda)z_\lambda - \xi_\lambda^r - \lambda\xi_\lambda^d\right]_+$. For the remainder of this proof, abbreviate $(a, e, z, \Delta) = (\xi_\lambda^r, \xi_\lambda^d, z_\lambda, \Delta_\lambda)$. If $\lambda \le 1$ and the argument of the positive part is nonpositive, then $[a+\Delta]_+ = a_+ \le |a|$. Otherwise the a terms cancel and

$$[a+\Delta]_+ = [(1+\lambda)z - \lambda e]_+ \le 2z + |e|.$$

Combining the two cases gives $\left[\widetilde{v}_{r,\lambda} - V_{1,r}^{\pi_\lambda}\right]_+ \le |a| + |e| + 2z$. The safety-utility coordinate is unchanged. If $\lambda > 1$, reward is unchanged. When the deficit is active, the e terms instead cancel:

$$[e + \Delta/\lambda]_+ = [(1+\lambda)z/\lambda - a/\lambda]_+ \le 2z + |a|.$$

When it is inactive, the expression is at most $|e|$. This proves lemma 4 with the stated bounds. □

C Proof of the weighted profile and width lemmas

C.1 Product-tape formalization

For every cell $x = (s, a, h)$, pre-generate an infinite visit-indexed tape of triples $(S'_{x,i}, R_{x,i}, C_{x,i})$ with the correct cell law. Entries with different indices (x, i) are mutually independent, and the initial-state draws are independent of all tapes and of one another. The three coordinates inside one entry may be dependent. Revealing the next unused entry on each visit is distributionally equivalent to standard fresh-noise episodic interaction. This representation permits reverse-stage exposure without an additional modeling assumption.

C.2 Gridwise scalar optimism

Put $\iota = c_\iota \log(c_\iota SAHK/\delta)$, $\bar{n}_q = n_q \vee 1$, and $\widehat{\mathrm{Var}}_q(W) = \widehat{P}_q W^2 - (\widehat{P}_q W)^2$. The exact bonus used in equation (7) is

$$\mathfrak{m}_q(W) = \max\left\{\frac{20}{3}\sqrt{\frac{\widehat{\mathrm{Var}}_q(W)\iota}{\bar{n}_q}}, \frac{400H\iota}{9\bar{n}_q}\right\}, \qquad B_q(W) = \mathfrak{m}_q(W) + 3\sqrt{\frac{\iota}{\bar{n}_q}}. \tag{31}$$

Here c_ι is a sufficiently large universal constant.

Lemma 13 (Fixed-target dyadic-block confidence)**.** *Assume $|\Lambda| \le K^2 + 1$ and use the product-tape construction in appendix C.1. With probability at least $1 - \delta$, simultaneously for every cell $x = (s, a, h)$, every dyadic block $q = (x, \ell)$ that can be installed in the first K candidate executions, and every $\lambda \in \Lambda$,*

$$\left|\widehat{f}_{\lambda,q} - f_{\lambda,h}(s,a)\right| \le 3\sqrt{\iota/n_q}, \tag{32}$$

$$|\widehat{g}_q - g_h(s,a)| \le 3\sqrt{\iota/n_q}, \qquad g \in \{r, d\}, \tag{33}$$

$$\widehat{P}_q V^\star_{h+1,f_\lambda} + \mathfrak{m}_q(V^\star_{h+1,f_\lambda}) \ge P_h V^\star_{h+1,f_\lambda}. \tag{34}$$

Here the target $V^\star_{h+1,f_\lambda}$ is a deterministic function of the true CMDP and λ, independent of block q.

Proof. For the scalar signal inequality, the observations

$$Z_i = \frac{R_{x,i} + \lambda(1 - C_{x,i})}{1 + \lambda} \in [0, 1]$$

in block q have mean $f_{\lambda,h}(s, a)$. The reward and utility observations also lie in $[0, 1]$ and have means $r_h(s, a)$ and $d_h(s, a)$. Two-sided Hoeffding inequalities therefore give equations (32) and (33) for each fixed label with failure probability at most $2\exp(-18\iota)$.

For the transition inequality, put $W = V^\star_{h+1,f_\lambda} \in [0, H]^S$. For $n_q \ge 2$, the empirical Bernstein inequality for bounded i.i.d. observations gives, with failure probability ε,

$$\begin{aligned}(P_h - \widehat{P}_q)W &\le \sqrt{\frac{2\widehat{\mathrm{Var}}_q(W)\log(2/\varepsilon)}{n_q - 1}} + \frac{7H\log(2/\varepsilon)}{3(n_q - 1)} \\ &\le 2\sqrt{\frac{\widehat{\mathrm{Var}}_q(W)\iota}{n_q}} + \frac{14H\iota}{3n_q},\end{aligned}$$

whenever $\log(2/\varepsilon) \le \iota$. If $x = \sqrt{\widehat{\mathrm{Var}}_q(W)\iota/n_q}$ and $y = H\iota/n_q$, then

$$\mathfrak{m}_q(W) = \max\{20x/3, 400y/9\} \ge \tfrac{1}{2}(20x/3 + 400y/9) \ge 2x + 14y/3,$$

which proves equation (34). When $n_q = 1$, the floor in $\mathfrak{m}_q$ is at least H. All claims are then deterministic after increasing c_ι if necessary.

There are at most $SAH\lceil\log_2(K + 1)\rceil(K^2 + 1)$ relevant pairs of blocks and grid points. Choose ε to be δ divided by a sufficiently large constant times this number and union bound all displayed inequalities. The definition $\iota = c_\iota \log(c_\iota SAHK/\delta)$, with a sufficiently large universal c_ι, dominates the resulting logarithm. This proves the simultaneous claim. □

Lemma 14 (Pathwise monotonicity of the MVP envelope)**.** *For every realized block q, put $G_q(W) = \widehat{P}_q W + \mathfrak{m}_q(W)$. Deterministically, $0 \le W \le W' \le H\mathbf{1}$ implies $G_q(W) \le G_q(W')$, simultaneously over all W, W'. In particular, W' may depend on $\widehat{P}_q$.*

Proof. Fix the realized empirical distribution $p = \widehat{P}_q$ and put

$$a = \frac{20}{3}\sqrt{\frac{\iota}{\bar{n}_q}}, \qquad \sigma_p(W) = \sqrt{\mathrm{Var}_p(W)}.$$

The linear branch of the envelope equals $a^2 H$, so $\mathfrak{m}_q(W) = a \max\{\sigma_p(W), aH\}$. For $0 \le W \le W' \le H\mathbf{1}$, let $D = W' - W$, $d = pD$, and abbreviate $\sigma = \sigma_p(W)$ and $\sigma' = \sigma_p(W')$. Since $D \ge 0$,

$$\begin{aligned}\sigma^2 - (\sigma')^2 &= -\mathrm{Var}_p(D) - 2\,\mathrm{Cov}_p(W, D) \\ &\le 2H\,d,\end{aligned}$$

where $\mathrm{Cov}_p(W, D) = p(WD) - (pW)d \ge -Hd$. If $\sigma \le aH$, then $\mathfrak{m}_q(W) = a^2 H \le \mathfrak{m}_q(W')$. Otherwise set $u = \max\{\sigma', aH\}$. The claim is again immediate when $u \ge \sigma$. In the remaining case $u < \sigma$, both u and σ are at least aH, and hence

$$\sigma - u = \frac{\sigma^2 - u^2}{\sigma + u} \le \frac{\sigma^2 - (\sigma')^2}{\sigma + u} \le \frac{2Hd}{2aH} = \frac{d}{a}.$$

Therefore $\mathfrak{m}_q(W) - \mathfrak{m}_q(W') \le d = p(W' - W)$, which implies $G_q(W') - G_q(W) \ge 0$. For each fixed realization of $\widehat{P}_q$, the argument is uniform over W, W' and uses no independence. □

