Proceedings of Machine Learning Research vol :1–46, 2026 39th Annual Conference on Learning Theory

Optimal Variance-Dependent Regret Bounds for Infinite-Horizon MDPs

Guy Zamir GZAMIR@WISC.EDU
Matthew Zurek MATTHEW.ZUREK@WISC.EDU
Yudong Chen YUDONGCHEN@CS.WISC.EDU
University of Wisconsin–Madison

Editors: Steve Hanneke and Tor Lattimore

Abstract

Online reinforcement learning in infinite-horizon Markov decision processes (MDPs) remains less theoretically and algorithmically developed than its episodic counterpart, with many algorithms suffering from high "burn-in" costs and failing to adapt to benign instance-specific complexity. In this work, we address these shortcomings for two infinite-horizon objectives: the classical average-reward regret and the γ-regret. We develop a single tractable UCB-style algorithm applicable to both settings, which achieves the first optimal variance-dependent regret guarantees. Our regret bounds in both settings take the form $\widetilde{O}(\sqrt{SA\,\mathrm{Var}}$+lower-order terms), where S, A are the state and action space sizes, and Var captures cumulative transition variance. This implies minimax-optimal average-reward and γ-regret bounds in the worst case but also adapts to easier problem instances, e.g. yielding nearly constant regret in deterministic MDPs. Furthermore, our algorithm enjoys significantly improved lower-order terms for the average-reward setting. With prior knowledge of the optimal bias span $\|h^\star\|_{\mathrm{sp}}$, our algorithm obtains lower-order terms scaling as $\|h^\star\|_{\mathrm{sp}}S^2A$, which we prove is optimal in both $\|h^\star\|_{\mathrm{sp}}$ and A. Without prior knowledge, we prove that no algorithm can have lower-order terms smaller than $\|h^\star\|_{\mathrm{sp}}^2SA$, and we provide a prior-free algorithm whose lower-order terms scale as $\|h^\star\|_{\mathrm{sp}}^2S^3A$, nearly matching this lower bound. Taken together, these results completely characterize the optimal dependence on $\|h^\star\|_{\mathrm{sp}}$ in both leading and lower-order terms, and reveal a fundamental gap in what is achievable with and without prior knowledge.

1. Introduction

We study online reinforcement learning (RL) in tabular Markov decision processes (MDPs), where an agent interacts with an unknown environment and aims to maximize cumulative reward. We specifically consider infinite-horizon continuing settings, where the environment does not contain a built-in reset mechanism. Despite its practical relevance and foundational significance, online infinite-horizon RL is much less well understood theoretically than the finite-horizon episodic setting. In this work we study two particular performance measures for infinite-horizon problems.

The most classical performance objective is the average-reward regret $\sum_t(\rho^\star - r_t)$ introduced by the seminal work Auer et al. (2008), which measures the instantaneous reward r_t of the agent against the optimal gain $\rho^\star$, which is the best long-term average reward per timestep of any policy. The reset-less nature of infinite-horizon online RL requires additional structural assumptions to permit sublinear regret bounds, such as for the MDP to be communicating, and these also ensure that $\rho^\star$ can be defined independent of an initial state. Auer et al. (2008) proved a regret lower bound of $\Omega(\sqrt{DSAT})$, where D is the MDP diameter, S and A are the numbers of states and actions, and T is the time horizon. Extensive research effort has since been dedicated to matching this regret lower bound and relaxing the communicativity assumptions by replacing D with $\|h^\star\|_{\text{sp}}$, the maximum gap in cumulative rewards between any two starting states, which is smaller than D and finite even in non-communicating MDPs (e.g., Bartlett and Tewari 2012; Fruit et al. 2018, 2019; Talebi and Maillard 2018; Zhang and Xie 2023), culminating in the minimax-optimal algorithms of Zhang and Ji (2019); Boone and Zhang (2024).

Another performance measure for infinite-horizon online RL is the γ-regret $\sum_t((1-\gamma)V_\gamma^\star(s_t)-r_t)$ introduced by Liu and Su (2021), where $V_\gamma^\star(s_t)$ denotes the γ-discounted optimal value function at the state s_t encountered by the agent. While γ-regret may appear to be a weaker regret notion due to its comparison with the agent's own trajectory rather than that of an optimal policy, this feature has the advantage of enabling sublinear regret bounds without the structural assumptions needed for the average-reward setting. Furthermore, when such communicativity assumptions do hold, we show the γ-regret can actually be used to control the average-reward regret in an optimal manner, which is the approach we take in this paper (see Lemma 4). Recent work has developed algorithms achieving the minimax-optimal γ-regret, $\widetilde{O}(\sqrt{SAT/(1-\gamma)})$ (He et al., 2021; Ji and Li, 2023).

Despite all of the aforementioned algorithmic progress on the average-reward and the γ-regret settings, there are several significant limitations of all existing work in both settings relative to episodic online RL. First, existing minimax-optimal algorithms incur a large burn-in cost, meaning that they only attain the optimal regret rate when T is very large. For example, the only computationally efficient and minimax-optimal algorithm for the average-reward regret, PMEVI-DT (Boone and Zhang, 2024), has a regret bound of $\widetilde{O}\Big(\sqrt{\|h^\star\|_{\text{sp}}SAT}+\|h^\star\|_{\text{sp}}S^{\frac{5}{2}}A^{\frac{3}{2}}T^{\frac{9}{20}}\Big)$, and thus it only matches the optimal rate for $T \geq \|h^\star\|_{\text{sp}}^{10}S^{40}A^{20}$. This contrasts the episodic setting, where significant effort has been expended towards remedying such issues (e.g., Zhang et al. 2021; Zhou et al. 2023; Zhang et al. 2024). Second, prior work for infinite-horizon settings fails to adapt to easier problem instances such as low-variance or deterministic MDPs, where substantially smaller regret should be possible. In episodic RL, this gap has been addressed through variance-dependent regret bounds, which interpolate between stochastic and deterministic environments and can be optimal in both regimes (Zhou et al., 2023). However, no optimal variance-dependent regret guarantees have been established for either infinite-horizon setting that we consider.

1.1. Contributions

In this paper, we establish the first optimal variance-dependent regret guarantees for infinite-horizon MDPs. Our main contribution is a single tractable algorithm that, in both the average-reward and γ-regret settings, attains a regret bound of the form

$$\widetilde{O}\big(\sqrt{\text{Var}_\gamma\, SA}+\text{lower-order terms}\big).$$

We handle both of these infinite-horizon settings in a unified way by treating $\gamma = 1-\frac{1}{T}$ as a tuning parameter in the average-reward case. Here Var_γ is a cumulative variance term that measures the stochasticity of the transition dynamics along the learner's trajectory. When the MDP is deterministic, we have $\text{Var}_\gamma = 0$, and thus the resulting regret is independent of T up to logarithmic factors. We also show $\text{Var}_\gamma \leq \widetilde{O}(\|V_\gamma^\star\|_{\text{sp}}T+\|V_\gamma^\star\|_{\text{sp}}^2)$, so by using that $\|V_\gamma^\star\|_{\text{sp}} \leq 2\|h^\star\|_{\text{sp}}$ in weakly communicating MDPs (Wei et al., 2020), or simply that $\|V_\gamma^\star\|_{\text{sp}} \leq \frac{1}{1-\gamma}$, we obtain minimax-optimal regret bounds in both the average-reward and γ-regret settings, respectively.

Focusing on the average-reward setting, another main contribution is that we significantly improve the lower-order terms relative to prior work. When given prior knowledge of the bias span

Algorithm	$\widetilde{O}(\cdot)$ Regret	Priorless?	Burn-In Cost	Tractable?	Type
UCRL2 (Auer et al., 2008)	$DS\sqrt{AT}$	✓	N/A	✓	EVI
REGAL (Bartlett and Tewari, 2012)	$\|h^\star\|_{\text{sp}}S\sqrt{AT}$	×	N/A	×	EVI
SCAL (Fruit et al., 2018)	$\|h^\star\|_{\text{sp}}S\sqrt{AT}$	×	N/A	✓	EVI
KL-UCRL (Talebi and Maillard, 2018)	$\sqrt{ST\sum_{s,a}\mathbb{V}^\star_{s,a}} + D\sqrt{T}$	✓	N/A	✓	EVI
EBF (Zhang and Ji, 2019)	$\sqrt{\|h^\star\|_{\text{sp}}SAT}$	×	(*)	×	EVI
UCB-AVG (Zhang and Xie, 2023)	$S^5A^2\|h^\star\|_{\text{sp}}\sqrt{T}$	×	N/A	✓	UCB
PMEVI-DT (Boone and Zhang, 2024)	$\sqrt{\|h^\star\|_{\text{sp}}SAT}$	✓	$\|h^\star\|_{\text{sp}}^{10}S^{40}A^{20}$	✓	EVI
γ-UCB-CVI (Hong et al., 2025)	$\|h^\star\|_{\text{sp}}S\sqrt{AT}$	×	N/A	✓	UCB
Corollary 6	$\sqrt{\text{Var}^\star_{1-1/T}SA} + \Gamma\|h^\star\|_{\text{sp}}SA$	×	$\|h^\star\|_{\text{sp}}S^3A$	✓	UCB
Corollary 7	$\sqrt{\|h^\star\|_{\text{sp}}SAT}$	✓	$\|h^\star\|_{\text{sp}}^2 S^3A$	✓	UCB
Lower Bound (Auer et al., 2008)	$\sqrt{DSAT}$, implies $\sqrt{\|h^\star\|_{\text{sp}}SAT}$				

Table 1: **Comparison of algorithms and regret bounds for average-reward MDPs.** Here $\|h^\star\|_{\text{sp}}$ is the span of the optimal bias function and D is the diameter, which satisfy $\|h^\star\|_{\text{sp}} \leq D$. We always have $\Gamma \leq S$, and $\Gamma = 1$ in deterministic MDPs. $\text{Var}^\star_{1-1/T}$ and $\mathbb{V}^\star_{sa}$ are instance-dependent variance parameters, which in particular are 0 for deterministic MDPs. Also $\text{Var}^\star_{1-1/T} \leq \widetilde{O}(\|h^\star\|_{\text{sp}}T + \|h^\star\|_{\text{sp}}^2)$. Only the leading terms as $T \to \infty$ of the regret bound are shown. The burn-in cost is defined as the smallest T for which the algorithm achieves a minimax-optimal regret of $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}}SAT})$, or N/A if this does not occur. Priorless means that an algorithm does not require prior knowledge about the value of $\|h^\star\|_{\text{sp}}$. (*) The burn-in cost of EBF (Zhang and Ji, 2019) is $\|h^\star\|_{\text{sp}}^6S^4A^4 + \|h^\star\|_{\text{sp}}^6S^6A^2 + \|h^\star\|_{\text{sp}}^3S^{12}A^3$.

$\|h^\star\|_{\text{sp}}$, our (variance-dependent and minimax-optimal) result contains lower-order terms scaling as $\|h^\star\|_{\text{sp}}S^2A$, and we show via matching lower bounds that this dependence on $\|h^\star\|_{\text{sp}}$ and A is unimprovable. Additionally, without prior knowledge of $\|h^\star\|_{\text{sp}}$, we obtain lower-order terms scaling as $\|h^\star\|_{\text{sp}}^2S^3A$. We also show a surprising hardness result that no algorithm without prior knowledge of $\|h^\star\|_{\text{sp}}$ can obtain lower-order terms better than $\|h^\star\|_{\text{sp}}^2SA$. Taken together, our results nearly completely characterize the optimal dependence on $\|h^\star\|_{\text{sp}}$ in both the leading and lower-order terms and reveal a fundamental separation between what is achievable with and without prior knowledge.

To obtain our improved bounds for γ-regret, we develop a model-based, upper-confidence-bound(UCB)-based algorithm called Fully Optimizing Clipped UCB Solver (FOCUS). We improve upon previous UCB-based algorithms for γ-regret by incorporating a sharp Bernstein-style bonus and span-clipping into our empirical Bellman operator. Crucially, instead of performing a single step of value iteration at each update, FOCUS repeatedly applies the empirical Bellman operator until convergence. This design ensures that the Q-estimates fully exploit the collected data at each update and is essential for obtaining variance-dependent bounds. Finally, FOCUS is the first UCB-style algorithm to achieve minimax-optimal regret guarantees for the online average-reward setting, contrasting previous optimal algorithms which instead depend on extended value iteration (EVI).

1.2. Related Work

Here we discuss related work, referring to Appendix A for additional discussion.

Online Average-Reward The average-reward setting is classical and well-studied for online infinite-horizon RL. See Table 1 for a comparison of important prior work; however, given the extensive history of this problem, this table is non-exhaustive. The seminal work of Auer et al. (2008) introduces UCRL2 and establishes a regret bound of $\widetilde{O}(DS\sqrt{AT})$, along with a lower bound of $\Omega(\sqrt{DSAT})$. Bartlett and Tewari (2012) establish regret bounds that depend on $\|h^\star\|_{\text{sp}}$ instead of D, but their algorithm REGAL is computationally intractable algorithm and requires prior knowledge of $\|h^\star\|_{\text{sp}}$. The computationally effcient algorithm SCAL by Fruit et al. (2018) utilizes the span-clipping technique to match the bound of REGAL. The EBF algorithm of Zhang and Ji (2019) is the first to achieve minimax optimality of $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}}SAT})$, but it is intractable and requires prior knowledge. Boone and Zhang (2024) resolves these issues by developing PMEVI-DT, which is tractable, minimax optimal and prior-knowledge-free. Talebi and Maillard (2018) obtain variance-aware guarantees for the KL-UCRL algorithm, but their bounds remain suboptimal in the worst-case and suffer $\sqrt{T}$ dependence even on deterministic MDPs. All aforementioned algorithms are based on EVI. UCB-based algorithms, which like ours also employ discounting to approximate the average-reward problem, include Wei et al. (2020); Zhang and Xie (2023); Hong et al. (2025).