Proof of lemma 3. For a frozen block define $G_q(W) = \widehat{P}_q W + \mathfrak{m}_q(W)$. By lemma 14, this operator is monotone for every realized empirical block. On the simultaneous event in lemma 13, for every block, grid point, and true optimal scalar suffix $V^\star_{h+1,f_\lambda}$,

$$f_{\lambda,h} + P_h V^\star_{h+1,f_\lambda} \le \widehat{f}_{\lambda,q} + 3\sqrt{\iota/\bar{n}_q} + G_q(V^\star_{h+1,f_\lambda}).$$

Assuming $\overline{V}_{h+1,f_\lambda} \ge V^\star_{h+1,f_\lambda}$, apply monotonicity followed by the confidence bound in equations (32) and (34). Monotonicity is pathwise at the data-dependent $\overline{V}$, while concentration is used only at the fixed $V^\star$. Clipping at $H - h + 1$ preserves the inequality because the true scalar Q-value is no larger than that threshold. Backward induction from $h = H$ proves

$$\overline{V}_{1,f_\lambda} \ge V^\star_{1,f_\lambda}.$$

Because $V^\star_{1,f_\lambda} = \max_\pi V^\pi_{1,f_\lambda}$, this same inequality immediately gives

$$\overline{V}_{1,f_\lambda} \ge V^\star_{1,f_\lambda} \ge V^{\pi_\lambda}_{1,f_\lambda},$$

even though π_λ is data-dependent. Thus $z_\lambda \ge 0$ without a second confidence argument. Multiplying the optimal-value inequality by $1 + \lambda$ gives the final inequality in equation (10). □

C.3 Profiles, measurability, and entropy

We now formalize the profile device. In particular, we never condition on the random profile realized by the adaptive run. Let $D = SAH$ and $L = \lceil \log_2(K + 1) \rceil$. Pre-generate the mutually independent cell tapes from appendix C.1 and partition each into the disjoint blocks $\mathcal{B}_{x,\ell}$ in equation (3). Write $\widehat{\mathcal{M}}^{(j_x:x)}$ for the mosaic that uses the placeholder at cell x when $j_x = 0$ and the record formed from block $\mathcal{B}_{x,j_x-1}$ when $j_x \ge 1$.

For the candidate-learner analysis, consider a length-K *candidate-only execution sequence* that always takes the candidate branch. The physical wrapper can be coupled to a prefix of this execution by using separate candidate and baseline tapes: baseline episodes reveal no candidate-tape entry and do not change the candidate state. Hence every prefix-uniform candidate-only statement also holds at the random number $N \le K$ of physical candidate executions.

Definition 15 (Total installed-block profile)**.** Immediately before candidate-only execution n, let $I_x^n = 0$ if cell x uses the placeholder and let $I_x^n = \ell + 1$ if its installed record was formed from block $\mathcal{B}_{x,\ell}$. The execution profile is $I^n = (I_x^n)_{x\in\mathcal{S}\times\mathcal{A}\times[H]}$, and the total profile is $\mathcal{I} = (I^1, \dots, I^K)$. Every realizable profile belongs to the following class, whose inequalities are coordinatewise:

$$\mathfrak{I}_K := \big\{(i^1, \dots, i^K) : 0 \le i^1 \le \cdots \le i^K \le L\mathbf{1}\big\}, \tag{35}$$

The charge vector $(m_q)_q$ is deliberately not part of $\mathcal{I}$. We union over it separately in lemma 5. Simultaneous completions merely make several coordinates of I^{n+1} increase after episode n.

Lemma 16 (Profile count and fixed-profile realization)**.** *For every $K \ge 1$,*

$$|\mathfrak{I}_K| \le \binom{K+L}{L}^D \le \big(e(K+L)\big)^{DL}. \tag{36}$$

For each deterministic $\mathcal{I} \in \mathfrak{I}_K$, there is an auxiliary version of the TRAPEZE *candidate learner, defined on the pre-generated tapes, whose mosaic in candidate-only execution n is $\widehat{\mathcal{M}}^{(I_x^n:x)}$. When $\mathcal{I}$ is the realized profile, all its component-indexed policies and value functions equal those of the actual candidate-only execution.*

Proof. For one cell, a coordinatewise nondecreasing length-K sequence with values in $\{0, \dots, L\}$ is determined by the multiplicities of its $L + 1$ values, so there are $\binom{K+L}{L}$ possibilities. Taking the product over the D cells and using $\binom{u}{v} \le (eu/v)^v \le (eu)^v$ proves equation (36).

For a fixed profile, initialize the auxiliary learner from I^1. Whenever $I^{n+1} \ne I^n$, replace the installed records by those prescribed by I^{n+1} and run the planning part of REPLAN, resetting the offline dual path exactly as in the online algorithm. If $I^{n+1} = I^n$, reuse the cached candidate and dual path. All referenced block records exist on the infinite tapes, even for a profile that is not realizable. For the realized profile the prescribed record at every cell and execution is exactly the record used online. Induction over candidate-only executions and backward induction over stages therefore couple every cached value, policy, dual iterate, and component law. Notice that this is a pathwise substitution after a uniform event, not conditioning on the realized profile. □

Lemma 17 (Profile admissibility of the TRAPEZE candidate learner)**.** *Fix a deterministic total profile $\mathcal{I}$. Every stage-$(h+1)$ optimistic or plain-evaluation suffix in $\mathcal{F}_q^{\mathcal{I}}$ is a deterministic function of $\mathcal{I}$, its grid-point label, and tapes at stages strictly later than h. Dependence of the executed dual component on current-stage data enters only through the data-dependent distribution ω_q in lemma 5. Consequently the families generated by the* TRAPEZE *candidate learner are profile-admissible.*

Proof. Fix $\mathcal{I} \in \mathfrak{I}_K$ and run the auxiliary learner in lemma 16. For a block $q = (s, a, h, \ell)$, form $\mathcal{F}_q^{\mathcal{I}}$ from all stage-$(h+1)$ optimistic suffixes $\overline{V}_{h+1,f_\lambda}^{n,\mathcal{I}}$ and plain suffixes $\widehat{V}_{h+1,g}^{\,n,\mathcal{I},\pi_{n,\lambda}^{\mathcal{I}}}$, for $n \in [K]$, $\lambda \in \Lambda$, and $g \in \{r, d\}$, together with zero, their complements, and their scaled squares W^2/H and corresponding complements.

Backward induction in equation (7) shows that every optimistic suffix in this family is a measurable function only of $\mathcal{I}$, λ, and block records at stages $h+1, \dots, H$. The same induction in equation (8) proves the claim for a plain suffix: at stages later than h, the actions of $\pi_{n,\lambda}^{\mathcal{I}}$ were selected using only those later-stage records. Thus, in a filtration that reveals complete block tapes in the order $H, H-1, \dots, 1$, every vector in $\mathcal{F}_q^{\mathcal{I}}$ is measurable before the stage-h block q is revealed and is independent of that block. Zero, complements, and scaled squares preserve this property.

More formally, let $\mathcal{R}_{q,i-1}$ contain all tapes at stages strictly later than h and the entries preceding $S'_{q,i}$ in a fixed within-stage ordering. A fixed family label produces a random vector W_q that is $\mathcal{R}_{q,i-1}$-measurable, and

$$\mathbb{E}[W_q(S'_{q,i}) - P_q W_q \mid \mathcal{R}_{q,i-1}] = 0,$$
$$\mathbb{E}[(W_q(S'_{q,i}) - P_q W_q)^2 \mid \mathcal{R}_{q,i-1}] = \mathrm{Var}_{P_q}(W_q).$$

The rounded dual path and the component actually selected for execution may depend on stage-h data. We do not claim otherwise: this dependence is kept outside the family members and appears

only in the post-event distribution ω_q. Since the concentration event below is simultaneous over every deterministic choice of a *label* from each $\mathcal{F}_q^{\mathcal{I}}$, arbitrary data-dependent choices and mixtures can be substituted afterward. This proves profile admissibility. □

Proof of lemma 5. Each local family has cardinality at most $cK|\Lambda| \le cK(K^2+1)$, and there are at most $Q_K = DL$ non-placeholder blocks. Since $n_{x,\ell} = 2^\ell$, the charge vectors satisfy

$$\prod_{x,\ell}(2n_{x,\ell}+1) \le 2^{D(L^2+3L)/2}. \tag{37}$$

Unioning equations (36) and (37), all joint family labels, both signs, and all K prefixes costs

$$\Gamma_K := c\left[DL\log\left(\frac{cSAHK}{\delta}\right) + \log\left(\frac{cK}{\delta}\right)\right] \le c'SAH\log^2\left(\frac{c'SAHK}{\delta}\right). \tag{38}$$

Initial no-data visits are handled deterministically at cost $O(SAH^2)$.

Deterministic selections and convexification. Fix a deterministic profile index, one deterministic choice of a label from each finite family $\mathcal{F}_q^{\mathcal{I}}$ (producing a reverse-predictable random vector W_q), and one integer $m_q \le 2n_q$ per completed local block. If $X_{q,i} = W_q(S'_{q,i}) - P_q W_q$, then the weighted transition error is

$$\sum_q \frac{m_q}{n_q}\sum_{i=1}^{n_q} X_{q,i}.$$

In the reverse-stage filtration, its increments have range at most $2H$ and predictable variance

$$\sum_q \frac{m_q^2}{n_q}\operatorname{Var}_{P_q}(W_q) \le 2\sum_q m_q \operatorname{Var}_{P_q}(W_q),$$

where both inequalities use $m_q/n_q \le 2$. Freedman's inequality in Bernstein form and the union just counted therefore give

$$\sum_q m_q\langle \widehat{P}_q - P_q, W_q\rangle \le c\left[\sqrt{\Gamma_K\sum_q m_q \operatorname{Var}_{P_q}(W_q)} + H\Gamma_K\right].$$

This bound holds simultaneously for every deterministic label selection and truncated count vector.

Fix an outcome in this simultaneous event, substitute $\mathcal{I} = \mathcal{I}^{\text{true}}$ using lemma 16, and then let independent auxiliary labels have product law $\prod_q \omega_q$. Averaging the pointwise inequality over these labels and applying Jensen gives

$$\mathbb{E}\sqrt{\sum_q m_q \operatorname{Var}_{P_q}(W_q)} \le \sqrt{\sum_q m_q \mathbb{E}_{\omega_q}\operatorname{Var}_{P_q}(W_q)},$$

proving lemma 5. Selecting between W and zero represents a positive-part selector. Since $\langle \widehat{P} - P, H\mathbf{1} - W\rangle = -\langle \widehat{P} - P, W\rangle$ and variances are unchanged, complements give the negative side. □

C.4 Width reductions and coherent total variance

We first make the filtration and the two combinatorial parameters explicit. Let $\mathcal{G}_{n,-1}$ contain the complete physical history available immediately before the nth candidate-component draw (including any intervening baseline episodes), the frozen mosaic, the component law p_n, and all cached policies and value functions. Let $\mathcal{G}_{n,-1/2} = \sigma(\mathcal{G}_{n,-1}, J_n)$, $\mathcal{G}_{n,0} = \sigma(\mathcal{G}_{n,-1/2}, S_{n,1})$, and let $\mathcal{G}_{n,h}$ additionally contain the action, signal observations, and next state revealed through step h. These fields are ordered lexicographically across episodes. Records of completed blocks are installed only after $\mathcal{G}_{n,H}$. Thus p_n and every component-indexed root quantity are $\mathcal{G}_{n,-1}$-measurable, while the suffix used at step h is fixed before that step's next state is exposed.

Put

$$Q_K := SAH\lceil\log_2(K+1)\rceil = DL. \tag{39}$$

There are at most Q_K installed local blocks, their charged counts satisfy $m_q \le 2n_q$, and $\sum_q m_q \le HN$. We take the admissible entropy parameter in lemma 5 large enough that

$$Q_K L_K \le \Gamma_K \le cSAHL_K^2. \tag{40}$$

This is consistent with the profile count above after changing only a universal constant. Let $N_0(N)$ be the number of visits among the first N candidate executions at which no record has yet been installed. The first block has size one, and its record is installed after that execution, so $N_0(N) \le SAH$. All primitive inequalities below are run at fixed constant-fraction allocations of δ and unioned over the two plain signals, the optimistic family, the two signs, and dyadic prefixes. Their intersection is included in $\mathcal{G}_{\mathrm{cand}}$.

Lemma 18 (Latent-mixture reduction)**.** *Suppose $x_{n,u} \in [0,B]$ is $\mathcal{G}_{n,-1}$-measurable for every $u \in \mathrm{supp}(p_n)$. On $\mathcal{G}_{\mathrm{cand}}$, simultaneously for every $N \le K$,*

$$\left|\sum_{n=1}^{N} x_{n,J_n} - \sum_{n=1}^{N} \mathbb{E}_{J\sim p_n} x_{n,J}\right| \le cB\big(\sqrt{NL_K} + L_K\big). \tag{41}$$

Consequently, equation (41) holds at every stopping time bounded by K.

Proof. Conditionally on $\mathcal{G}_{n,-1}$, J_n has law p_n. Hence the difference at time n in equation (41) is a martingale difference of magnitude at most B. Apply the maximal Freedman inequality on horizons $1, 2, 4, \dots, 2^{\lceil \log_2 K\rceil}$ and union bound over those horizons. For $2^{i-1} < N \le 2^i$, increasing c makes the bound at 2^i no larger than the displayed right-hand side. Prefix uniformity then proves the stopping time claim. □

For the remaining reductions, after drawing J_n write $\pi_n := \pi_{n,J_n}$, let λ_n be the grid value carried by that component, and let $q(n,h)$ be the installed block used at $(S_{n,h}, A_{n,h}, h)$, when one exists. For $g \in \{r,d\}$ abbreviate

$$\widehat{V}_{n,h,g} := \widehat{V}_{h,g}^{\pi_n}, \qquad V_{n,h,g} := V_{h,g}^{\pi_n},$$

and define the coherent visit-wise variance

$$\mathcal{V}_{g,N} := \sum_{n=1}^{N}\sum_{h=1}^{H} \mathrm{Var}_{P_h(\cdot|S_{n,h},A_{n,h})}(\widehat{V}_{n,h+1,g}). \tag{42}$$

Lemma 19 (Signed plain-evaluation reduction)**.** *For either $g \in \{r,d\}$, on $\mathcal{G}_{\mathrm{cand}}$ and uniformly for $N \le K$,*

$$\begin{aligned}\sum_{n=1}^{N}\left|\widehat{V}_{1,g}^{\pi_n} - V_{1,g}^{\pi_n}\right| \le{}& c\left[\sqrt{\Gamma_K \mathcal{V}_{g,N}} + H\Gamma_K\right] \\ &+ c\left[\sqrt{Q_K HNL_K} + Q_K L_K\right] \\ &+ c\left[\sqrt{H^3 NL_K} + HL_K\right] + cHN_0(N).\end{aligned} \tag{43}$$

Proof. Let $\Delta_{n,h} = \widehat{V}_{n,h,g} - V_{n,h,g}$ and choose $\sigma_n = \mathrm{sign}(p\Delta_{n,1})$ after J_n is drawn but before $S_{n,1}$ is revealed. Bellman subtraction for an installed visit gives

$$\Delta_{n,h}(s) = \widehat{g}_q - g_h(s,a) + \langle \widehat{P}_q - P_q, \widehat{V}_{n,h+1,g}\rangle + \langle P_q, \Delta_{n,h+1}\rangle, \qquad a = \pi_{n,h}(s). \tag{44}$$

Multiply equation (44) by σ_n, evaluate it along the true trajectory, and telescope over h. Writing $\mathcal{J}_N$ for installed visits gives the exact signed reduction

$$\begin{aligned}\sum_{n=1}^{N}|p\Delta_{n,1}| = {}& \sum_{(n,h)\in\mathcal{J}_N} \sigma_n\left[\widehat{g}_{q(n,h)} - g_h(S_{n,h},A_{n,h})\right] \\ &+ \sum_{(n,h)\in\mathcal{J}_N} \sigma_n\left\langle \widehat{P}_{q(n,h)} - P_{q(n,h)}, \widehat{V}_{n,h+1,g}\right\rangle + \mathcal{M}_N + \mathcal{R}_{0,N},\end{aligned} \tag{45}$$

where $\mathcal{M}_N$ is the sum of the initial-state and transition sampling differences and $|\mathcal{R}_{0,N}| \le cHN_0(N)$. Predictability of σ_n is crucial here. The increments of $\mathcal{M}_N$ are bounded by $2H$ and there are at most $HN + N$ of them. A dyadic maximal martingale bound gives the third bracket in equation (43).