γ-regret Liu and Su (2021) introduce the γ-regret notion and prove the first upper bound with their Double Q-learning algorithm. Subsequent work of He et al. (2021) proposes UCBVI-γ, a model-based algorithm with Bernstein-style bonuses and achieving a γ-regret bound that is minimax-optimal in the leading term. Ji and Li (2023) develop Q-SlowSwitch-Adv, a model-free algorithm that matches the optimal leading term while improving lower-order dependence on SA. Recently, Ma and Lee (2026) take a Bayesian approach with the EUBRL algorithm, which leverages epistemic uncertainty for directed exploration and achieves state-of-the-art lower-order dependence on $\frac{1}{1-\gamma}$. We note that the definition of γ-regret in our paper matches that of Liu and Su (2021), whereas He et al. (2021), Ji and Li (2023), and Ma and Lee (2026) use slightly different definitions. While there is not an immediate translation between these definitions, they are closely related and algorithms with good guarantees for one should have good guarantees for the others. For further discussion on this issue, see Appendix A.2 of He et al. (2021) and Appendix A of Ji and Li (2023).

2. Preliminaries

MDP Basics A Markov Decision Process is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \mu_0)$,[1] where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel with $\Delta(\mathcal{S})$ denoting the probability simplex over $\mathcal{S}$, $r\colon \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, and $\mu_0 \in \Delta(\mathcal{S})$ is the initial state distribution. We assume $\mathcal{S}$ and $\mathcal{A}$ are finite sets with $S \coloneqq |\mathcal{S}|$ and $A \coloneqq |\mathcal{A}|$. A (stationary) policy is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. When π is a deterministic policy, we treat π as a mapping $\mathcal{S} \to \mathcal{A}$. Let Π be the set of all stationary deterministic policies. An initial state distribution $\mu \in \Delta(\mathcal{S})$ and a policy π induce a distribution over trajectories $(s_0, a_0, s_1, a_1, \dots)$, where $s_0 \sim \mu$, $a_t \sim \pi(s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$. We let $\mathbb{E}_s^\pi$ denote the expectation with respect to this distribution when μ satisfies $\mu(s) = 1$.

For a policy π and discount factor $\gamma \in (0,1)$, the discounted value function $V_\gamma^\pi \in [0, \frac{1}{1-\gamma}]^{\mathcal{S}}$ is defined by $V_\gamma^\pi(s) = \mathbb{E}_s^\pi[\sum_{t=0}^\infty \gamma^t r_t]$, where $r_t = r(s_t, a_t)$. Also define the optimal value function $V_\gamma^\star \in [0, \frac{1}{1-\gamma}]^{\mathcal{S}}$ by $V_\gamma^\star(s) = \sup_{\pi \in \Pi} V_\gamma^\pi(s)$. We often write $P_{s,a}$ to denote the row vector such that $P_{s,a}(s') = P(s'|s,a)$. Then for any $V \in \mathbb{R}^{\mathcal{S}}$ we have $P_{s,a}V = \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]$. The gain of a

1. When considering classes of MDPs with the same $\mathcal{S}, \mathcal{A}, r$, and μ_0, we simply denote the MDP by its kernel P.

policy π, $\rho^\pi \in [0,1]^{\mathcal{S}}$, is defined by $\rho^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^\pi[\sum_{t=0}^{T-1} r_t]$. We define the optimal gain $\rho^\star = \sup_{\pi\in\Pi} \rho^\pi$. The bias function of a policy π, $h^\pi \in \mathbb{R}^{\mathcal{S}}$, is $h^\pi(s) = \text{C-lim}_{T\to\infty}\mathbb{E}_s^\pi[\sum_{t=0}^{T-1}(r_t - \rho^\pi(s_t))]$. The optimal bias function is $h^\star := h^{\pi^\star}$, where $\pi^\star$ is a Blackwell-optimal policy, which satisfies $V_\gamma^{\pi^\star} = V_\gamma^\star$ for all sufficiently large γ.

The diameter of an MDP is $D = \max_{s_1\neq s_2} \inf_{\pi\in\Pi} \mathbb{E}_{s_1}^\pi[\eta_{s_2}]$, where η_s denotes the hitting time of a state $s \in \mathcal{S}$. An MDP is (strongly) communicating if its diameter is finite; that is, any state is reachable from any other state under some policy. An MDP is weakly communicating if the states can be partitioned into two disjoint subsets $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ such that all states in $\mathcal{S}_1$ are transient under all policies, and within $\mathcal{S}_2$ any state is reachable from any other state under some policy. In weakly communicating MDPs, the optimal gain $\rho^\star$ is constant and thus, with a slight abuse of notation, treated as a scalar. We say an MDP is deterministic if the transition probabilities $P(\cdot|s,a)$ are one-hot vectors, i.e., from each state-action pair the agent will transit to a certain state with probability 1. We also define $\Gamma := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\text{supp}(P(\cdot|s,a))|$.

Online RL and Regrets The learner interacts with the MDP for T steps, starting from a state $s_1 \sim \mu_0$. At each step $t = 1,\ldots,T$, the learner at state s_t chooses an action a_t and observes the next state $s_{t+1} \sim P(\cdot|s_t,a_t)$. The learner aims to maximize the total reward $\sum_{t=1}^T r(s_t,a_t)$ it receives. We consider two different notions of regret, which measure the disparity between the learner's reward and that of an optimal policy. Given a discount factor $\gamma \in (0,1)$, define $\text{Regret}_\gamma(T) := \sum_{t=1}^T \big((1-\gamma)V_\gamma^\star(s_t) - r(s_t,a_t)\big)$. When the MDP is weakly communicating, further define $\text{Regret}(T) := \sum_{t=1}^T \big(\rho^\star - r(s_t,a_t)\big)$. Throughout the paper, γ-regret or discounted setting refers to $\text{Regret}_\gamma(T)$, and regret or average-reward setting refers to $\text{Regret}(T)$.

The regret and γ-regret are both functions of the underlying MDP and the trajectory $s_1, a_1, \ldots, s_T, a_T$. The trajectory is random, with a distribution determined by the MDP, the learning algorithm, and the time horizon T. Oftentimes we only explicitly write T as a parameter in regret because the MDP and learning algorithm are usually clear from context. In situations where the underlying MDP P and learning algorithm $\texttt{Alg}$ are not obvious, we write $\text{Regret}(T, P, \texttt{Alg})$.

Burn-In Cost We say that an algorithm achieves a burn-in cost of $f(\|h^\star\|_{\text{sp}}, S, A)$ if for all MDPs M and any $T \geq f(\|h_M^\star\|_{\text{sp}}, S_M, A_M)$, the regret can be bounded (with high probability) by $\widetilde{O}\big(\sqrt{\|h_M^\star\|_{\text{sp}} S_M A_M T}\big)$, where $h_M^\star, S_M$, and A_M are the optimal bias function, state space size, and action space size, respectively, of M. We use lower-order terms to refer to any of the additive terms in the regret bound besides the leading term, which is typically $\sqrt{\text{Var}_{1-1/T}^\star SA}$ or $\sqrt{\|h^\star\|_{\text{sp}} SAT}$. Note that the lower-order terms may sometimes dominate the regret, including when the MDP is deterministic so that $\text{Var}_{1-1/T} = 0$.

To illustrate the difference between burn-in cost and lower-order terms, consider an algorithm that obtains a regret bound of $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}} SAT} + \|h^\star\|_{\text{sp}} S^2 A)$. Then the algorithm has a lower-order term of $\|h^\star\|_{\text{sp}} S^2 A$ and achieves a burn-in cost of $\|h^\star\|_{\text{sp}} S^3 A$. There is an conversion — albeit a lossy one — between the two notions. If an algorithm has a burn-in cost of f, then its regret can be bounded by $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}} SAT} + f(\|h^\star\|_{\text{sp}}, S, A))$, because for large T the leading term dominates, and for small T the regret is at most $T < f(\|h^\star\|_{\text{sp}}, S, A)$. If an algorithm achieves a regret of $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}} SAT} + f(\|h^\star\|_{\text{sp}}, S, A))$, then the algorithm has a burn-in cost of $\frac{f(\|h^\star\|_{\text{sp}}, S, A)^2}{\|h^\star\|_{\text{sp}} SA}$, since for T larger than this quantity the leading term dominates.

Additional Notation Let $[m]$ denote $\{1,\ldots,m\}$ for any positive integer m. Let $\mathbf{0}, \mathbf{1}$ be the all-zero and all-one vectors. For $x \in \mathbb{R}^{\mathcal{S}}$, define the span semi-norm $\|x\|_{\text{sp}} := \max_{s\in\mathcal{S}} x(s) -$

$\min_{s\in\mathcal{S}} x(s)$. For $x, y \in \mathbb{R}^n$, we define $\mathbb{V}(x, y) := \sum_{i=1}^n x_i y_i^2 - (\sum_{i=1}^n x_i y_i)^2$. Observe that when x is a probability vector, $\mathbb{V}(x, y)$ is the variance of a random variable that takes value y_i with probability x_i. For $H \geq 0$, we define the clipping operator $\mathrm{Clip}_H : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ by $(\mathrm{Clip}_H(V))(s) = \min\{V(s), \min_{s'\in\mathcal{S}} V(s') + H\}$. We also define the maximum operator $M\colon \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$ by $(MQ)(s) = \max_{a\in\mathcal{A}} Q(s, a)$. The notation $O(\cdot)$ hides constant factors. The notation $\widetilde{O}(\cdot)$ hides constant factors and possible polylog factors of S, A, T, and $\frac{1}{\delta}$.

3. Main Results

In this section, we first describe our algorithm and discuss improvements over prior work. We then define our variance-dependent term $\mathrm{Var}_\gamma^\star$ and relate it to other relevant quantities. Next, we provide our main bound for γ-regret and implications for the discounted setting. We then reduce the average-reward setting to the discounted setting and derive optimal bounds on the regret with and without prior knowledge. Finally, we state lower bounds showing that our lower-order dependence on $\|h^\star\|_{\mathrm{sp}}$ is optimal in both the prior knowledge and prior-free settings.

3.1. Algorithm

Algorithm 1: Fully Optimizing Clipped UCB Solver (FOCUS)

Input: run time $T \geq 1$, discount factor $\gamma \in (0, 1)$, failure probability $\delta \in (0, 1)$, span clipping parameter $H \geq 1$

```
k ← 1, U ← log(1/δ'), where δ' = δ/(9S²AT)
N(s,a) ← 0, N(s,a,s') ← 0 for all (s,a,s') ∈ S × A × S
Q̂_1(s,a) ← 1/(1−γ) for all (s,a) ∈ S × A
Observe s_1
for t = 1, ..., T do
    Take action a_t ∈ argmax_{a∈A} Q̂_k(s_t, a) and observe s_{t+1}
    N(s_t,a_t) ← N(s_t,a_t) + 1, N(s_t,a_t,s_{t+1}) ← N(s_t,a_t,s_{t+1}) + 1
    if N(s_t,a_t) = 2^i for some integer i ≥ 0 then
        k ← k + 1, ε_k ← 1/(t(1−γ))
        N_k(s,a) ← N(s,a), N_k(s,a,s') ← N(s,a,s') for all (s,a,s')
        P̂^k_{s,a,s'} ← N_k(s,a,s')/N_k(s,a) for all (s,a,s') such that N_k(s,a) > 0
        P̂^k_{s,a,s'} ← 1/S for all (s,a,s') such that N_k(s,a) = 0
        Compute Q̂_k = T̂_k^(m)(0) where m = ⌈ (1/(1−γ)) log((1+32HU)/(ε_k(1−γ))) ⌉
    end
end
```

We present our algorithm, Fully Optimizing Clipped UCB Solver (FOCUS), in Algorithm 1. The algorithm uses a model-based approach: it keeps track of state-action visitation counts and uses these to maintain an empirical estimate of the transition kernel. The algorithm runs through episodes, with a new episode starting when the number of visits to a state-action pair doubles. At the start of the kth episode, the algorithm updates the empirical transition kernel $\widehat{P}^k$, then uses a clipped optimistic value iteration procedure to compute $\widehat{Q}_k$, which is an optimistic estimate of $Q^\star$. At each time step t through the remainder of the episode, the algorithm observes state $s_t \in \mathcal{S}$ and takes a greedy action $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}_k(s_t, a)$.

FOCUS uses the following clipped optimistic empirical Bellman operators $\widehat{\mathcal{T}}_k$. Fixing the episode number k, for any $Q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, define

$$\widehat{\mathcal{T}}_k(Q)(s,a) := r(s,a) + \gamma \widehat{P}^k_{s,a}\big(\text{Clip}_H(MQ)\big) + \gamma b\big(s, a, \text{Clip}_H(MQ)\big)$$

where $b(s,a,V) := \max\left\{4\sqrt{\frac{\mathbb{V}(\widehat{P}^k_{s,a}, V)U}{\max\{N_k(s,a),1\}}}, 32\frac{HU}{\max\{N_k(s,a),1\}}\right\}$ for $V \in \mathbb{R}^{\mathcal{S}}$, and U, H, γ are defined within Algorithm 1. $\widehat{\mathcal{T}}_k$ incorporates three key features: span-clipping (ensuring all value estimates have span $\leq H$), a sharp Bernstein-style bonus analogous to that of MVP (Zhang et al., 2021), and full fixed-point computation via iterated application until convergence. All three ingredients are essential for obtaining bounds tight enough to perform our average-to-discount reduction, for reasons that we further discuss in Section 4.

Computational Complexity We now consider the computational complexity of FOCUS. By the doubling rule, there are at most $O(SA \log T)$ episodes. At the start of each episode, we update our empirical counts and transition kernel, which takes $O\left(S^2 A\right)$ time. We then run $O(\frac{1}{1-\gamma} \log T)$ steps of value iteration, and since each step of value iteration takes $O(S^2 A)$ time, this takes a total of $O(\frac{S^2 A}{1-\gamma} \log T)$ time. So across all episodes, the total run time of FOCUS is $O\big(\frac{S^3 A^2}{1-\gamma} (\log T)^2 + T\big)$ (the additional T is from updating $N(s_t, a_t)$ and $N(s_t, a_t, s_{t+1})$ at each time step t). For the average-reward setting where we set $\gamma = 1 - \frac{1}{T}$, the run time is $O(S^3 A^2 T)$.

3.2. Variance Parameters

Next, we introduce the main variance-dependent term that will later appear in our regret bounds.