For the model drift in equation (45), use $W = \widehat{V}_{n,h+1,g}$ when $\sigma_n = 1$ and $W = H\mathbf{1} - \widehat{V}_{n,h+1,g}$ when $\sigma_n = -1$. Constants cancel under $\widehat{P}_q - P_q$, and complements preserve variance. Grouping repeated uses of block q makes ω_q the empirical distribution of these selected suffixes, so the drift is exactly

$$\sum_q m_q \mathbb{E}_{W\sim\omega_q}\langle \widehat{P}_q - P_q, W\rangle.$$

The first bracket of equation (43) follows from lemma 5, because its variance term is at most $\mathcal{V}_{g,N}$. Finally, for $q = (s, a, h, \ell)$, simultaneous blockwise signal concentration in equation (33) gives $|\widehat{g}_q - g_h(s, a)| \le c(\sqrt{L_K/n_q} + L_K/n_q)$. Using $m_q/n_q \le 2$, $\sum_q m_q \le HN$, and at most Q_K blocks, Cauchy's inequality gives the second bracket. This treats a reused frozen signal mean as one block error, rather than incorrectly as fresh noise on every use. □

For the selected optimistic component put

$$W^{\mathrm{op}}_{n,h} := \overline{V}_{h+1,f_{\lambda_n}}, \qquad \mathcal{V}^{\mathrm{op}}_N := \sum_{n=1}^{N}\sum_{h=1}^{H} \mathrm{Var}_{P_h(\cdot|S_{n,h},A_{n,h})}(W^{\mathrm{op}}_{n,h}).$$

On installed visits, group the $W^{\mathrm{op}}_{n,h}$ by block and denote their empirical distribution by ω^{op}_q. Define the charged bonus

$$X_N := \sum_q m_q \mathbb{E}_{W\sim\omega^{\mathrm{op}}_q} B_q(W) = \sum_{\substack{n\le N,\, h\le H:\\ q(n,h)\neq\perp}} B_{q(n,h)}(W^{\mathrm{op}}_{n,h}). \tag{46}$$

Lemma 20 (Coherent variance recursion). *On $\mathcal{G}_{\mathrm{cand}}$, for $g \in \{r, d\}$,*

$$\mathcal{V}_{g,N} \le c\left[H^2N + H\sqrt{\Gamma_K\mathcal{V}_{g,N}} + H^2\Gamma_K\right], \tag{47}$$

$$\mathcal{V}_{g,N} \le c\left[H^2N + H^2\Gamma_K\right]. \tag{48}$$

For the optimistic family,

$$\mathcal{V}^{\mathrm{op}}_N \le c\left[H^2N + HX_N + H\sqrt{\Gamma_K\mathcal{V}^{\mathrm{op}}_N} + H^2\Gamma_K\right]. \tag{49}$$

Proof. For one plain installed backup write $v = \widehat{g}_q + \widehat{P}_q w$. Since $v, P_q w \in [0, H]$,

$$\mathrm{Var}_{P_q}(w) \le P_q w^2 - v^2 + 2H\widehat{g}_q + 2H\left[\langle \widehat{P}_q - P_q, w\rangle\right]_+. \tag{50}$$

Along a coherent sampled trajectory, the realized terms $w(S_{n,h+1})^2 - v(S_{n,h})^2$ telescope. The difference between $P_q w^2$ and $w(S_{n,h+1})^2$ is a martingale whose predictable variance is at most $4H^2\mathcal{V}_{g,N}$, because $\mathrm{Var}(w^2) \le 4H^2\,\mathrm{Var}(w)$. More explicitly, summing equation (50) over installed visits gives

$$\mathcal{V}_{g,N} \le \mathcal{M}^{(2)}_N + 2H\sum_{(n,h)\in\mathcal{J}_N}\widehat{g}_{q(n,h)} + 2H\sum_{(n,h)\in\mathcal{J}_N}\left[\langle \widehat{P}_{q(n,h)} - P_{q(n,h)}, w_{n,h}\rangle\right]_+ + cH^2N_0(N), \tag{51}$$

where the coherent square-value terms telescope to a nonpositive terminal boundary, and $\mathcal{M}^{(2)}_N$ is the just-described martingale. Maximal Freedman bounds it by $c[H\sqrt{L_K\mathcal{V}_{g,N}} + H^2L_K]$. Apply lemma 5 after selecting w or zero for each positive part, and use $\sum_{n,h}\widehat{g}_q \le HN$. Together with $L_K \le \Gamma_K$, equation (51) yields equation (47). The $O(H^2N_0(N))$ omitted-block boundary charge is absorbed by $H^2\Gamma_K$. Solving $x \le a + b\sqrt{x}$ gives equation (48).

For an installed optimistic backup let $v = \overline{V}_{h,f_{\lambda_n}}(S_{n,h})$, $w = W^{\mathrm{op}}_{n,h}$, and $\widehat{g}^{\mathrm{eff}} = v - \widehat{P}_q w$. The horizon cap is at least $\widehat{P}_q w$, so

$$0 \le \widehat{g}^{\mathrm{eff}} \le \widehat{f}_{\lambda,q} + B_q(w).$$

Apply equation (50) with $\widehat{g}^{\mathrm{eff}}$ in place of $\widehat{g}_q$. The same square telescoping gives the analogue of equation (51). Its effective-increment sum is at most $HN + X_N$. Its martingale and positive model-drift terms are bounded by the same maximal-Freedman and profile inequalities. This proves equation (49). □

Lemma 21 (Empirical variance and bonus bound)**.** *For any one of the plain or optimistic suffix families, let $\mathcal{V}_N$ denote its true visit-wise variance sum and define*

$$\widehat{\mathcal{V}}_N := \sum_q m_q \mathbb{E}_{W\sim\omega_q} \widehat{\mathrm{Var}}_q(W)$$

over its installed visits. On $\mathcal{G}_{\mathrm{cand}}$,

$$\widehat{\mathcal{V}}_N \le c\left(\mathcal{V}_N + H^2\Gamma_K\right). \tag{52}$$

For the optimistic family, its bonus sum in equation (46) obeys

$$X_N \le c\left[\sqrt{Q_K L_K \widehat{\mathcal{V}}_N} + HQ_K L_K + \sqrt{Q_K HNL_K}\right], \tag{53}$$

$$\mathcal{V}^{\mathrm{op}}_N \le c\left(H^2 N + H^2\Gamma_K\right), \tag{54}$$

$$X_N \le c\left[H\sqrt{Q_K NL_K} + H\Gamma_K\right]. \tag{55}$$

Proof. For a fixed deterministic profile selection,

$$\widehat{\mathrm{Var}}_q(W) - \mathrm{Var}_{P_q}(W) \le |\langle \widehat{P}_q - P_q, W^2\rangle| + 2H|\langle \widehat{P}_q - P_q, W\rangle|.$$

Apply lemma 5 to W, its complement, W^2/H, and the complement of W^2/H, using $\mathrm{Var}(W^2/H) \le 4\,\mathrm{Var}(W)$. Only after these simultaneous deterministic-selection bounds are established do we average with respect to the corresponding product distribution $\prod_q \omega_q$. This preserves the pairing of W and W^2/H and yields

$$\widehat{\mathcal{V}}_N \le \mathcal{V}_N + cH\sqrt{\Gamma_K \mathcal{V}_N} + cH^2\Gamma_K,$$

which implies equation (52) by Young's inequality.