Definition 1 *Let $\gamma \in (0,1)$. For a particular MDP with transition kernel P and trajectory $\{s_t, a_t\}_{t \in [T]}$, define the cumulative variance as* $\mathrm{Var}^\star_\gamma := \sum_{t=1}^T \mathbb{V}\big(P_{s_t, a_t}, V^\star_\gamma\big)$.

One can interpret $\mathrm{Var}^\star_\gamma$ as a measure of stochasticity in an MDP. When the transition probabilities $P(\cdot|s,a)$ are more concentrated on states of similar value, $\mathrm{Var}^\star_\gamma$ will generally be smaller. If the MDP is completely deterministic, it is easy to see that $\mathrm{Var}^\star_\gamma = 0$ for any trajectory. For a more detailed discussion on related variance parameters (in the episodic setting), see Zhou et al. (2023).

Existing variance-dependent guarantees for infinite-horizon MDPs (Talebi and Maillard, 2018) involve terms similar to $T\mathbb{V}^\star_\gamma$, where $\mathbb{V}^\star_\gamma := \max_{s,a} \mathbb{V}(P_{s,a}, V^\star_\gamma)$ is the maximum per-step variance. Our bounds will instead depend on $\mathrm{Var}^\star_\gamma$, which is a random but much sharper quantity. It is easy to see that $\mathrm{Var}^\star_\gamma \leq T\mathbb{V}^\star_\gamma$. Furthermore, while $T\mathbb{V}^\star_\gamma$ can be as large as $T\|V^\star_\gamma\|_{\text{sp}}^2$, the following lemma shows that $\mathrm{Var}^\star_\gamma$ scales with $T\|V^\star_\gamma\|_{\text{sp}}$. Thus this lemma can be used to derive minimax-optimal span-based bounds from our variance-dependent bounds in both the discounted and average-reward settings, extending a successful approach from offline settings (e.g., Zurek and Chen 2025b) to the variance measure $\mathrm{Var}^\star_\gamma$ relevant for online learning. We defer the proof to Appendix F.

Lemma 2 *For $\delta \in (0,1)$, we have with probability $1-\delta$ that* $\mathrm{Var}^\star_\gamma \leq O\big(\|V^\star_\gamma\|_{\text{sp}} T + \|V^\star_\gamma\|_{\text{sp}}^2 \log(T/\delta)\big)$.

3.3. Main Results for Discounted Setting

We now present our main theorem on the performance of FOCUS in the discounted setting.

Theorem 3 (Variance-Dependent γ-Regret Bound) *Let $T \geq 1, \gamma \in (0,1), \delta \in (0,1)$. For any $H \geq \|V^\star_\gamma\|_{\text{sp}}$, Algorithm 1 with input (T, γ, δ, H) achieves, with probability at least $1-\delta$,*

$$\text{Regret}_\gamma(T) \leq O\Big(\sqrt{SA\text{Var}^\star_\gamma \log(SAT/\delta)} + \Gamma HSA \log^2(SAT/\delta)\Big).$$

We provide a complete proof in Appendix E. Theorem 3 establishes the first variance-dependent γ-regret bound. Unlike prior works that fail to exploit easier environments and thereby necessarily scale with $\sqrt{T}$, the leading term of our bound depends on $\text{Var}^\star_\gamma$, which as previously mentioned captures the stochasticity of transition dynamics in the MDP. Consequently, our bound interpolates between stochastic and deterministic environments and is significantly sharper in the latter case.

To illustrate these improvements, we consider implications when one has prior knowledge of the span $\|V^\star_\gamma\|_{\text{sp}}$. In this case, one can set $H = \|V^\star_\gamma\|_{\text{sp}}$ to obtain a regret bound of $\widetilde{O}\big(\sqrt{SA\text{Var}^\star_\gamma} + \Gamma\|V^\star_\gamma\|_{\text{sp}}SA\big)$. When the MDP is deterministic, the γ-regret is T-independent up to logarithmic factors, scaling as $\widetilde{O}\big(\|V^\star_\gamma\|_{\text{sp}}SA\big)$. For stochastic MDPs, the leading term $\widetilde{O}\big(\sqrt{\|V^\star_\gamma\|_{\text{sp}}SAT}\big)$ matches the minimax lower bound for γ-regret. Although this rate was previously achieved by He et al. (2021) and Ji and Li (2023), their bounds depend on $\frac{1}{1-\gamma}$ in both the leading term and the lower-order terms. Now, $\|V^\star_\gamma\|_{\text{sp}}$ can be as large as $\frac{1}{1-\gamma}$, but it has the potential to be bounded independently of γ, such as in weakly communicating MDPs where $\|V^\star_\gamma\|_{\text{sp}} \leq 2\|h^\star\|_{\text{sp}}$ (Wei et al., 2020).

We remark that while Theorem 3 is the the first explicit span-dependent bound for γ-regret, the analysis in Hong et al. (2025) can be slightly modified to yield a span-dependent bound of $\widetilde{O}\big(\|V^\star_\gamma\|_{\text{sp}}S\sqrt{AT} + \frac{S}{1-\gamma}\big)$ (see Appendix C). However, this result is not minimax optimal and still has a lower-order dependence on $\frac{1}{1-\gamma}$. Their algorithm also requires prior knowledge of the span. Theorem 3, on the other hand, attains a span-based bound without prior knowledge. In particular, by setting $H = \frac{1}{1-\gamma}$, we obtain a bound of $\widetilde{O}\big(\sqrt{\|V^\star_\gamma\|_{\text{sp}}SAT} + \frac{S^2A}{1-\gamma}\big)$. We can even remove the lower-order dependence on $\frac{1}{1-\gamma}$ by setting $H = \sqrt{T/(S^3A)}$, similar to Corollary 7 below.

3.4. Main Results for Average-Reward Setting

In this section we use our results from the discounted setting with a properly tuned γ to approximate the average-reward setting. That is, by setting the discount factor γ to be large enough, the γ-regret bound achieved by FOCUS implies an optimal variance-dependent regret. Our reduction hinges on the following lemma. The proof follows in a straightforward manner from standard average-to-discounted reduction techniques (Wei et al., 2020; Zurek and Chen, 2025a), and we defer the complete proof to Section G.

Lemma 4 *Suppose the MDP is weakly communicating. For any $T \geq 1$ and $\gamma \in (0,1)$, it holds that* $\text{Regret}(T) \leq (1-\gamma)\big\|V^\star_\gamma\big\|_{\text{sp}}T + \text{Regret}_\gamma(T)$.

Lemma 4 decomposes regret into γ-regret and an approximation error term $(1-\gamma)\|V^\star_\gamma\|_{\text{sp}}T$. By choosing γ close to 1, this term becomes negligible, and the regret is bounded by γ-regret. This observation reinforces the notion that a span-based γ-regret bound is crucial for deriving optimal guarantees in the average-reward setting. Indeed, when γ is close to 1, applying Lemma 4 to a γ-regret bound that scales with $\frac{1}{1-\gamma}$ instead of the span would imply a vacuous regret bound. We can now state our main theorem for the average-reward setting. It follows by combining Theorem 3 and Lemma 4, taking $\gamma = 1 - \frac{1}{T}$, and using the fact that $\|V_\gamma^\star\|_{\text{sp}} \le 2\|h^\star\|_{\text{sp}}$ for all $\gamma \in (0,1)$.

Theorem 5 (Variance-Dependent Regret Bound) *Suppose the MDP is weakly communicating. Let $T \ge 1$ and $\delta \in (0,1)$. For any $H \ge 2\|h^\star\|_{\text{sp}}$, Algorithm 1 with input $(T, \gamma = 1 - \frac{1}{T}, \delta, H)$ achieves, with probability at least $1-\delta$,*

$$\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\mathrm{Var}^\star_{1-\frac{1}{T}} SA} + \Gamma HSA\Big).$$

Theorem 5 establishes the first minimax optimal variance-dependent regret bound for the average-reward setting. Similar to our γ-regret result, the leading term depends on $\mathrm{Var}^\star_{1-\frac{1}{T}}$, so the regret adapts to the stochasticity of the environment. We remark that Talebi and Maillard (2018) provide a variance-dependent regret bound, but it only implies a suboptimal $\widetilde{O}(DS\sqrt{AT})$ regret bound in the worst case. Furthermore, their bound includes a lower order term of $\widetilde{O}(D\sqrt{T})$, which means it cannot be optimal for deterministic MDPs. Thus, as demonstrated by the following corollary, we have the first regret guarantee that is simultaneously optimal for stochastic and deterministic MDPs.

Corollary 6 (Regret Bound with Prior Knowledge) *Suppose the MDP is weakly communicating. Let $T \ge 1$ and $\delta \in (0,1)$. Algorithm 1 with input $(T, \gamma = 1 - \frac{1}{T}, \delta, H = 2\|h^\star\|_{\text{sp}})$ satisfies $\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\mathrm{Var}^\star_{1-\frac{1}{T}} SA} + \Gamma \|h^\star\|_{\text{sp}} SA\Big)$ with probability at least $1-\delta$. Consequently, by Lemma 2, with probability at least $1-2\delta$, $\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\|h^\star\|_{\text{sp}} SAT} + \|h^\star\|_{\text{sp}} S^2 A\Big)$. Provided that $T \ge \|h^\star\|_{\text{sp}} S^3 A$, it follows that with probability at least $1-2\delta$, $\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\|h^\star\|_{\text{sp}} SAT}\Big)$.*

Corollary 6 shows the optimal bounds that FOCUS attains across different regimes. When the underlying MDP is deterministic, the regret scales as $\widetilde{O}\big(\|h^\star\|_{\text{sp}} SA\big)$, which is optimal and T-independent up to logarithmic factors. For stochastic MDPs, the leading term is minimax optimal, while the lower-order term is significantly smaller than those incurred by existing algorithms. We later show that this $\|h^\star\|_{\text{sp}} S^2 A$ lower order term is nearly optimal — it could be improved at most by a factor of S to $\|h^\star\|_{\text{sp}} SA$. Note that Corollary 6 applies Theorem 5 with span bound $H = 2\|h^\star\|_{\text{sp}}$, which requires prior knowledge of $\|h^\star\|_{\text{sp}}$. We next consider the performance of FOCUS when we do not have prior knowledge of $\|h^\star\|_{\text{sp}}$.

Corollary 7 (Regret Bound without Prior Knowledge) *Suppose the MDP is weakly communicating. Let $T \ge 1$ and $\delta \in (0,1)$. Algorithm 1 with input $(T, \gamma = 1 - \frac{1}{T}, \delta, H = \sqrt{T/(S^3 A)})$ satisfies, with probability at least $1-\delta$, $\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\mathrm{Var}^\star_{1-\frac{1}{T}} SA} + \sqrt{SAT}\Big)$, provided that $T \ge \|h^\star\|_{\text{sp}}^2 S^3 A$. Consequently, Lemma 2 implies that with probability at least $1-2\delta$, $\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\big(\|h^\star\|_{\text{sp}} + 1\big) SAT}\Big)$ provided that $T \ge \|h^\star\|_{\text{sp}}^2 S^3 A$. It follows that with probability at least $1-2\delta$, we have $\mathrm{Regret}(T) \le \widetilde{O}\Big(\sqrt{\big(\|h^\star\|_{\text{sp}} + 1\big) SAT} + \|h^\star\|_{\text{sp}}^2 S^3 A\Big)$.*

We show how the last two regret bounds follow from the first in Appendix I. Corollary 7 shows that our algorithm achieves significantly improved burn-in cost compared to that of previous work on priorless algorithms. Indeed, PMEVI-DT attains minimax optimality for $T \ge \|h^\star\|_{\text{sp}}^{10} S^{40} A^{20}$, whereas our algorithm attains minimax optimality for $T \geq \|h^\star\|_{\text{sp}}^2 S^3 A$. Furthermore, we prove a matching lower bound in Theorem 8 showing that the lower-order dependence on the bias span cannot be improved beyond $\|h^\star\|_{\text{sp}}^2$. We remark that our choice of H is optimized for the more interesting scenario that $\|h^\star\|_{\text{sp}} \geq 1$ and leads to lower-order terms scaling quadratically in $\|h^\star\|_{\text{sp}}$. If one cares about the possibility of $\|h^\star\|_{\text{sp}} \ll 1$, we could instead choose H to be lower-order in T, such as $H = (T/(S^3 A))^{1/(2+\varepsilon)}$, so that the leading term is $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}} SAT})$ for large T albeit with slightly worse lower-order terms.

Unlike the case with prior knowledge of $\|h^\star\|_{\text{sp}}$, Theorem 5 does not immediately imply a simultaneously optimal bound for stochastic and deterministic MDPs. We claim it is still possible to obtain some form of T-independent bound for certain MDP instances. Towards this end, consider running a diameter estimation procedure (i.e., Tarbouriech et al. 2021; Tuynman et al. 2024) for at most $\sqrt{T/(S^3 A)}$ steps. With high probability, it will either terminate within $\text{poly}(DSA)$ steps and output $\widehat{D}$ satisfying $D \leq \widehat{D} \leq 2D$, or it will not terminate, in which case we set $\widehat{D} = \infty$. For the remainder of the time steps, we run Algorithm 1 with $H = \min\{\widehat{D}, \sqrt{T/(S^3 A)}\}$, recovering the same bound as Corollary 7 for stochastic MDPs and a bound of $\widetilde{O}(\min\{\text{poly}(DSA), \sqrt{SAT} + \|h^\star\|_{\text{sp}}^2 S^3 A\})$ for deterministic MDPs, which is T-independent for strongly communicating deterministic MDPs, and still finite when $D = \infty$.

3.5. Lower Bounds for Average-Reward Regret

We now turn to lower bounds on the average-reward regret. Prior work in Auer et al. (2008) establishes a regret lower bound of $\Omega(\sqrt{\|h^\star\|_{\text{sp}} SAT})$ when $T \geq \|h^\star\|_{\text{sp}} SA$,[2] which has been matched by several algorithms (Zhang and Ji 2019; Boone and Zhang 2024; our Corollaries 6 and 7) when the horizon T is sufficiently large. In contrast to the large-T regime, here we focus on lower bounds applicable for all T, in order to characterize the optimal burn-in cost of any algorithm.

We begin by formalizing the definition of algorithms to which our lower bounds apply. We define a *horizon-T algorithm* `Alg` to be a function from histories of length $\leq T$ to a distribution over actions, that is, a function $\bigcup_{0 \leq t \leq T} \left(\mathcal{S} \times (\mathcal{A} \times \mathcal{S})^t\right) \to \Delta(\mathcal{A})$. We note that as defined, such an algorithm only takes as input a sequence of elements of $\mathcal{S}$ and $\mathcal{A}$; intuitively speaking, any other data, such as the value of T or prior knowledge of $\|h^\star\|_{\text{sp}}$, must already be "baked in" to the algorithm. Hence by our definition, an "algorithm" (for horizon T) with prior knowledge of $\|h^\star\|_{\text{sp}}$ is actually a family of horizon-T algorithms, one for each value of $\|h^\star\|_{\text{sp}}$.

The main theorem of this section is a lower bound on lower-order terms incurred by any algorithm without prior knowledge of $\|h^\star\|_{\text{sp}}$. This also implies a lower bound on the burn-in cost.

Theorem 8 (Burn-In Lower Bound For Prior-Free Algorithms) *There is a universal constant $c \geq 1$ such that the following holds. Let $S \geq 2$ and $A \geq 2$ be integers. Fix $\alpha \in [1, 2)$ and a function $t \mapsto \beta_t$. Suppose that $T > SA(c\beta_T)^{\frac{4}{2-\alpha}}$ and $\beta_T \geq 1$. Then there exist two communicating MDPs P_1 and P_2, each with S states and A actions, such that no horizon-T algorithm `Alg` can satisfy $\mathbb{E}[\text{Regret}(T, P_i, \texttt{Alg})] \leq \sqrt{\beta_T \left\|h^\star_{P_i}\right\|_{\text{sp}} SAT} + \beta_T SA \left\|h^\star_{P_i}\right\|_{\text{sp}}^\alpha$ for both $i = 1$ and $i = 2$.*

We provide a proof sketch in Section 4.3 and a complete proof in Appendix K. Here we intend β_T to be used to encapsulate $\widetilde{O}(1)$ terms; Theorem 8 states that for any $\alpha < 2$, we can find T such that

2. This lower bound was originally stated in terms of the diameter D, but for their hard instances $\|h^\star\|_{\text{sp}}$ and D differ by only a constant factor.

no single horizon-T algorithm can enjoy regret bounds of the form $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}} SAT} + \|h^\star\|_{\text{sp}}^\alpha SA)$ simultaneously for two certain MDPs P_1 and P_2. The two MDPs have $\|h^\star_{P_1}\|_{\text{sp}} \gg \|h^\star_{P_2}\|_{\text{sp}}$. With prior knowledge, formally a different horizon-T algorithm would be applied to each MDP P_1, P_2, and furthermore, we would not expect a horizon-T algorithm designed for MDPs with bias span $\leq \|h^\star_{P_2}\|_{\text{sp}}$ to enjoy a nonvacuous regret bound when deployed on P_1. So in short, Theorem 8 is not a counterexample to the type of theorem that one proves when designing algorithms that use prior knowledge, and in particular does not contradict our Corollary 6. However, this lower bound does prohibit any minimax-optimal (for large T) algorithm without prior knowledge from obtaining a better $\|h^\star\|_{\text{sp}}$ dependence in its lower-order terms than $\|h^\star\|_{\text{sp}}^2$. This is matched by our prior-knowledge-free Corollary 7.

Theorem 8 demonstrates a "price of adaptivity" for the burn-in cost, that is, a gap between what is achievable with and without prior knowledge. In particular, this gap is established by combining the above lower bound and the strictly smaller regret upper bound provided by our prior-knowledge-based Corollary 6, which achieves $\widetilde{O}(\sqrt{\|h^\star\|_{\text{sp}} T} + \|h^\star\|_{\text{sp}})$ when applied to instances with $S, A \leq O(1)$. Note that previous results are insufficient for establishing this gap, as the only other algorithm which uses prior knowledge and achieves minimax-optimal regret for large T has a burn-in cost scaling with $\|h^\star\|_{\text{sp}}^6$ even when $S, A \leq O(1)$ (Zhang and Ji, 2019). The only result of a similar nature for the average-reward regret of which we are aware is Fruit et al. (2019, Lemma 3), which shows an exponential lower bound on the burn-in cost but only for algorithms achieving a logarithmic regret. One particularly interesting feature of the gap implied by Theorem 8 is that this contrasts the simulator (a.k.a. generative model) setting, where recent work has characterized the sample complexity in terms of $\|h^\star\|_{\text{sp}}$ and demonstrated no gap between algorithms which do and do not possess prior knowledge (Wang et al., 2022; Zurek and Chen, 2025b,a).

Next, we show a burn-in cost lower bound applicable even to algorithms with prior knowledge.

Theorem 9 (General Burn-In Lower Bound) *Let $S \geq 2$ and $A \geq 2$ be integers, and let $D \geq 4\lceil \log_A S \rceil$. For any horizon-$T$ algorithm $\texttt{Alg}$, there exists an MDP P with S states, A actions, and diameter at most D such that for all $T \leq \frac{1}{32} DSA$, $\mathbb{E}[\text{Regret}(T, P, \texttt{Alg})] \geq \frac{T}{4}$.*

The proof uses standard constructions and is deferred to Appendix J. We emphasize that this result applies to any horizon-T algorithm. Since $D \geq \|h^\star\|_{\text{sp}}$ (in fact they are equal up to a constant factor for this MDP), Theorem 9 implies that any algorithm with a sublinear-in-T regret bound must have a burn-in requirement of $T \geq \Omega(\|h^\star\|_{\text{sp}} SA)$. In particular, combining with the lower bound from Auer et al. (2008), we see that no algorithm can have regret below $\Omega(\sqrt{\|h^\star\|_{\text{sp}} SAT} + \|h^\star\|_{\text{sp}} SA)$, matching our Corollary 6 up to an additional S in the additive burn-in term and $\widetilde{O}(1)$ factors. We conjecture that this lower bound is nearly tight and that the factor of S in our upper bound could be removed, although we believe this may be challenging and the techniques used to do so in the inhomogeneous episodic setting (Zhang et al., 2024) would not apply in the infinite horizon case.

4. Technical Highlights

In this section, we discuss our algorithmic and analytical contributions in the context of related work. We also include a proof sketch for Theorem 8.

4.1. Algorithmic Improvements Over Prior UCB-based Approaches

Our algorithm builds on prior model-based algorithms for the discounted setting, particularly UCBVI-γ (He et al., 2021) and γ-UCB-CVI (Hong et al., 2025), but introduces several crucial modifications that enable variance-dependent bounds and eliminate extraneous dependence on $\frac{1}{1-\gamma}$. One main novelty is how the optimistic Q-estimate is updated. Previous algorithms initialize the estimate using $\widehat{Q}_1 \leftarrow \frac{1}{1-\gamma}\mathbf{1}$, and then at each time step $t \in [T]$ perform the following one-step value iteration:

$$\widehat{Q}_{t+1}(s,a) \leftarrow r(s,a) + \gamma P^t_{s,a}\widehat{V}_t + \gamma b_t(s,a).$$

Here, $\widehat{V}_t = M\widehat{Q}_t$ (in UCBVI-γ) or $\widehat{V}_t = \text{Clip}_H\big(M\widehat{Q}_t\big)$ (in γ-UCB-CVI). We first discuss the bonus term $b_t(s,a)$. UCBVI-γ uses a Bernstein-style bonus similar to that of the UCBVI algorithm for the episodic setting (Azar et al., 2017). UCBVI obtains a regret bound which is optimal in the leading term but suboptimal in lower-order terms. This suboptimality is due to an extra term in the bonus; indeed, in the episodic setting, Zhang et al. (2021) show a similar term to be unnecessary for optimism. Hence, UCBVI-γ incurs suboptimal lower-order terms for the same reason. The MVP algorithm (Zhang et al., 2021) removes this extra term from the bonus to achieve significantly improved lower order terms in the episodic setting.

Secondly, as mentioned above, γ-UCB-CVI uses a clipping step to ensure $\|\widehat{V}_t\|_{\text{sp}} \le H$. Their analysis has steps that involve upper bounding $\|\widehat{V}_t\|_{\text{sp}}$, and clipping allows one to replace some factors of $\frac{1}{1-\gamma}$ with H. However, γ-UCB-CVI uses a Hoeffding-style bonus, resulting in a suboptimal leading term. Still, the result of clipping is that the leading term of their γ-regret bound depends on H instead of $\frac{1}{1-\gamma}$. Combining clipping with the sharp Bernstein-style bonus would yield a γ-regret of $\widetilde{O}\big(\sqrt{\|V^\star_\gamma\|_{\text{sp}}SAT} + HS^2A + \frac{SA}{1-\gamma}\big)$, which improves the state-of-the-art but retains a lower-order factor of $\frac{1}{1-\gamma}$. Applying our average-to-discounted reduction (Lemma 4) would then require $\gamma = 1 - \sqrt{HT/(SA)}$. The $\widetilde{O}\big(\sqrt{HSAT} + HS^2A\big)$. This bound is not variance-dependent, and moreover the leading term depends on the tuning parameter H instead of $\|h^\star\|_{\text{sp}}$, which means the algorithm would only be optimal with prior knowledge of $\|h^\star\|_{\text{sp}}$.

The core issue with one-step updates is that the estimate $\widehat{Q}_{t+1}$ may remain significantly larger than the fixed point of $\widehat{\mathcal{T}}_t$, especially for small t or γ close to 1. In particular, value iteration requires on the order of $\frac{1}{1-\gamma}$ steps to approximately converge. On the other hand, the statistical error converges to 0 at an unrelated and potentially much faster rate, especially for low-variance MDPs or average-reward settings where γ is tuned to be very large. To address this issue, our algorithm FOCUS fully optimizes the Q-estimate at the beginning of each episode k. Concretely, the algorithm iteratively applies the empirical Bellman operator $\widehat{\mathcal{T}}_k$ until convergence, producing an estimate $\widehat{Q}_k$ that fully exploits all data collected to that point. This mechanism of fully optimizing is also a feature of algorithms that utilize the EVI subroutine, suggesting that full exploitation of available data is crucial for achieving optimal span-based bounds in the average-reward setting. With this strategy, we obtain a γ-regret bound without dependence on $\frac{1}{1-\gamma}$, which allows us to solve the average-reward problem with a simple, UCB-based approach (see Appendix B for further comparison between FOCUS and EVI-based algorithms).

4.2. How Full Optimization Helps in Regret Analysis

A key technical challenge that illustrates the necessity of fully optimizing lies in controlling $\mathscr{T}_{\text{ind}}$, a term in our regret decomposition which accounts for changes in the value estimate along the learner's trajectory. To see this, we first note that the prior work in Hong et al. (2025) bounds an analogous quantity in the analysis of γ-UCB-CVI according to

$$\sum_{t=1}^{T}\left(\widehat{V}_{t-1}(s_{t+1})-\widehat{V}_t(s_t)\right)\leq\sum_{t=1}^{T}\left(\widehat{V}_{t-1}(s_{t+1})-\widehat{V}_{t+1}(s_{t+1})\right)+\frac{1}{1-\gamma}$$
$$\leq\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\left(\widehat{V}_{t-1}(s)-\widehat{V}_{t+1}(s)\right)+\frac{1}{1-\gamma}\leq O\left(\frac{S}{1-\gamma}\right).$$

Here, the second inequality holds because their value estimates are monotonically decreasing with t, a property they enforce by taking minimums. These calculations show that the scale of this term depends on the possible range of value estimates, which, under one-step value iteration updates, can be as large as $\frac{1}{1-\gamma}$. In contrast, by fully running the value iteration procedure, along with the use of clipping, our value estimates lie in a range of H times the number of episodes, which is only logarithmic in T due to the doubling trick. Formally, letting m be the total number of episodes and t_k be the time at the start of the kth episode, a telescoping argument yields

$$\mathscr{T}_{\text{ind}}=\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\widehat{V}_k(s_{t_{k+1}})-\widehat{V}_k(s_{t_k}\right)\leq\sum_{k=1}^{m}\left(\widehat{V}_k(s_{t_{k+1}})-\widehat{V}_k(s_{t_k}\right)\leq mH\leq\widetilde{O}(HSA).$$

4.3. Proof Sketch for Theorem 8

We sketch the construction underlying Theorem 8, which shows that without prior knowledge of the bias span, any algorithm must incur a burn-in cost of order $\|h^\star\|_{\text{sp}}^2 SA$. Consider the MDPs P_1 and P_2 in Figure 1. They differ only in the `stay` action at state 2. In both, state 1 has `stay` (self-loop, reward 1/2) and `leave` (transits to state 2 with probability $1/B$, reward 0), and state 2 has `leave` (transits to state 1, reward 0). Now, in both MDPs `stay` in state 2 has reward 1, but in P_1 it is a self-loop while in P_2 it is a deterministic transition to state 1. It follows that in P_1 the optimal gain is 1, obtained by reaching and staying in state 2. On the other hand, in P_2 state 2 offers no long-term benefit, so the optimal gain is 1/2, obtained by staying in state 1. Additionally, a direct calculation confirms that the span of the optimal bias function in P_1 is B (reflecting the long delay before being able to collect reward 1), while that of P_2 is 1/2.

Suppose a prior-less algorithm promises a sublinear regret on P_1. The algorithm must reach state 2; otherwise it would incur at least regret 1/2 per time step. Moreover, reaching state 2 requires taking the `leave` action B times in expectation. On P_2, however, these exploration attempts are wasteful since `leave` yields reward 0 with no potential to recoup reward in state 2. Yet, since P_1 and P_2 only differ on the `stay` action in state 2, even when the true MDP is P_2, the algorithm must still reach state 2 and collect data there to guard against the possibility that the MDP is actually P_1. Note that this argument is only valid without prior knowledge, because knowing that the bias span is at most $1/2$, one could eliminate the possibility of P_1 without reaching state 2 since $\|h^\star_{P_2}\|_{\text{sp}} > 1/2$. Thus, any prior-less algorithm which achieves less than $T/2$ regret on P_1 must incur $\Omega(B)$ regret on P_2. We summarize the preceding argument in the following intermediate result.