Next use $\max\{x, y\} \le x + y$, Jensen within each block, and then Cauchy's inequality across blocks. Since $m_q/n_q \le 2$, there are at most Q_K blocks, and $\sum_q m_q \le HN$, the definition of B_q gives

$$\begin{aligned} X_N &\le c\sum_q m_q \left[\sqrt{\frac{L_K \mathbb{E}_{\omega^{\mathrm{op}}_q} \widehat{\mathrm{Var}}_q(W)}{n_q}} + \frac{HL_K}{n_q} + \sqrt{\frac{L_K}{n_q}}\right] \\ &\le c\left[\sqrt{Q_K L_K \widehat{\mathcal{V}}_N} + HQ_K L_K + \sqrt{Q_K HNL_K}\right], \end{aligned}$$

which is equation (53). Substitute equations (52) and (53) into equation (49). By equation (40), all terms containing $\sqrt{\mathcal{V}^{\mathrm{op}}_N}$ are absorbed with Young's inequality, giving equation (54). Substitution back into equation (53) gives equation (55). This order of substitution closes the empirical-variance/optimistic-variance recursion without assuming the desired bound. □

Lemma 22 (Optimistic width assembly)**.** *On $\mathcal{G}_{\mathrm{cand}}$, uniformly for $N \le K$,*

$$\begin{aligned} \sum_{n=1}^{N} z_{n,J_n} \le X_N &+ c\left[\sqrt{\Gamma_K \mathcal{V}^{\mathrm{op}}_N} + H\Gamma_K\right] \\ &+ c\left[\sqrt{Q_K HNL_K} + Q_K L_K\right] + c\left[\sqrt{H^3 NL_K} + HL_K\right] + cHN_0(N). \end{aligned} \tag{56}$$

Together with lemmas 18 to 21, this proves lemma 6.

Proof. For an installed visit of the selected optimistic policy, the clipped backup and the true Bellman equation give the one-sided recursion

$$\overline{V}_{h,f_{\lambda_n}}(S_{n,h}) - V^{\pi_n}_{h,f_{\lambda_n}}(S_{n,h}) \le \widehat{f}_{\lambda_n,q} - f_{\lambda_n,h}(S_{n,h},A_{n,h}) + \langle \widehat{P}_q - P_q, W^{\mathrm{op}}_{n,h}\rangle + B_q(W^{\mathrm{op}}_{n,h}) + \left\langle P_q, \overline{V}_{h+1,f_{\lambda_n}} - V^{\pi_n}_{h+1,f_{\lambda_n}} \right\rangle.$$

Evaluate this inequality along the true trajectory. Since $z_{n,J_n} \ge 0$ by lemma 3, the root terms may be summed one-sidedly. Telescoping the continuation gaps gives

$$\sum_{n=1}^{N} z_{n,J_n} \le \sum_{(n,h)\in\mathcal{J}_N} \left(\widehat{f}_{\lambda_n,q(n,h)} - f_{\lambda_n,h}(S_{n,h},A_{n,h})\right) + \sum_{(n,h)\in\mathcal{J}_N} \langle \widehat{P}_{q(n,h)} - P_{q(n,h)}, W^{\mathrm{op}}_{n,h}\rangle + X_N + \mathcal{M}^{\mathrm{op}}_N + cHN_0(N), \tag{57}$$

where $\mathcal{M}^{\mathrm{op}}_N$ contains the initial-state and true-transition sampling differences. Select $W^{\mathrm{op}}_{n,h}$ or zero according to the sign of each model drift and group by installed block. Then lemma 5 gives the first statistical bracket in equation (56). The simultaneous scalar signal bound in equation (32) gives the next bracket, and a maximal martingale inequality applied to $\mathcal{M}^{\mathrm{op}}_N$ gives the third. The charged bonuses in equation (57) sum exactly to X_N, while no-data visits cost at most cH each. This proves equation (56).

Apply equations (48), (54) and (55) in equations (43) and (56) for the reward, utility, and optimistic families. Using $N_0(N) \le SAH$ and equation (40) gives

$$\sum_{n=1}^{N} \left(|\xi^r_{n,J_n}| + |\xi^d_{n,J_n}| + z_{n,J_n}\right) \le c\left[\sqrt{SAH^3N}\, L_K + SAH^2 L_K^2\right]. \tag{58}$$

Finally apply lemma 18 to the globally defined $x_{n,u} = |\xi^r_{n,u}| + |\xi^d_{n,u}| + [z_{n,u}]_+ \in [0,3H]$. On the scalar-optimism event, $[z_{n,u}]_+ = z_{n,u}$. Its error is dominated by the right side of equation (58), so equation (16) follows after enlarging c_0. Every bound is prefix uniform. Hence the conclusion holds for all stopping times $N \le K$, as asserted in lemma 6. □

D TRAJECTORY CERTIFICATION AND TRANSFER ANALYSIS

Theorem 23 (Safe switching transfer)**.** *Under assumption 1, apply the wrapper in equation (13) to a candidate-clock learner with at most M segments. At each segment start, its finite latent candidate distribution is history measurable and then frozen. Before each physical episode, the mixing weight is history measurable, the branch coin has the corresponding conditional Bernoulli law, and any component draw uses fresh conditional randomness. Only a candidate execution may advance the candidate learner, update the certificate, or end a segment. A baseline execution changes none of these objects. Suppose also that, on $\mathcal{G}_{\mathrm{cand}}$, the candidate learner uniformly satisfies $\mathrm{SRegret}(N) \le \mathfrak{R}_N$ and $\mathrm{SCV}(N) \le \mathfrak{C}_N$ for $N \le K$. There is an event $\mathcal{G}_{\mathrm{wrap}}$ of probability at least $1-2\delta$ such that, on $\mathcal{G}_{\mathrm{cand}} \cap \mathcal{G}_{\mathrm{wrap}}$, every deployed policy is safe and, for the random candidate count $N \le K$,*

$$\mathrm{Regret}(K) \le \widetilde{O}\Big(\mathfrak{R}_N + H\sqrt{K} + \frac{\Delta_0}{\kappa}\big[\mathfrak{C}_N + H\sqrt{MN} + HM + H\sqrt{K}\big]\Big). \tag{59}$$

Lemma 24 (Safety by scalar certification)**.** *If $V^{\rho_e}_{1,c} \le U^c_{e,j}(\rho_e)$ for every created segment and attained sample size, every deployed policy in equation (13) satisfies equation (2).*

Proof. If $U^c_{e,j}(\rho_e) \le b$, then equation (13) sets $\alpha_{e,j} = 1$ and the candidate is certified directly. Otherwise, linearity of the latent mixture, $V^{\pi_0}_{1,c} = b - \kappa$, and equation (13) gives

$$V^{\mu_k|\mathcal{H}_{k-1}}_{1,c} \le \alpha_{e,j} U^c_{e,j}(\rho_e) + (1-\alpha_{e,j}) V^{\pi_0}_{1,c} = b.$$

□

Proof of lemma 8. For compactness, put

$$L_{\mathrm{cert}}(M,K,\delta) := \log\left(\frac{4(M+1)^2(K+1)^2}{\delta}\right).$$

Let τ_e be the physical episode at which segment e is created. Set $\tau_e = K+1$ if it is not created by episode K, and extend its candidate arbitrarily on that event. Abbreviate $C_e := V_{1,c}^{\rho_e}$. The random policy ρ_e, its component list $(\pi_{e,u})_{u=1}^T$, and its component law p_e are $\mathcal{H}_{\tau_e-1}$-measurable and remain fixed until the segment ends. For TRAPEZE, $p_e(u) = 1/T$.

On one common enlargement of the probability space, conditionally on $\mathcal{H}_{\tau_e-1}$ append to every created segment an infinite ghost tape of candidate executions under its frozen ρ_e. The algorithm reveals an initial prefix of this tape. Intervening baseline episodes merely delay successive revelations. Ghost observations after the actual segment ends are never used by the algorithm. They only let us state one prefix-uniform event whose validity is unaffected by the random segment length.