Lemma 10 (Simplified Version of Theorem 26) *There is a constant $c \in (0,1)$ so that the following holds. Fix T, B, and let `Alg` be any horizon-T algorithm. There exist MDPs P_1 and P_2 such that $\|h^\star_{P_1}\|_{\text{sp}} = B$, $\|h^\star_{P_2}\|_{\text{sp}} = 1/2$, and $\mathbb{E}_{P_1}[\text{Regret}(T)] < T/4 \implies \mathbb{E}_{P_2}[\text{Regret}(T)] \geq cB$.*

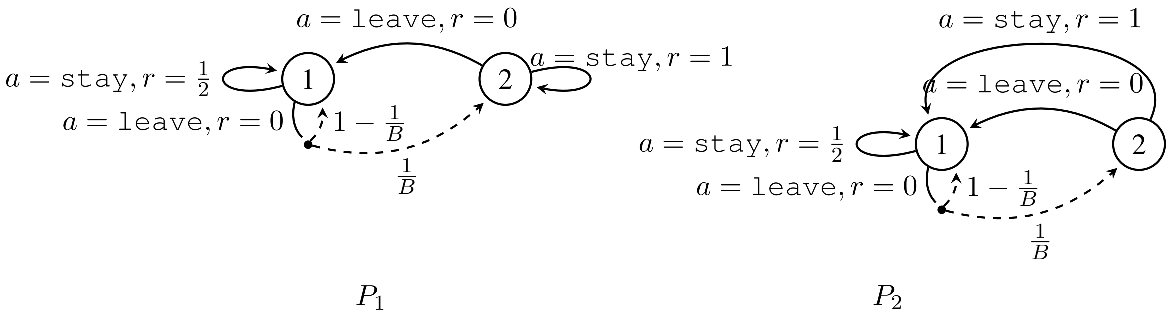

Figure 1: An example of the MDPs used in the proof of Theorem 8. Each state-action pair is annotated with its reward. A deterministic transitions is represented by a solid arrow. A stochastic transition is represented by a solid line splitting into dashed arrows to different states, each annotated with the associated transition probability. The MDPs are parameterized by $B > 2$, both have starting state 1, and differ only in the transition distribution of the `stay` action of state 2. In P_1 an optimal stationary policy traverses to state 2 and stays there, while in P_2 an optimal stationary policy remains in state 1.

Ignoring S and A factors, Theorem 8 follows from Lemma 10 by appropriately tuning B. Fix arbitrary $\alpha \in [1, 2)$ and $\beta \geq 1$. For a sufficiently large T suppose there exists a horizon-T algorithm that obtains a $\beta\big(\sqrt{\|h^\star\|_{\text{sp}}T} + \|h^\star\|_{\text{sp}}^{\alpha}\big)$ regret bound in expectation for any MDP with $SA = O(1)$. Choosing $B = \frac{\beta}{c}\sqrt{T}$ and letting MDPs P_1 and P_2 be as in Lemma 10, the algorithm simultaneously obtains regret less than $T/4$ on P_1 and less than cB on P_2 — a contradiction.

Note that he role of no prior knowledge is implicit but crucial. The contradiction arises because a single horizon-T algorithm must handle both MDPs simultaneously. Prior work (Fruit et al., 2018) shows how prior knowledge allows the learner to aggressively utilize span-clipping and exploit earlier. Our hard instances illustrate that without prior knowledge, such aggressive exploitation is impossible in general, and the algorithm must explore significantly longer to perform optimally in instances with large bias spans, resulting in worse burn-in cost in instances with small bias spans.

5. Conclusion

We developed the first algorithm for both average-reward regret and γ-regret that is simultaneously minimax-optimal and variance-dependent. Our average-reward regret bounds have optimal lower-order dependence on $\|h^\star\|_{\text{sp}}$, and we proved lower bounds which revealed a fundamental gap in what is achievable with and without prior knowledge. One open problem is to eliminate the Γ factor from the lower-order terms of Theorems 3 and 5, which has recently been done in the inhomogeneous episodic setting (Zhang et al., 2024) but appears more challenging in infinite-horizon settings.

Acknowledgments

G. Zamir, M. Zurek, and Y. Chen acknowledge support by National Science Foundation (NSF) grants CCF-2233152 and DMS-2023239, by a Cisco Systems Fellowship, and by a Vilas Associates Award. Part of this research was performed while M. Zurek was visiting the Institute for Mathematical and Statistical Innovation (IMSI) supported by the NSF (Grant No. DMS-2425650).

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://papers.nips.cc/paper_files/paper/2008/hash/e4a6222cdb5b34375400904f03d8e6a5-Abstract.html.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. PMLR, July 2017. URL https://proceedings.mlr.press/v70/azar17a.html. ISSN: 2640-3498.

Peter L. Bartlett and Ambuj Tewari. REGAL: A Regularization based Algorithm for Reinforcement Learning in Weakly Communicating MDPs, May 2012. URL http://arxiv.org/abs/1205.2661. arXiv:1205.2661.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward mdps. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 26728–26769. Curran Associates, Inc., 2024. doi: 10.52202/079017-0840. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f0bb736ccc8551ef5bcc9165c2a4d9e-Paper-Conference.pdf.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning, July 2018. URL http://arxiv.org/abs/1802.04020. arXiv:1802.04020 [cs, stat].

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes, March 2019. URL http://arxiv.org/abs/1807.02373. arXiv:1807.02373 [cs, stat].

Germano Gabbianelli, Gergely Neu, Nneka Okolo, and Matteo Papini. Offline Primal-Dual Reinforcement Learning for Linear MDPs, May 2023. URL http://arxiv.org/abs/2305.12944. arXiv:2305.12944 [cs].

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly Minimax Optimal Reinforcement Learning for Discounted MDPs. In *Advances in Neural Information Processing Systems*, volume 34, pages 22288–22300. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bb57db42f77807a9c5823bd8c2d9aaef-Abstract.html.

Kihyuk Hong, Woojin Chae, Yufan Zhang, Dabeen Lee, and Ambuj Tewari. Reinforcement Learning for Infinite-Horizon Average-Reward Linear MDPs via Approximation by Discounted-Reward MDPs, March 2025. URL http://arxiv.org/abs/2405.15050. arXiv:2405.15050 [stat].

Xiang Ji and Gen Li. Regret-Optimal Model-Free Reinforcement Learning for Discounted MDPs with Short Burn-In Time. *Advances in Neural Information Processing Systems*, 36:80674–80689, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ff887781480973bd3cb6026feb378d1e-Abstract-Conference.html.

Yujia Jin and Aaron Sidford. Efficiently Solving MDPs with Stochastic Mirror Descent, August 2020. URL http://arxiv.org/abs/2008.12776. arXiv:2008.12776.

Yujia Jin and Aaron Sidford. Towards Tight Bounds on the Sample Complexity of Average-reward MDPs, June 2021. URL http://arxiv.org/abs/2106.07046. arXiv:2106.07046 [cs, math].

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-Optimal Sample Complexity for MDPs via Anchoring, June 2025. URL http://arxiv.org/abs/2502.04477. arXiv:2502.04477 [math].

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic First-Order Methods for Average-Reward Markov Decision Processes. *Mathematics of Operations Research*, December 2024. ISSN 0364-765X. doi: 10.1287/moor.2022.0241. URL https://pubsonline.informs.org/doi/full/10.1287/moor.2022.0241.

Shuang Liu and Hao Su. Regret Bounds for Discounted MDPs, May 2021. URL http://arxiv.org/abs/2002.05138. arXiv:2002.05138 [cs].

Jianfei Ma and Wee Sun Lee. Eubrl: Epistemic uncertainty directed bayesian reinforcement learning, 2026. URL https://arxiv.org/abs/2512.15405.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization, July 2009. URL http://arxiv.org/abs/0907.3740. arXiv:0907.3740 [stat].

Gergely Neu and Nneka Okolo. Dealing with unbounded gradients in stochastic saddle-point optimization, June 2024. URL http://arxiv.org/abs/2402.13903. arXiv:2402.13903 [cs, math, stat] version: 2.

Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Offline Reinforcement Learning via Linear-Programming with Error-Bound Induced Constraints, December 2024. URL http://arxiv.org/abs/2212.13861. arXiv:2212.13861 [cs].

Charles Chapman Pugh. *Real mathematical analysis*. Undergraduate texts in mathematics. Springer, Cham Heidelberg, 2. ed edition, 2015. ISBN 978-3-319-17770-0.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/talebi18a.html.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7611–7624. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3e98410c45ea98addec555019bbae8eb-Paper.pdf.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward markov decision processes without prior knowledge. In *Advances in Neural Information Processing Systems*, volume 37, pages 109948–109979, 2024.

Jinghan Wang, Mengdi Wang, and Lin F. Yang. Near Sample-Optimal Reduction-based Policy Learning for Average Reward MDP, December 2022. URL http://arxiv.org/abs/2212.00603. arXiv:2212.00603 [cs].

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal Sample Complexity for Average Reward Markov Decision Processes, February 2024. URL http://arxiv.org/abs/2310.08833. arXiv:2310.08833.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10170–10180. PMLR, November 2020. URL https://proceedings.mlr.press/v119/wei20c.html. ISSN: 2640-3498.

Andrea Zanette and Emma Brunskill. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds, November 2019. URL http://arxiv.org/abs/1901.00210. arXiv:1901.00210 [cs, stat] version: 4.

Zihan Zhang and Xiangyang Ji. Regret Minimization for Reinforcement Learning by Evaluating the Optimal Bias Function. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/9e984c108157cea74c894b5cf34efc44-Abstract.html.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5476–5477. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/zhang23b.html.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 4528–4531. PMLR, July 2021. URL https://proceedings.mlr.press/v134/zhang21b.html. ISSN: 2640-3498.

Zihan Zhang, Yuxin Chen, Jason D. Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning. In *Proceedings of Thirty Seventh Conference on Learning Theory*, pages 5213–5219. PMLR, June 2024. URL https://proceedings.mlr.press/v247/zhang24a.html. ISSN: 2640-3498.

Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments. In *Proceedings of the 40th International Conference on Machine Learning*, pages 42878–42914. PMLR, July 2023. URL https://proceedings.mlr.press/v202/zhou23t.html. ISSN: 2640-3498.

Matthew Zurek and Yudong Chen. The Plug-in Approach for Average-Reward and Discounted MDPs: Optimal Sample Complexity Analysis, October 2024. URL http://arxiv.org/abs/2410.07616. arXiv:2410.07616 [cs].

Matthew Zurek and Yudong Chen. Span-Agnostic Optimal Sample Complexity and Oracle Inequalities for Average-Reward RL. In *Proceedings of Thirty Eighth Conference on Learning Theory*, pages 6156–6209. PMLR, July 2025a. URL https://proceedings.mlr.press/v291/zurek25a.html.

Matthew Zurek and Yudong Chen. Span-Based Optimal Sample Complexity for Weakly Communicating and General Average Reward MDPs. *Advances in Neural Information Processing Systems*, 37:33455–33504, January 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3acbe9dc3a1e8d48a57b16e9aef91879-Abstract-Conference.html.

Matthew Zurek, Guy Zamir, and Yudong Chen. Optimal single-policy sample complexity and transient coverage for average-reward offline RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MjOf5qnEX7.

Appendix A. More Related Work

Here we discuss additional related work.

Average-Reward Simulator and Offline Settings A complementary problem setting to online RL is the offline/simulator setting, where the goal is to learn a ε-optimal policy π (such that $\rho^\pi(s) \geq \rho^\star(s) - \varepsilon$ for all $s \in \mathcal{S}$) from a fixed/simulator-generated dataset with the minimum number of samples. A sequence of works obtained sharper sample complexity bounds with relaxed assumptions on the environmental structure (e.g. Jin and Sidford 2020, 2021; Wang et al. 2022, 2024; Li et al. 2024), culminating with the optimal $\|h^\star\|_{\mathrm{sp}}$-based sample complexity bound of $\widetilde{O}(SA\frac{\|h^\star\|_{\mathrm{sp}}}{\varepsilon^2})$ shown by Zurek and Chen (2025b), matching a lower bound due to Wang et al. (2022). However, this result required prior knowledge of $\|h^\star\|_{\mathrm{sp}}$, leaving the question open of whether the optimal sample complexity could be obtained by algorithms without prior knowledge. After extensive research effort (Neu and Okolo, 2024; Tuynman et al., 2024; Zurek and Chen, 2024; Lee et al., 2025), this question was answered affirmatively by Zurek and Chen (2025a). Tuynman et al. (2024) and Zurek and Chen (2025b) show various hardness results related to estimating $\|h^\star\|_{\mathrm{sp}}$ and $\|h^\star\|_{\mathrm{sp}}$-based PAC guarantees with online environment access. A very common approach throughout these works is to reduce the average-reward problem to a discounted one; obtaining sharp variance bounds for the discounted problem, somewhat analogous to our Lemma 2, plays a key role in all minimax-optimal approaches (Zurek and Chen, 2025b, 2024, 2025a). All of the aforementioned work is for the generative model setting, where a dataset with uniform coverage can be sampled. Recent works have also studied offline settings with more general data sampling patterns (Gabbianelli et al., 2023; Ozdaglar et al., 2024; Zurek et al., 2025).

Episodic Online RL Studying the theoretical limits of regret for episodic online RL has been one of the most fundamental problems in RL theory. Hence, we cannot provide a comprehensive review of work on this topic, but we discuss a few related works with strong connections to our own. Azar et al. (2017) establishes minimax regret bounds in the episodic setting with the UCBVI algorithm, which uses a Bernstein-style bonus. Zanette and Brunskill (2019) developed the EULER algorithm which obtains both minimax-optimal and variance-dependent regret bounds. Zhang et al. (2021) greatly improves lower order terms by using an even sharper Bernstein-style bonus in their MVP algorithm. Later refinements to the MVP algorithm in Zhou et al. (2023) and Zhang et al. (2024) yield optimal lower order terms and variance-dependent regret bounds that adapt to the difficulty of the environment.

Appendix B. Comparison of FOCUS to EVI-Based Approaches

The two existing minimax-optimal algorithms for the average-reward setting, EBF (Zhang and Ji, 2019) and PMEVI-DT (Boone and Zhang, 2024), are both based on EVI. The way that these algorithms refine previous EVI-based approaches suggest that obtaining an optimal span-dependent regret requires exploiting structural information encoded in the optimal bias function $h^\star$. We elaborate on this point by examining the state-of-the-art PMEVI-DT algorithm, whose name reflects two central modifications to standard EVI. We remark that EBF is an earlier attempt in this direction, in that it estimates bias differences to shrink the confidence set, but incorporates this restriction through a step that is not efficiently computable.

At each episode, PMEVI-DT runs a "BiasEstimation" subroutine to construct a confidence set for $h^\star$. The extended Bellman operator used in EVI is then combined with a "projection" step (the P in PMEVI-DT) that constrains the possible models to those with optimal bias in the confidence set. Their extended Bellman operator also incorporates a "mitigation" step (the M in PMEVI-DT) that uses the bias confidence region to tighten a Bernstein-type variance constraint. Despite this machinery, PMEVI-DT still requires either prior knowledge of $\|h^\star\|_{\text{sp}}$ or a condition like $T \geq \|h^\star\|_{\text{sp}}^5$ to achieve minimax-optimal regret.

Our results show that under the same type of assumptions, our UCB-based algorithm sufficiently exploits $h^\star$ without explicitly estimating it. The span-clipping component of the empirical Bellman update replaces the projection and mitigation steps of PMEVI-DT with a simple, easily interpretable operation. Specifically, span-clipping prevents the value estimate from being overly optimistic; a smaller clipping threshold H reduces exploration and increases immediate exploitation of known high-reward actions. This span-clipping technique was introduced by Fruit et al. (2018) as part of the EVI-based algorithm SCAL. While SCAL obtains a span-based regret bound, it does not maintain sharp confidence regions and consequently its regret suffers from extra factors of $\|h^\star\|_{\text{sp}}$ and S. The more involved bias estimation techniques of EBF and PMEVI-DT circumvent these issues and produce sharp confidence regions with bounded bias spans, but our algorithm achieves optimal regret bounds with the simpler combination of span-clipping and a sharp Bernstein-style bonus.

Finally, given the success of UCB-based approaches in the episodic setting, we remark that it is at least somewhat surprising that such algorithms have yet to be thoroughly studied in the infinite-horizon average-reward setting. We suggest two contributing factors. First, our approach utilizes an average-to-discounted reduction, a strategy which only recently has been shown to yield optimal span-based bounds (Zurek and Chen, 2025b,a). In these works which derive an optimal span-based bound via an average-to-discounted reduction, a crucial step is tight analysis of variance-dependent quantities to remove factors of $\frac{1}{1-\gamma}$. The analogous step in our results is Lemma 2, which was vital in successfully applying the reduction. Secondly, since the seminal work of Auer et al. (2008), the most well-studied and successful algorithms for the online infinite-horizon average-reward setting have been EVI-based, so the most natural route for obtaining a minimax-optimal algorithm was to refine these existing works.

Appendix C. γ-Regret Bound of γ-UCB-CVI

In this section we show how the analysis of γ-UCB-CVI in Hong et al. (2025) can be modified slightly to obtain a span-based γ-regret bound. The quantities P^t, V_t and N_t refer to the empirical transition kernel, value estimate, and empirical state-action counts, respectively, of γ-UCB-CVI at time t. $\beta = \widetilde{\Theta}(\left\|V_\gamma^\star\right\|_{\text{sp}} \sqrt{S})$ is a constant in the bonus term which is large enough to ensure optimism. Additionally, in their analysis they often immediately bound $\left\|V_\gamma^\star\right\|_{\text{sp}} \leq 2 \left\|h^\star\right\|_{\text{sp}}$ and end up with factors of $\left\|h^\star\right\|_{\text{sp}}$ in intermediate steps. Below, we leave in the factors of $\left\|V_\gamma^\star\right\|_{\text{sp}}$.

Via a concentration argument on $|(P^{t-1}_{s_t,a_t} - P_{s_t,a_t})V_{t-1}|$ (their Lemma 3), Hong et al. obtains that with high probability

$$r(s_t, a_t) \geq V_t(s_t) - \gamma P_{s_t,a_t} V_{t-1} - \frac{2\beta}{\sqrt{N_{t-1}(s_t, a_t)}}.$$

Subsequently, Hong et al. decompose the average-reward regret as

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^T (\rho^\star - r(s_t, a_t)) \\
&\leq \sum_{t=1}^T \left(\rho^\star - V_t(s_t) + \gamma P_{s_t,a_t} V_{t-1} + \frac{2\beta}{\sqrt{N_{t-1}(s_t, a_t)}} \right) \\
&= \underbrace{\sum_{t=1}^T (\rho^\star - (1-\gamma) V_t(s_t))}_{(a)} \quad + \quad \underbrace{\gamma \sum_{t=1}^T (V_{t-1}(s_{t+1} - V_t(s_t)))}_{(b)} \\
&\qquad\qquad + \underbrace{\gamma \sum_{t=1}^T (P_{s_t,a_t} V_{t-1} - V_{t-1}(s_{t+1}))}_{(c)} \quad + \quad \underbrace{2\beta \sum_{t=1}^T \frac{1}{\sqrt{N_{t-1}(s_t, a_t}}}_{(d)}.
\end{aligned}$$

To instead analyze $\text{Regret}_\gamma(T) = \sum_t ((1-\gamma) V_\gamma^\star(s_t) - r(s_t, a_t)$, we simply replace $\rho^\star$ with $(1-\gamma) V_\gamma^\star(s_t)$. With this replacement, term (a) vanishes due to optimism (their Lemma 4). For the other terms, the bounds in the original proof still hold. In particular, term (b) is bounded by $O(\frac{S}{1-\gamma})$, term

(c) is bounded by $\widetilde{O}(\left\|V_\gamma^\star\right\|_{\text{sp}}\sqrt{T})$, and term (d) is bounded by $O(\beta\sqrt{SAT}) \le \widetilde{O}(\left\|V_\gamma^\star\right\|_{\text{sp}} S\sqrt{AT})$. Recombining terms, we obtain that with high probability,

$$\text{Regret}_\gamma(T) \le \widetilde{O}\left(\left\|V_\gamma^\star\right\|_{\text{sp}} S\sqrt{AT} + \frac{S}{1-\gamma}\right).$$

Appendix D. Technical Lemmas

Lemma 11 (Bernstein's Inequality, Theorem 3 in Maurer and Pontil 2009) *Let $Z, Z_1, \dots, Z_n$ be i.i.d. random variables with values in $[c_{\min}, c_{\max}]$ for some constants $c_{\min} < c_{\max}$. Set $c = c_{\max} - c_{\min}$. Let $\delta > 0$. Then we have with probability at least $1-\delta$ that*

$$\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i\right| \le \sqrt{\frac{2\text{Var}(Z)\log(2/\delta)}{n}} + \frac{c\log(2/\delta)}{3n}.$$

Lemma 12 (Empirical Bernstein's Inequality, Theorem 4 in Maurer and Pontil 2009) *Let Z, $Z_1, \dots, Z_n$ be i.i.d. random variables with values in $[c_{\min}, c_{\max}]$ for some constants $c_{\min} < c_{\max}$. Set $c = c_{\max} - c_{\min}$. Denote $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\widehat{\text{Var}}_n = \frac{1}{n}\sum_{i=1}^n \left(Z_i - \overline{Z}\right)^2$. Let $\delta > 0$. Then we have with probability at least $1-\delta$ that*

$$\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i\right| \le \sqrt{\frac{2\widehat{\text{Var}}_n\log(2/\delta)}{n-1}} + \frac{7c\log(2/\delta)}{3(n-1)}.$$

Lemma 13 (Lemma 22 in Zhou et al. 2023) *For any two nonnegative constants c_1, c_2 satisfying $2c_1^2 \le c_2$, let $f : \Delta([S]) \times [0, 2C]^S \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be defined by*

$$f(p, v, n, u) = pv + \max\left\{c_1\sqrt{\frac{\mathbb{V}(p, v)u}{n}}, c_2\frac{Cu}{n}\right\}.$$

Then for all $p \in \Delta([S]), v \in [0, 2C]^S$, and $n, u > 0$, f is non-decreasing in v, i.e.

$$f(p, v, n, u) \ge f(p, v', n, u) \quad \forall v, v' \in [0, 2C]^S \text{ satisfying } v \ge v'.$$

Lemma 14 (Lemma 19 in Zhou et al. 2023) *Let X be a random variable with $\sup|X| \le c$ for some constant $c \ge 0$. Then*

$$\text{Var}\left(X^2\right) \le 4c^2\text{Var}(X).$$

Lemma 15 *If $x \le a\sqrt{x} + b$ for $a, b > 0$, then $x \le 2a^2 + 2b$.*

Lemma 16 (Bernoulli's Inequality) *Let $r \ge 1$ and $x \ge -1$. Then $(1+x)^r \ge 1 + rx$.*

Lemma 17 (Rearrangement Inequality) *Let $x_1 \le \dots \le x_n$ and $y_1 \le \dots \le y_n$ be real numbers. For every permutation σ of $1, \dots, n$, we have*

$$x_1y_n + \dots + x_ny_1 \le x_1y_{\sigma(1)} + \dots + x_ny_{\sigma(n)} \le x_1y_1 + \dots + x_ny_n.$$

Appendix E. Proof of Theorem 3

We provide an overview of notation used in the proof. We let $\mathbb{I}$ denote the indicator function, meaning that for an event $\mathcal{E}$, we have $\mathbb{I}(\mathcal{E}) = 1$ if $\mathcal{E}$ holds and $\mathbb{I}(\mathcal{E}) = 0$ otherwise. We denote by m the number of episodes. For each $k \in [m]$, we write t_k to denote the time at the start of the kth episode, and we set $t_{m+1} = T + 1$. Analogously, for each $t \in [T]$, we write k_t to denote the current episode at time t. $\widehat{Q}_k$ is the Q-estimate used during the kth episode, and $\widehat{V}_k = \text{Clip}_H(MQ)$. We will frequently use the fact that $\|\widehat{V}_k\|_{\text{sp}} \leq H$ for all $k \in [m]$. $N_k(s, a, s')$ and $N_k(s, a)$ are the counts of (s, a, s') and (s, a), respectively, at the start of the kth episode, and we let N_k denote the length of the kth episode. We write $n_k(s, a)$ as shorthand for $\max\{N_k(s, a), 1\}$. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $V \in \mathbb{R}^{\mathcal{S}}$, and $k \in [m]$, we let $b_k(s, a, V)$ denote the bonus used for the kth episode, namely

$$b_k(s, a, V) := \max\left\{4\sqrt{\frac{\mathbb{V}\left(\widehat{P}^k_{s,a}, V\right)U}{n_k(s, a)}}, 32\frac{HU}{n_k(s, a)}\right\}.$$

For any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $V \in \mathbb{R}^{\mathcal{S}}$ we define

$$\mathbb{V}_{s,a}(V) := \mathbb{V}(P_{s,a}, V).$$

Further recall that $\delta' = \frac{\delta}{9S^2AT}$ and $U = \log\left(\frac{1}{\delta'}\right)$. It is easy to see by the doubling trick of Algorithm 1 that the number of episodes is bounded as $m \leq SAU$. We also have $\varepsilon_k = \frac{1}{t_k(1-\gamma)}$.

Proof Let $T \geq 1$ and $\gamma, \delta \in (0, 1)$ be arbitrary. For the optimal discounted value function we will drop the γ and simply write $V^\star$. By assumption $H \geq \|V^\star\|_{\text{sp}}$.

As is standard in the analysis of optimistic algorithms for online RL, a key step in our analysis is to establish an optimism property, which enables the rest of our regret decomposition.

Step 1: Optimism We state the fact that with high probability, our Q-estimate $\widehat{Q}_k$ and value estimate $\widehat{V}_k$ are indeed optimistic across all episodes. We defer the proof to Appendix H.1.

Lemma 18 (Optimism) *With probability $1 - SAT\delta'$, both of the following hold:*

1. $\widehat{Q}_k(s, a) \geq Q^\star(s, a) - \varepsilon_k$ *and* $\widehat{V}_k(s) \geq V^\star(s) - \varepsilon_k$ *for all* $(s, a, k) \in \mathcal{S} \times \mathcal{A} \times [m]$.
2. *For any* $k \in [m]$ *and* $t \in \{t_k, \ldots, t_{k+1} - 1\}$, $\widehat{V}_k(s_t) \leq r(s_t, a_t) + \gamma\widehat{P}^k_{s_t,a_t}\widehat{V}_k + b_k\left(s_t, a_t, \widehat{V}_k\right)$.

Equipped with optimism, we turn to decomposing the regret.

Step 2: Regret Decomposition Under the successful events of Lemma 18, observe that we have the following bound for any $t \in \{t_k, \ldots, t_{k+1} - 1\}$:

$$(1-\gamma)V^\star(s_t) - r(s_t, a_t) \leq \gamma\left(\widehat{P}^k_{s_t,a_t}\widehat{V}_k - \widehat{V}_k(s_t)\right) + \gamma b_k\left(s_t, a_t, \widehat{V}_k\right) + (1-\gamma)\varepsilon_k. \tag{1}$$

Now, it is not immediately obvious how we should bound the first term of the RHS, so we will relate it to something we do know how to bound. Observe that this term vaguely looks like $P_{s_t,a_t}\widehat{V}_k - \widehat{V}_k(s_{t+1})$. Defining X_t to be this expression, one can verify that $\{X_t\}_{0 \leq t \leq T}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma(s_1, a_1, \ldots, s_{t+1}, a_{t+1})$. Consequently, we can bound $\sum_{t=1}^T X_t$ with the following martingale concentration result.