Let $\mathcal{F}_{e,i}^-$ be the sigma-field immediately before the latent component of the ith candidate execution is drawn. It contains the complete preceding physical history and the fact that the candidate branch has been selected, but not the fresh component $J_{e,i}$ or any randomness from its trajectory. Let $\mathcal{F}_{e,i}$ be the corresponding post-execution sigma-field, so that $\mathcal{F}_{e,i-1} \subseteq \mathcal{F}_{e,i}^- \subseteq \mathcal{F}_{e,i}$. Conditionally on $\mathcal{F}_{e,i}^-$, the algorithm draws $J_{e,i} \sim p_e$ and follows $\pi_{e,J_{e,i}}$ for the whole episode. The tower property and the definition of the latent mixture therefore give

$$\mathbb{E}[Y_{e,i} \mid \mathcal{F}_{e,i}^-, J_{e,i} = u] = V_{1,c}^{\pi_{e,u}}, \tag{60}$$

$$\mathbb{E}[Y_{e,i} \mid \mathcal{F}_{e,i}^-] = \sum_{u=1}^{T} p_e(u) V_{1,c}^{\pi_{e,u}} = V_{1,c}^{\rho_e} = C_e. \tag{61}$$

The conditioning boundary in equation (61) is essential: after $J_{e,i}$ is revealed, the conditional mean is generally the component value $V_{1,c}^{\pi_{e,J_{e,i}}}$, not the mixture value C_e.

Consequently, for $X_{e,i} := Y_{e,i} - C_e$, another application of the tower property gives $\mathbb{E}[X_{e,i} \mid \mathcal{F}_{e,i-1}] = 0$. Thus $(X_{e,i})_{i\ge1}$ is a martingale-difference sequence at candidate-execution boundaries, and $|X_{e,i}| \le H$. For each fixed $j \ge 1$, the conditional Hoeffding inequality for martingale differences and equation (12) give

$$\begin{aligned}\mathbb{P}\big(|\overline{Y}_{e,j} - C_e| > \beta_{e,j} \,\big|\, \mathcal{H}_{\tau_e-1}\big) \\ \le 2\exp\left(-\frac{j\beta_{e,j}^2}{2H^2}\right) = \frac{\delta}{2(e+1)^2(j+1)^2}.\end{aligned}$$

The ghost construction makes this statement valid for every prefix, including all prefixes attained before the trajectory-dependent replan time.

It remains to justify the adaptive creation of the segments themselves. Let $\mathcal{A}_e = \{\tau_e \le K\}$ and let $\mathcal{E}_e$ be the event that the deviation bound holds for every prefix of segment e's ghost tape. By the conditional bound above and the tower property,

$$\begin{aligned}\mathbb{P}(\mathcal{A}_e \cap \mathcal{E}_e^c) &= \mathbb{E}[\mathbf{1}\{\mathcal{A}_e\}\mathbb{P}(\mathcal{E}_e^c \mid \mathcal{H}_{\tau_e-1})] \\ &\le \frac{\delta}{2(e+1)^2}\sum_{j=1}^{\infty}\frac{1}{(j+1)^2}.\end{aligned}$$

Summing over $e \ge 1$ uses $\sum_{m=2}^{\infty} m^{-2} < 1$ twice and gives total failure probability below δ. Thus $|\overline{Y}_{e,j} - C_e| \le \beta_{e,j}$ simultaneously for all adaptively created segments and attained prefixes.

On this event, both H and $\overline{Y}_{e,j} + \beta_{e,j}$ are at least C_e. Hence, for every $j \ge 1$, clipping preserves coverage and gives

$$0 \le U_{e,j}^c(\rho_e) - C_e \le \overline{Y}_{e,j} + \beta_{e,j} - C_e \le 2\beta_{e,j}. \tag{62}$$

At $j = 0$, equation (12) instead gives $0 \le U_{e,0}^c(\rho_e) - C_e \le H$.

Finally, for the first M created segments and every attained $j \le K$,

$$\beta_{e,j} \le H\sqrt{\frac{2L_{\mathrm{cert}}(M,K,\delta)}{j}}.$$

If segment e contains n_e candidate executions, the certificates used on those executions have pre-execution sample counts $j = 0, \dots, n_e - 1$. Using equation (62) and $\sum_{j=1}^{n-1} j^{-1/2} \le 2\sqrt{n}$ gives

$$\begin{aligned}
&\sum_{e=1}^{M} \sum_{j=0}^{n_e-1} \left(U^c_{e,j}(\rho_e) - C_e\right) \\
&\quad \le HM + 2H\sqrt{2L_{\mathrm{cert}}(M,K,\delta)} \sum_{e=1}^{M} \sum_{j=1}^{n_e-1} \frac{1}{\sqrt{j}} \\
&\quad \le HM + 4\sqrt{2}\, H\sqrt{L_{\mathrm{cert}}(M,K,\delta)} \sum_{e=1}^{M} \sqrt{n_e} \\
&\quad \le HM + 4\sqrt{2}\, H\sqrt{MNL_{\mathrm{cert}}(M,K,\delta)},
\end{aligned}$$

where the last line follows from Cauchy's inequality and $\sum_{e=1}^{M} n_e = N$. This implies equation (20). □

Proof of theorem 23. For physical episode k, let e_k be its segment and let j_k be the number of candidate executions already completed in that segment before the branch coin is drawn. Put

$$\alpha_k := \alpha_{e_k,j_k}, \qquad r_k^+ := \left[V^\star_{1,r} - V^{\rho_{e_k}}_{1,r}\right]_+, \qquad v_k := \left[V^{\rho_{e_k}}_{1,c} - b\right]_+,$$
$$u_k := U^c_{e_k,j_k}(\rho_{e_k}) - V^{\rho_{e_k}}_{1,c}, \qquad \chi_k := (1-\alpha_k)/\alpha_k,$$

and let $Z_k \mid \mathcal{H}_{k-1} \sim \mathrm{Bernoulli}(\alpha_k)$ indicate the candidate branch. On the certification event, $u_k \ge 0$ and equation (13) implies

$$\chi_k \le (v_k + u_k)/\kappa. \tag{63}$$

The physical pseudo-regret is bounded by

$$\mathrm{Regret}(K) \le \sum_{k=1}^{K} \left[\alpha_k r_k^+ + (1-\alpha_k)\Delta_0\right] = \sum_k \alpha_k r_k^+ + \Delta_0 \sum_k \alpha_k \chi_k.$$

The partial-sum processes associated with the one-step differences

$$M_r := \sum_k (\alpha_k - Z_k) r_k^+, \qquad M_\chi := \sum_k (\alpha_k - Z_k)\chi_k$$

are predictable Bernoulli martingales. Since $r_k^+ \le H$ and $\chi_k \le H/\kappa$, one-sided Hoeffding inequalities for martingales and a union bound give, with probability at least $1-\delta$,

$$M_r \le H\sqrt{2K\log(2/\delta)}, \qquad M_\chi \le \frac{H}{\kappa}\sqrt{2K\log(2/\delta)}. \tag{64}$$

On candidate branches, the virtual sequence is exactly the history seen by the candidate learner. Consequently,

$$\begin{aligned}
\sum_k \alpha_k r_k^+ &= \sum_k Z_k r_k^+ + M_r \le \mathfrak{R}_N + M_r, \\
\sum_k \alpha_k \chi_k &= \sum_k Z_k \chi_k + M_\chi \\
&\le \frac{1}{\kappa}\left(\sum_k Z_k v_k + \sum_k Z_k u_k\right) + M_\chi \\
&\le \frac{1}{\kappa}\left(\mathfrak{C}_N + \mathrm{Cert}_N\right) + M_\chi,
\end{aligned}$$

where, because the candidate executions in segment e encounter the pre-execution counts $0, \dots, n_e - 1$ exactly once,

$$\sum_{k=1}^{K} Z_k u_k = \sum_e \sum_{j=0}^{n_e-1} \left(U^c_{e,j}(\rho_e) - V^{\rho_e}_{1,c}\right) =: \mathrm{Cert}_N\,.$$

It is bounded by equation (20). Substitution in the physical-regret display and equation (64) proves equation (59), with only one inverse-slack factor. Intersecting this thinning event with the probability-$1-\delta$ certificate event gives $\mathcal{G}_{\mathrm{wrap}}$. □

E PROOF OF THE MATCHING LOWER BOUND

E.1 SMALL-SLACK CONSTRUCTION

We first prove theorem 9 for $0 < \kappa \le H/4$ by establishing

$$\mathbb{P}_\nu\left(\mathrm{Regret}(K) \ge c\min\left\{HK, \frac{H}{\kappa}\sqrt{SAH^3K}\right\}, \quad V_{1,r}^\star - V_{1,r}^{\mu_k|\mathcal{H}_{k-1}} \ge 0 \ \ \forall k \le K\right) \ge 0.3. \tag{65}$$

Proof of the small-slack case. The construction is a contextual direct sum. Let

$$L := S - 4, \qquad m := \left\lfloor \frac{H-2}{3} \right\rfloor, \qquad J := m(A-1)+1, \qquad R := H - m - 1. \tag{66}$$

Here $L \ge S/2$, $m \ge H/6$, $J \ge AH/12$, and $R \ge H/2$. The initial distribution is $p = \delta_{s_\mathrm{w}}$. The states are the root s_w, L context states, and absorbing states $s_0, s_\mathrm{g}, s_\mathrm{b}$. At stage one, action a_0 enters the baseline state s_0, whereas every other action enters the exploratory branch and reveals a uniformly random context at stage two. The gate therefore precedes the context revelation. Any pair of a state and a stage not reached by the transitions below is completed identically across instances using an arbitrary fixed kernel and fixed signals.