Lemma 19 (Martingale Concentration; Adapted from Lemma 13 in Zhang et al. 2021) *Let $\{M_n\}_{n\geq 0}$ be a martingale with respect to some filtration $\{\mathcal{F}_n\}_{n\geq 0}$ such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ almost surely for some $c \geq 0$ and all $n \geq 1$. Let $\mathrm{Var}_n = \sum_{k=1}^{n} \mathbb{E}\left[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}\right]$. Then for any positive integer n and any $\delta \in (0,1)$, we have with probability at least $1 - 3n\delta$ that*

$$|M_n| < 2\sqrt{2}\sqrt{\mathrm{Var}_n \log\left(\frac{1}{\delta}\right)} + 4c\log\left(\frac{1}{\delta}\right).$$

We now have motivation for transforming $\widehat{P}^k_{s_t,a_t}\widehat{V}_k - \widehat{V}_k(s_t)$ into X_t. We accomplish this feat by introducing additional terms:

$$\widehat{P}^k_{s_t,a_t}\widehat{V}_k - \widehat{V}_k(s_t) = \left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + X_t + \left(\widehat{V}_k(s_{t+1}) - \widehat{V}_k(s_t)\right).$$

Substituting back into (1) and summing over t gives us the following decomposition:

$$\begin{aligned}
&\sum_{t=1}^{T}((1-\gamma)V^\star(s_t) - r(s_t,a_t)) \\
&= \sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}((1-\gamma)V^\star(s_t) - r(s_t,a_t)) \\
&\leq \sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\widehat{P}^k_{s_t,a_t}\widehat{V}_k - \widehat{V}_k(s_t) + b_k\left(s_t,a_t,\widehat{V}_k\right) + (1-\gamma)\varepsilon_k\right) \\
&= \underbrace{\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + b_k\left(s_t,a_t,\widehat{V}_k\right) + (1-\gamma)\varepsilon_k\right)}_{=:\mathscr{T}_{\mathrm{model}}} \\
&\qquad + \underbrace{\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(P_{s_t,a_t}\widehat{V}_k - \widehat{V}_k(s_{t+1})\right)}_{=:\mathscr{T}_{\mathrm{mart}}} \\
&\qquad + \underbrace{\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\widehat{V}_k(s_{t+1}) - \widehat{V}_k(s_t)\right)}_{=:\mathscr{T}_{\mathrm{ind}}}.
\end{aligned}$$

We call the first term $\mathscr{T}_{\mathrm{model}}$ because it is a function of the error of our model estimate. Note that we include the bonus term in $\mathscr{T}_{\mathrm{model}}$ mainly due to technical reasons, but intuitively our bound for the model error term will involve a Bernstein inequality and this will absorb the Bernstein-style bonus. The second term is $\mathscr{T}_{\mathrm{mart}}$ because we will bound it using martingale concentration. Lastly, we call the third term $\mathscr{T}_{\mathrm{ind}}$ because it arises due to our need to shift indices.

Step 3: Bounding $\mathscr{T}_{\mathrm{model}}$ and $\mathscr{T}_{\mathrm{mart}}$ by cumulative variance terms Our next step is to bound each term in the decomposition. Before doing so, we introduce some cumulative variance terms that arise in the analysis. We define

$$\mathrm{Var}^\star_\gamma := \sum_{t=1}^{T} \mathbb{V}_{s_t,a_t}\left(V^\star\right), \qquad \mathrm{Var}^{\mathrm{diff}} := \sum_{t=1}^{T} \mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t} - V^\star\right).$$

We start with $\mathscr{T}_{\mathrm{model}} = \sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + b_k\left(s_t,a_t,\widehat{V}_k\right) + (1-\gamma)\varepsilon_k\right)$. We would like to bound the term involving model error via a Bernstein-like concentration inequality. We cannot immediately do so, however, because $\widehat{P}^k_{s_t,a_t}$ and $\widehat{V}_k$ are not statistically independent. To overcome this hurdle, we perform the following decomposition:

$$\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k = \left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)V^\star + \left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\left(\widehat{V}_k - V^\star\right).$$

Since $V^\star$ is fixed, we can apply Bernstein to the first term. Bounding the second term is also doable because $\|\widehat{V}_k - V^\star\|_\infty$ decreases as k increases. The bonus term is designed so that it is subsumed by the model error term. We state the bound for $\mathscr{T}_{\mathrm{model}}$ in the following lemma, which we formally prove in Appendix H.2.

Lemma 20 *With probability at least $1 - 2S^2AT\delta'$, we have*

$$\mathscr{T}_{\mathrm{model}} \le O\left(\sqrt{SAU\mathrm{Var}^\star_\gamma} + \sqrt{\Gamma SAU^2\mathrm{Var}^{\mathrm{diff}}} + \Gamma HSAU^2\right).$$

We now move on to $\mathscr{T}_{\mathrm{mart}}$. As discussed above, we can apply Lemma 19, which gives us that with probability $1 - 3T\delta'$,

$$\mathscr{T}_{\mathrm{mart}} \le 2\sqrt{2}\sqrt{\sum_{t=1}^{T}\mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t}\right)U} + 4HU \le 4\sqrt{\mathrm{Var}^\star_\gamma U} + 4\sqrt{\mathrm{Var}^{\mathrm{diff}}U} + 4HU, \tag{2}$$

where the second inequality holds because

$$\sqrt{\sum_{t=1}^{T}\mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t}\right)} = \sqrt{\sum_{t=1}^{T}\mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t} - V^\star + V^\star\right)} \le \sqrt{\sum_{t=1}^{T}\left(2\mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t} - V^\star\right) + 2\mathbb{V}_{s_t,a_t}(V^\star)\right)}$$
$$= \sqrt{2\mathrm{Var}^\star_\gamma + 2\mathrm{Var}^{\mathrm{diff}}} \le \sqrt{2\mathrm{Var}^\star_\gamma} + \sqrt{2\mathrm{Var}^{\mathrm{diff}}}.$$

Step 4: Bounding cumulative variance terms We can bound $\mathrm{Var}^{\mathrm{diff}}$ with the following lemma, whose proof we defer to Appendix H.3.

Lemma 21 *Let $\mathcal{E}_{\mathrm{opt}}$ be the event that Lemma 18 holds. Under the event $\mathcal{E}_{\mathrm{opt}} \cap \mathcal{E}_{\mathrm{diff}}$, we have*

$$\mathrm{Var}^{\mathrm{diff}} \le O\left(\mathscr{T}_{\mathrm{model}}H + H^2SAU\right),$$

where $\mathcal{E}_{\mathrm{diff}}$ is an event that occurs with probability at least $1 - 3T\delta'$.

Subsequently, under the successful events of Lemma 20, (2), and Lemma 21, we have

$$\mathscr{T}_{\mathrm{model}} + \mathscr{T}_{\mathrm{mart}} \le O\left(\sqrt{\Gamma HSAU^2}\sqrt{\mathscr{T}_{\mathrm{model}} + \mathscr{T}_{\mathrm{mart}}} + \sqrt{\mathrm{Var}^\star_\gamma SAU} + \Gamma HSAU^2\right),$$

which by Lemma 15 implies that

$$\mathscr{T}_{\mathrm{model}} + \mathscr{T}_{\mathrm{mart}} \le O\left(\sqrt{\mathrm{Var}^\star_\gamma SAU} + \Gamma HSAU^2\right). \tag{3}$$

Step 5: Bounding $\mathscr{T}_{\text{ind}}$ It remains to bound $\mathscr{T}_{\text{ind}}$, which turns out to be the most straightforward argument. We compute

$$\mathscr{T}_{\text{ind}} = \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(\widehat{V}_k(s_{t+1}) - \widehat{V}_k(s_t)\right) = \sum_{k=1}^{m} \left(\widehat{V}_k(s_{t_{k+1}}) - \widehat{V}_k(s_{t_k})\right) \leq mH \leq HSAU.$$

Step 6: Recombining Finally, we recombine terms. Under the successful events of Lemma 18 and (3), we have

$$\text{Regret}_\gamma(T) \leq \gamma(\mathscr{T}_{\text{model}} + \mathscr{T}_{\text{mart}} + \mathscr{T}_{\text{ind}}) \leq O\left(\sqrt{\text{Var}_\gamma^\star SAU} + \Gamma HSAU^2\right).$$

Via a union bound (over the high probability events of Lemma 18, Lemma 20, (2), and Lemma 21), this occurs with probability at least $1 - 9S^2AT\delta' = 1 - \delta$. ■

Appendix F. Proof of Lemma 2

Proof Let $\delta \in (0, 1)$ be arbitrary, and set $\delta' = \frac{\delta}{6T}$. For the optimal value function we will drop γ and simply write $V^\star$. Our goal is to bound $\text{Var}_\gamma^\star = \sum_{t=1}^T \mathbb{V}(P_{s_t,a_t}, V^\star)$.

Now, observe that if we set $\widetilde{V}^\star := V^\star - \min_s V^\star(s)$, then we have $\widetilde{V}^\star \in \left[0, \|V^\star\|_{\text{sp}}\right]^S$ and $\sum_{t=1}^T \mathbb{V}_{s_t,a_t}(V^\star) = \sum_{t=1}^T \mathbb{V}_{s_t,a_t}\left(\widetilde{V}^\star\right)$. Subsequently, we perform the following decomposition.

$$\begin{aligned}
\sum_{t=1}^T \mathbb{V}_{s_t,a_t}\left(\widetilde{V}^\star\right) &= \sum_{t=1}^T \left(P_{s_t,a_t}\left(\widetilde{V}^\star\right)^2 - \left(P_{s_t,a_t}\widetilde{V}^\star\right)^2\right) \\
&= \underbrace{\sum_{t=1}^T \left(P_{s_t,a_t}\left(\widetilde{V}^\star\right)^2 - \left(\widetilde{V}^\star(s_{t+1})\right)^2\right)}_{=:\mathscr{T}_1^\star} + \underbrace{\sum_{t=1}^T \left(\left(\widetilde{V}^\star(s_t)\right)^2 - \left(P_{s_t,a_t}\widetilde{V}^\star\right)^2\right)}_{=:\mathscr{T}_2^\star} \\
&\qquad + \underbrace{\sum_{t=1}^T \left(\left(\widetilde{V}^\star(s_{t+1})\right)^2 - \left(\widetilde{V}^\star(s_t)\right)^2\right)}_{=:\mathscr{T}_3^\star}.
\end{aligned}$$

With probability at least $1 - 3T\delta'$, we have that

$$\begin{aligned}
\mathscr{T}_1^\star &\leq 2\sqrt{2}\sqrt{\sum_{t=1}^T \mathbb{V}_{s_t,a_t}\left(\left(\widetilde{V}^\star\right)^2\right)\log\left(\frac{1}{\delta'}\right)} + 4\|V^\star\|_{\text{sp}}^2 \log\left(\frac{1}{\delta'}\right) \\
&\leq 4\sqrt{2}\|V^\star\|_{\text{sp}}\sqrt{\text{Var}_\gamma^\star \log\left(\frac{1}{\delta'}\right)} + 4\|V^\star\|_{\text{sp}}^2 \log\left(\frac{1}{\delta'}\right),
\end{aligned}$$

where the first inequality holds by Lemma 19 and the second inequality holds by Lemma 14.

Next, we compute that with probability at least $1 - 3T\delta'$, we have

$$
\begin{aligned}
\mathscr{T}_2^\star &= \sum_{t=1}^{T} \left(\left(\widetilde{V}^\star(s_t) \right)^2 - \left(P_{s_t,a_t} \widetilde{V}^\star \right)^2 \right) \\
&\overset{(i)}{\leq} 2 \left\| V^\star \right\|_{\mathrm{sp}} \sum_{t=1}^{T} \max \left\{ \widetilde{V}^\star(s_t) - P_{s_t,a_t} \widetilde{V}^\star, 0 \right\} \\
&\leq 2 \left\| V^\star \right\|_{\mathrm{sp}} \sum_{t=1}^{T} \max \left\{ \widetilde{V}^\star(s_t) - P_{s_t,a_t} \widetilde{V}^\star + 1, 0 \right\} \\
&\overset{(ii)}{=} 2 \left\| V^\star \right\|_{\mathrm{sp}} \sum_{t=1}^{T} \left(\widetilde{V}^\star(s_t) - P_{s_t,a_t} \widetilde{V}^\star + 1 \right) \\
&= 2 \left\| V^\star \right\|_{\mathrm{sp}} \left(T + \sum_{t=1}^{T} \left(\widetilde{V}^\star(s_t) - P_{s_t,a_t} \widetilde{V}^\star \right) \right) \\
&\leq 2 \left\| V^\star \right\|_{\mathrm{sp}} \left(T + \widetilde{V}^\star(s_1) + \sum_{t=1}^{T} \left(\widetilde{V}^\star(s_{t+1}) - P_{s_t,a_t} \widetilde{V}^\star \right) \right) \\
&\overset{(iii)}{\leq} 2 \left\| V^\star \right\|_{\mathrm{sp}} \left(T + \widetilde{V}^\star(s_1) + 2\sqrt{2} \sqrt{\mathrm{Var}_\gamma^\star \log \left(\frac{1}{\delta'} \right)} + 4 \left\| V^\star \right\|_{\mathrm{sp}} \log \left(\frac{1}{\delta'} \right) \right) \\
&\leq 4\sqrt{2} \left\| V^\star \right\|_{\mathrm{sp}} \sqrt{\mathrm{Var}_\gamma^\star \log \left(\frac{1}{\delta'} \right)} + 2T \left\| V^\star \right\|_{\mathrm{sp}} + 10 \left\| V^\star \right\|_{\mathrm{sp}}^2 \log \left(\frac{1}{\delta'} \right).
\end{aligned}
$$

Inequality (i) holds because $a^2 - b^2 = (a+b)(a-b) \leq (a+b)\max\{a-b, 0\} \leq 2\left\| V^\star \right\|_{\mathrm{sp}} \max\{a-b, 0\}$ for $a, b \in \left[0, \left\| V^\star \right\|_{\mathrm{sp}}\right]$. Equality (ii) holds because $\widetilde{V}^\star(s_t) - P_{s_t,a_t}\widetilde{V}^\star + 1 \geq 0$. Indeed, we have

$$
\widetilde{V}^\star(s_t) - P_{s_t,a_t}\widetilde{V}^\star = V^\star(s_t) - P_{s_t,a_t}V^\star \geq Q^\star(s_t,a_t) - P_{s_t,a_t}V^\star = r(s_t,a_t) - (1-\gamma)P_{s_t,a_t}V^\star \geq -1.
$$

Finally, inequality (iii) holds with probability at least $1 - 3T\delta'$ by Lemma 19.