In a context, at stages $2, \dots, m$, action a_0 waits while every action $a \ne a_0$ probes. At stage $m+1$, every action probes. A probe transitions to s_g or s_b, which then absorb. Thus each exploratory trajectory selects exactly one of $(m-1)(A-1)+A = J$ arms indexed by time and action. The stages $m+2, \dots, H$ form a common tail of length R irrespective of the selected arm.

An instance is indexed by $\theta = (\theta_1, \dots, \theta_L) \in [J]^L$. In context s, every ordinary arm reaches s_g with probability $1/2$, whereas arm θ_s reaches it with probability $1/2 + \varepsilon$. The complementary probability leads to s_b. Set

$$\varepsilon := \frac{1}{100}\min\left\{\frac{\kappa}{H}, \sqrt{\frac{LJ}{K}}\right\}. \tag{67}$$

It remains to define the signals. Set

$$b = H/2, \qquad u_0 = \frac{1}{2} - \frac{\kappa}{H}, \qquad u = \frac{1}{2} + \frac{\kappa}{H}, \qquad v = \frac{1}{4}.$$

At stage one, the baseline action has reward zero and cost u_0. Every action in s_0 keeps the process there with reward zero and cost u_0. Thus the baseline has known value $b - \kappa$. Every reachable state-action pair on the exploratory branch has reward one. Its cost is u outside the common tail, including visits to either outcome state before the tail. Every action in s_g and s_b keeps that state absorbing. During the common tail its cost is respectively $u - v$ and $u + v$. The assumption $\kappa \le H/4$ keeps all signals in $[0, 1]$. Every exploratory policy has reward H, while one that selects the distinguished arm with context-averaged probability q has cost

$$b + \kappa - \gamma q, \qquad \gamma := 2vR\varepsilon = R\varepsilon/2. \tag{68}$$

Fix an episode k. Let z_k be its pre-episode probability of taking the exploratory gate. Conditional on taking that gate and observing context s, let $q_{k,s,j}$ be the policy's probability of selecting arm j and define

$$q_k := \frac{1}{L}\sum_{s=1}^{L} q_{k,s,\theta_s}. \tag{69}$$

These are regular conditional probabilities under the full pre-episode filtration, including the learner's past private randomization. The policy is understood as a full stochastic kernel at every history. If $z_k = 0$, we use a fixed canonical completion on the off-path exploratory histories. This keeps $q_{k,s,j}$ measurable with respect to $\mathcal{H}_{k-1}$ and preserves $\sum_j q_{k,s,j} = 1$.

Under instance θ, the optimal feasible policy selects θ_s in each revealed context and mixes the resulting exploratory policy with the baseline at mass

$$p_\star := \frac{\kappa}{2\kappa - \gamma} \ge \frac{1}{2}. \tag{70}$$

Indeed, every exploratory continuation has reward H and its cost is minimized at $q = 1$. Moreover, equation (67) gives $\gamma \le \kappa/200 < \kappa$, so even this cheapest continuation is unsafe and $p_\star \in (0, 1)$. No feasible policy can use more exploratory mass. On the learner's simultaneous-safety event $\mathcal{E}_\theta$, episode k must satisfy

$$z_k(2\kappa - \gamma q_k) \le \kappa = (2\kappa - \gamma)p_\star. \tag{71}$$

Because $q_k \le 1$, this first gives $z_k \le p_\star$. Rearranging again,

$$2\kappa(p_\star - z_k) \ge \gamma(p_\star - z_k q_k) \ge \gamma p_\star(1 - q_k). \tag{72}$$

Consequently, on $\mathcal{E}_\theta$, every reward gap is nonnegative and

$$\mathrm{Regret}(K) = H\sum_{k=1}^{K}(p_\star - z_k) \ge \frac{H\gamma}{4\kappa}\sum_{k=1}^{K}(1 - q_k). \tag{73}$$

We next show that some θ has linear cumulative prediction error with constant probability. Fix a context s and the other coordinates θ_{-s}. Let $\mathbb{P}^{s,0}_{\theta_{-s}}$ denote the hybrid law in which every arm in context s has success probability $1/2$, while all other contexts retain their distinguished arms. Let $N_{s,j}$ count actual probes of arm j in context s and put $Q_{s,j} := \sum_{k=1}^{K} q_{k,s,j}$. Under the hybrid law,

$$\sum_{j=1}^{J}\mathbb{E}^{s,0}_{\theta_{-s}} N_{s,j} \le K/L, \qquad \sum_{j=1}^{J} Q_{s,j} = K \quad \text{pathwise.} \tag{74}$$

Only the probed transition differs between the hybrid and the alternative whose sth coordinate is j. The adaptive KL chain rule and $\mathrm{kl}(\mathrm{Ber}(1/2)\|\,\mathrm{Ber}(1/2+\varepsilon)) \le 4\varepsilon^2$ therefore give

$$\mathrm{KL}\Big(\mathbb{P}^{s,0}_{\theta_{-s}} \| \mathbb{P}_{(\theta_{-s},j)}\Big) \le 4\varepsilon^2 \mathbb{E}^{s,0}_{\theta_{-s}} N_{s,j}.$$

Applying Pinsker's inequality to $Q_{s,j}/K \in [0, 1]$, averaging over j, and using Cauchy's inequality with equation (74) yields

$$\begin{aligned}\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}_{(\theta_{-s},j)} Q_{s,j} &\le \frac{K}{J} + \frac{K\sqrt{2}\varepsilon}{J}\sum_{j=1}^{J}\sqrt{\mathbb{E}^{s,0}_{\theta_{-s}} N_{s,j}} \\ &\le \frac{K}{J} + K\varepsilon\sqrt{\frac{2K}{LJ}}.\end{aligned} \tag{75}$$

This coordinatewise comparison is essential: directly comparing an all-null instance with a product alternative would sum the KL over all L coordinates.

Average equation (75) over s, θ_{-s}, and a uniform product prior on θ. All expectations below are joint over this instance draw and the trajectory data. Since equation (69) implies $\sum_k q_k = L^{-1}\sum_s Q_{s,\theta_s}$, we obtain

$$\mathbb{E}_{\theta\sim\mathrm{Unif}([J]^L)}\mathbb{E}_\theta \sum_{k=1}^{K} q_k \le \frac{K}{J} + K\varepsilon\sqrt{\frac{2K}{LJ}}. \tag{76}$$

Here $J \ge 3$, and equation (67) makes the right-hand side at most $(1/3 + \sqrt{2}/100)K$. Thus, for $D_\theta := \sum_k(1 - q_k) \in [0, K]$, the product-prior mean exceeds $0.65K$. The bounded reverse-Markov inequality

$$\mathbb{E}D_\theta \le K\mathbb{P}(D_\theta \ge K/3) + \frac{K}{3}\mathbb{P}(D_\theta < K/3)$$

shows that the product-prior average of $\mathbb{P}_\theta(D_\theta \ge K/3)$ exceeds 0.47. Thus some fixed θ satisfies $\mathbb{P}_\theta(D_\theta \ge K/3) > 0.47$. Together with $\mathbb{P}_\theta(\mathcal{E}_\theta) \ge 0.9$, this gives intersection probability at least 0.3.

On this intersection, equation (73) and $R \ge H/2$ give

$$\mathrm{Regret}(K) \ge \frac{H^2\varepsilon K}{48\kappa} \ge c\min\left\{HK, \frac{H}{\kappa}\sqrt{SAH^3K}\right\},$$

where the final step uses $LJ \ge SAH/24$. This proves equation (65). □

E.2 LARGE-SLACK CONSTRUCTION

Completion of the proof of theorem 9. For $\kappa \le H/4$, the preceding construction suffices because $1 + H/\kappa \le (5/4)H/\kappa$. For $\kappa > H/4$, use the same contextual dynamics, set every cost to zero and $b = \kappa$, and give the learner the zero-cost baseline. Put reward one in s_g during the R common-tail stages and zero elsewhere. Now choose

$$\varepsilon = \frac{1}{100} \min\left\{1, \sqrt{\frac{LJ}{K}}\right\}.$$

The coordinatewise hybrid argument equations (74) to (76) is unchanged, so some θ satisfies $\mathbb{P}_\theta(D_\theta \ge K/3) > 0.47$.

The optimal policy always takes the exploratory gate and selects the distinguished arm, obtaining $R(1/2 + \varepsilon)$. If episode k has exploratory mass z_k and conditional success probability q_k, its reward gap is

$$R\left[\frac{1 - z_k}{2} + \varepsilon(1 - z_k q_k)\right] \ge R\varepsilon(1 - q_k).$$

Hence, on the preceding event,

$$\mathrm{Regret}(K) \ge \frac{R\varepsilon K}{3} \ge c'' \min\{HK, \sqrt{SAH^3K}\}.$$

All policies are safe and every per-episode gap is nonnegative. Finally, $1 + H/\kappa < 5$ in this regime, which proves the stated result after changing the universal constant. □

Theorems 2 and 9 show that

$$\min\left\{HK, \left(1 + \frac{H}{\kappa}\right)\sqrt{SAH^3K}\right\}$$

is the minimax leading order over the full slack range, with no reachability qualification. The additive term in the upper bound creates no gap. If $K < SAH$, both rates clip at HK. If $K \ge SAH$, then $SAH^2 \le \sqrt{SAH^3K}$. The encoding by time and action is related to the hard MDP architecture of Domingues et al. (2021), while the post-gate uniform context and coordinatewise hybrid argument are specific to our joint construction coupling dimension and slack. To our knowledge, theorem 9 is the first lower bound showing that inverse safety slack multiplies the full tabular exploration difficulty. Whether inverse-slack dependence can be removed was left open by Kitamura et al. (2025).

F ONE-STEP SPECIAL CASE

Theorem 25 (Safe-bandit lower bound)**.** *Fix $A \ge 3$, $K \ge A$, $b = 1/2$, and $0 < \kappa \le 1/8$. Suppose that, on every instance ν, a learner guarantees with probability 0.9 that $V_{1,c}^{\mu_k|\mathcal{H}_{k-1}} \le b$ simultaneously for all $k \le K$. For every $\Delta \in (0, 1]$, there is a one-step A-action CMDP with a known baseline of true cost $b - \kappa$ such that*

$$\mathbb{P}_\nu\left(\mathrm{Regret}(K) \ge \frac{\Delta}{40} \min\left\{K, \frac{\sqrt{(A-1)K}}{\kappa}\right\}\right) \ge 0.7. \tag{77}$$

A deterministic-transition absorbing-state embedding multiplies the bound by H for every $H \ge 1$, without changing the budget, slack, or statistical observation law.

At $H = 1$, theorem 25 uses the expected-policy model introduced by Pacchiano et al. (2021) and developed further by Pacchiano et al. (2025). When the baseline reward gap is constant, Theorem 34 of Pacchiano et al. (2025) specializes to $\Omega(\max\{\sqrt{(A-1)K}, \kappa^{-2}\})$ in its stated regimes. Our multiplicative inverse-slack term complements that result and resolves, for this tabular submodel, the necessity question highlighted by Kitamura et al. (2025). The preceding contextual construction is stronger because the dimension and inverse slack factors multiply.

Proof of theorem 25. There is one baseline action and $A-1$ risky actions. The baseline has deterministic reward zero and deterministic cost $b-\kappa$. Every risky action has deterministic reward Δ. Set

$$q_+ = b+\kappa+\varepsilon, \qquad q_- = b+\kappa-\varepsilon, \qquad \varepsilon = \min\left\{\frac{\kappa}{2}, \frac{\sqrt{A-1}}{20\sqrt{K}}\right\}.$$

Under the null instance ν_0, every risky action has cost $\mathrm{Ber}(q_+)$. Under alternative ν_j, risky action j has cost $\mathrm{Ber}(q_-)$, while all other risky actions retain cost $\mathrm{Ber}(q_+)$.

Let $p_k(\mathcal{H}_{k-1})$ be the total conditional probability that μ_k assigns to risky actions. Under ν_0, feasibility is equivalent to

$$(1-p_k)(b-\kappa)+p_k q_+ \le b,$$

and hence to

$$p_k \le p_+ := \frac{\kappa}{2\kappa+\varepsilon}.$$

Consequently, for

$$\mathcal{E} := \{p_k(\mathcal{H}_{k-1}) \le p_+ \text{ for every } k \le K\},$$

the assumed simultaneous safety guarantee gives $\mathbb{P}_{\nu_0}(\mathcal{E}) \ge 0.9$.

Let

$$N_j := \sum_{k=1}^{K} \mathbf{1}\{A_k = j\}$$

be the actual number of pulls of risky action j. Since $\sum_{j=1}^{A-1} \mathbb{E}_{\nu_0} N_j \le K$, there exists j with $\mathbb{E}_{\nu_0} N_j \le K/(A-1)$. The adaptive KL chain rule gives

$$\mathrm{KL}(\mathbb{P}_{\nu_0} \| \mathbb{P}_{\nu_j}) = \mathbb{E}_{\nu_0} N_j \ \mathrm{kl}(\mathrm{Ber}(q_+) \| \mathrm{Ber}(q_-)).$$

Here $q_- \in [1/2, 5/8]$, and therefore

$$\mathrm{kl}(\mathrm{Ber}(q_+) \| \mathrm{Ber}(q_-)) \le \frac{(q_+-q_-)^2}{q_-(1-q_-)} \le 18\varepsilon^2.$$

Using $\varepsilon^2 \le (A-1)/(400K)$ yields

$$\mathrm{KL}(\mathbb{P}_{\nu_0} \| \mathbb{P}_{\nu_j}) \le \frac{18}{400} < \frac{1}{20}.$$

Pinsker's inequality consequently implies

$$\mathbb{P}_{\nu_j}(\mathcal{E}) \ge 0.9 - \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbb{P}_{\nu_0} \| \mathbb{P}_{\nu_j})} > 0.7.$$

Under ν_j, the optimal feasible policy mixes only risky action j with the baseline, assigning it probability $p_- := \kappa/(2\kappa-\varepsilon)$. On $\mathcal{E}$, the learner obtains reward at most Δp_+ in every episode, whereas the optimal feasible value is Δp_-. Hence

$$\begin{aligned}\mathrm{Regret}(K) &\ge \Delta K(p_- - p_+) = \Delta K \frac{2\kappa\varepsilon}{4\kappa^2-\varepsilon^2} \ge \Delta K \frac{\varepsilon}{2\kappa} \\ &\ge \frac{\Delta}{40}\min\left\{K, \frac{\sqrt{(A-1)K}}{\kappa}\right\}.\end{aligned}$$

This proves equation (77).

For the horizon-H embedding, retain the same first-stage action, reward Δ, and cost experiment. A risky action enters an absorbing state that earns another reward Δ at each stage $2, \dots, H$, while the baseline enters a zero-reward absorbing state. All later costs are zero. Thus a risky first-stage choice earns $H\Delta$, while the cost observations and KL calculation are unchanged. □