Observing that $\mathscr{T}_3^\star$ is a telescoping sum, we have

$$
\begin{aligned}
\mathscr{T}_3^\star &= \sum_{t=1}^{T} \left(\left(\widetilde{V}^\star(s_{t+1}) \right)^2 - \left(\widetilde{V}^\star(s_t) \right)^2 \right) \\
&\leq \left(\widetilde{V}^\star(s_{T+1}) \right)^2 \\
&\leq \left\| V^\star \right\|_{\mathrm{sp}}^2.
\end{aligned}
$$

Recombining terms, we have with probability $1 - 6T\delta'$ that

$$
\mathrm{Var}_\gamma^\star \leq \mathscr{T}_1^\star + \mathscr{T}_2^\star + \mathscr{T}_3^\star \leq O\left(\left\| V^\star \right\|_{\mathrm{sp}} \sqrt{\mathrm{Var}_\gamma^\star \log \left(\frac{1}{\delta'} \right)} + T \left\| V^\star \right\|_{\mathrm{sp}} + \left\| V^\star \right\|_{\mathrm{sp}}^2 \log \left(\frac{1}{\delta'} \right) \right),
$$

which by Lemma 15 implies

$$\mathrm{Var}_\gamma^\star \le O\left(\|V^\star\|_{\mathrm{sp}} T + \|V^\star\|_{\mathrm{sp}}^2 \log\left(\frac{1}{\delta'}\right)\right).$$

Substituting back $\delta' = \frac{\delta}{6T}$ gives us that

$$\mathrm{Var}_\gamma^\star \le O\left(\|V^\star\|_{\mathrm{sp}} T + \|V^\star\|_{\mathrm{sp}}^2 \log(T/\delta)\right)$$

with probability at least $1-\delta$. ■

Appendix G. Proof of Lemma 4

The proof relies on the following lemma.

Lemma 22 (Lemma 6 in Zurek and Chen 2025a) *Suppose the underlying MDP is weakly communicating. Let $\gamma \in (0,1)$. The optimal value function $V_\gamma^\star$ satisfies*

$$\left\|\rho^\star - (1-\gamma)V_\gamma^\star\right\|_\infty \le (1-\gamma)\left\|V_\gamma^\star\right\|_{\mathrm{sp}}.$$

Proof For any $T \ge 1$ and $\gamma \in (0,1)$, we have

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{t=1}^T (\rho^\star - r(s_t,a_t)) \\
&= \sum_{t=1}^T \left(\rho^\star - (1-\gamma)V_\gamma^\star(s_t)\right) + \sum_{t=1}^T \left((1-\gamma)V_\gamma^\star(s_t) - r(s_t,a_t)\right) \\
&\le \sum_{t=1}^T (1-\gamma)\left\|V_\gamma^\star\right\|_{\mathrm{sp}} + \sum_{t=1}^T \left((1-\gamma)V_\gamma^\star(s_t) - r(s_t,a_t)\right) \\
&= (1-\gamma)\left\|V_\gamma^\star\right\|_{\mathrm{sp}} T + \mathrm{Regret}_\gamma(T),
\end{aligned}$$

where the inequality is due to Lemma 22. ■

Appendix H. Missing Proofs from Appendix E

H.1. Proof of Lemma 18

Proof We denote by $\widehat{\mathcal{T}}_k$ the empirical Bellman operator used in Algorithm 1 during the kth episode. In particular, we have

$$\left(\widehat{\mathcal{T}}_k Q\right)(s,a) = r(s,a) + \gamma \widehat{P}_{s,a}^k \mathrm{Clip}_H(MQ) + \gamma b_k\left(s,a,\mathrm{Clip}_H(MQ)\right).$$

We also write iters_k to be the number of value iterations used Algorithm 1 during the kth episode, so that

$$\mathrm{iters}_k = \left\lceil \frac{1}{1-\gamma} \log \frac{1+32HU}{\varepsilon_k(1-\gamma)} \right\rceil.$$

Observe that

$$\gamma^{\text{iters}_k} \leq \exp(-(1-\gamma)\text{iters}_k) \leq \frac{\varepsilon_k(1-\gamma)}{1+32HU},$$

where the first inequality is due to $x \leq e^{-(1-x)}$. The following Lemma, which states some crucial properties of $\widehat{\mathcal{T}}_k$, is proved in Appendix H.4.

Lemma 23 *With probability $1 - SAT\delta'$, the following hold for all $k \in [m]$.*

1. *$\widehat{\mathcal{T}}_k$ is monotonic: $\widehat{\mathcal{T}}_k Q \geq \widehat{\mathcal{T}}_k Q'$ for any $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that $Q \geq Q'$.*
2. *$\widehat{\mathcal{T}}_k$ is a γ-contraction: $\left\|\widehat{\mathcal{T}}_k Q - \widehat{\mathcal{T}}_k Q'\right\|_\infty \leq \gamma \left\|Q - Q'\right\|_\infty$ for any $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.*
3. *$Q^\star \leq \widehat{\mathcal{T}}_k Q^\star$.*

Now, assume that the events of Lemma 23 hold, and fix an arbitrary k. We first prove Part 1 of Lemma 18. By the Banach fixed point theorem (Pugh, 2015), the fact that $\widehat{\mathcal{T}}_k$ is a γ-contraction implies that $\widehat{\mathcal{T}}_k$ has a unique fixed point, which we will denote $\widehat{Q}_k^\star$. By monotonicity of $\widehat{\mathcal{T}}_k$, the sequence $Q^\star, \widehat{\mathcal{T}}_k Q^\star, \widehat{\mathcal{T}}_k\widehat{\mathcal{T}}_k Q^\star, \ldots$ is nondecreasing and converges to $\widehat{Q}_k^\star$, which implies

$$Q^\star \leq \widehat{Q}_k^\star. \tag{4}$$

Furthermore, we have

$$\left\|\widehat{Q}_k - \widehat{Q}_k^\star\right\|_\infty = \left\|\left(\widehat{\mathcal{T}}_k\right)^{(\text{iters}_k)}\mathbf{0} - \left(\widehat{\mathcal{T}}_k\right)^{(\text{iters}_k)}\widehat{Q}_k^\star\right\|_\infty \leq \gamma^{\text{iters}_k}\left\|\widehat{Q}_k^\star\right\|_\infty \leq \varepsilon_k, \tag{5}$$

where the final inequality holds by the above bound on γ_k^{iters} and

$$\begin{aligned}
&\left(\widehat{\mathcal{T}}_k Q\right)(s,a) \leq 1 + \gamma\|Q\|_\infty + \gamma 32HU \\
&\implies \left\|\widehat{\mathcal{T}}_k\mathbf{0}\right\|_\infty \leq 1 + \gamma 32HU \\
&\implies \left\|\widehat{\mathcal{T}}_k\widehat{\mathcal{T}}_k\mathbf{0}\right\|_\infty \leq 1 + \gamma 32HU + \gamma(1+\gamma 32HU) \\
&\quad\vdots \\
&\implies \left\|\widehat{Q}_k^\star\right\|_\infty \leq \frac{1+\gamma 32HU}{1-\gamma}.
\end{aligned}$$

Combining (4) and (5) gives us that (elementwise)

$$\widehat{Q}_k \geq Q^\star - \mathbf{1}\varepsilon_k.$$

Subsequently, for any $s \in \mathcal{S}$,

$$\begin{aligned}
\widehat{V}_k(s) &= \min\left\{\left(M\widehat{Q}_k\right)(s), \min_{s'}\left(M\widehat{Q}_k\right)(s') + H\right\} \\
&\geq \min\left\{\left(M\left(Q^\star - \mathbf{1}\varepsilon_k\right)\right)(s), \min_{s'}\left(M\left(Q^\star - \mathbf{1}\varepsilon_k\right)\right)(s') + H\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \min\left\{(MQ^\star)(s), \min_{s'}(MQ^\star)(s') + H\right\} - \varepsilon_k \\
&= \min\left\{V^\star(s), \min_{s'} V^\star(s') + H\right\} - \varepsilon_k \\
&= V^\star(s) - \varepsilon_k,
\end{aligned}
$$

so the desired result holds.

It remains to prove Part 2 of Lemma 18. We remark that $\widehat{Q}_k \le \widehat{\mathcal{T}}_k \widehat{Q}_k$ because monotonicity implies

$$
\mathbf{0} \le \widehat{\mathcal{T}}_k \mathbf{0} \le \widehat{\mathcal{T}}_k \widehat{\mathcal{T}}_k \mathbf{0} \le \cdots \le \underbrace{\left(\widehat{\mathcal{T}}_k\right)^{(\text{iters}_k)} \mathbf{0}}_{=\widehat{Q}_k} \le \underbrace{\left(\widehat{\mathcal{T}}_k\right)^{(\text{iters}_k+1)} \mathbf{0}}_{=\widehat{\mathcal{T}}_k \widehat{Q}_k}.
$$

It follows that for any $t \in \{t_k, \ldots, t_{k+1}-1\}$,

$$
\begin{aligned}
\widehat{V}_k(s_t) &\le \left(M\widehat{Q}_k\right)(s_t) \\
&= \widehat{Q}_k(s_t, a_t) \\
&\le \widehat{\mathcal{T}}_k \widehat{Q}_k(s_t, a_t) \\
&= r(s_t, a_t) + \gamma \widehat{P}^k_{s_t,a_t} \widehat{V}_k + \gamma b_k\left(s_t, a_t, \widehat{V}_k\right),
\end{aligned}
$$

so the desired result holds. ■

H.2. Proof of Lemma 20

Proof As explained in the main section of the proof, we decompose

$$
\begin{aligned}
\mathscr{T}_{\text{model}} &= \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + b_k\left(s_t, a_t, \widehat{V}_k\right) + (1-\gamma)\varepsilon_k\right) \\
&= \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \Bigg(\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right) V^\star + \left(P^k_{s_t,a_t} - P_{s_t,a_t}\right)\left(\widehat{V}_k - V^\star\right) \\
&\qquad + b_k\left(s_t, a_t, \widehat{V}_k\right) + (1-\gamma)\varepsilon_k \Bigg).
\end{aligned}
$$

We will bound each of the first three terms inside the sum separately under high probability events. Recall that for ease of notation we denote $n_k(s,a) \coloneqq \max\{N_k(s,a), 1\}$. Let $\mathcal{E}_1$ be the event that

$$
\left|\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right| \le \sqrt{\frac{2P_{s,a,s'}U}{n_k(s,a)}} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right)U}{3n_k(s,a)} \qquad \forall (s,a,s',k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [m],
$$

and let $\mathcal{E}_2$ be the event that

$$
\left|\left(\widehat{P}^k_{s,a} - P_{s,a}\right)V^\star\right| \le \sqrt{\frac{2\mathbb{V}_{s,a}\left(V^\star\right)U}{n_k(s,a)}} + \frac{HU}{3n_k(s,a)} \qquad \forall (s,a,k) \in \mathcal{S} \times \mathcal{A} \times [m]. \tag{6}
$$

We will later confirm that $\mathcal{E}_1$ and $\mathcal{E}_2$ are indeed high probability events. Under event $\mathcal{E}_1$, for any $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times [m]$ we have

$$
\begin{aligned}
&\left(\widehat{P}^k_{s,a} - P_{s,a}\right)\left(\widehat{V}_k - V^\star\right) \\
&= \sum_{s' \in \mathcal{S}} \left(\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right)\left(\widehat{V}_k(s') - V^\star(s')\right) \\
&\overset{(i)}{=} \sum_{s' \in \mathcal{S}} \left(\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right)\left(\widehat{V}_k(s') - V^\star(s') - P_{s,a}(\widehat{V}_k - V^\star)\right) \\
&\overset{(ii)}{\le} \sum_{s' \in \mathcal{S}} \left| \sqrt{\frac{2P_{s,a,s'}U}{n_k(s,a)}} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right)U}{3n_k(s,a)} \right| \cdot \left| \widehat{V}_k(s') - V^\star(s') - P_{s,a}\left(\widehat{V}_k - V^\star\right) \right| \\
&\overset{(iii)}{\le} \sum_{s' \in \mathcal{S}} \sqrt{\frac{2P_{s,a,s'}U}{n_k(s,a)}} \cdot \left| \widehat{V}_k(s') - V^\star(s') - P_{s,a}\left(\widehat{V}_k - V^\star\right) \right| + \frac{2\Gamma HU}{3n_k(s,a)} \\
&\overset{(iv)}{\le} \sqrt{\frac{2\Gamma U \sum_{s' \in \mathcal{S}} P_{s,a,s'}\left(\widehat{V}_k(s') - V^\star(s') - P_{s,a}\left(\widehat{V}_k - V^\star\right)\right)^2}{n_k(s,a)}} + \frac{2\Gamma HU}{3n_k(s,a)} \\
&= \sqrt{\frac{2\Gamma U \mathbb{V}_{s,a}\left(\widehat{V}_k - V^\star\right)}{n_k(s,a)}} + \frac{2\Gamma HU}{3n_k(s,a)}.
\end{aligned} \tag{7}
$$

Equality (i) holds because $\sum_{s'}\left(\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right)c = 0$ for any $c \in \mathbb{R}$. Inequality (ii) holds under event $\mathcal{E}_1$. We obtain inequality (iii) by bounding $\left|\widehat{V}_k(s') - V^\star(s') - P_{s,a}\left(\widehat{V}_k - V^\star\right)\right| \le 2H$ and summing over s'. Inequality (iv) follows from Cauchy-Schwarz. Moreover, under $\mathcal{E}_1$, for any $(s,a,k) \in \mathcal{S} \times \mathcal{A} \times [m]$ we have

$$
\begin{aligned}
b_k\left(s,a,\widehat{V}_k\right) &= \max\left\{4\sqrt{\frac{\mathbb{V}\left(\widehat{P}^k_{s,a},\widehat{V}_k\right)U}{n_k(s,a)}}, 32\frac{HU}{n_k(s,a)}\right\} \\
&\le 4\sqrt{\frac{\mathbb{V}\left(\widehat{P}^k_{s,a},\widehat{V}_k\right)U}{n_k(s,a)}} + 32\frac{HU}{n_k(s,a)} \\
&\le 4\sqrt{3/2}\sqrt{\frac{\mathbb{V}_{s,a}\left(\widehat{V}_k\right)U}{n_k(s,a)}} + 4\sqrt{4/3}\frac{\sqrt{\Gamma H^2 U}}{n_k(s,a)} + 32\frac{HU}{n_k(s,a)} \\
&\le 5\sqrt{\frac{\mathbb{V}_{s,a}\left(\widehat{V}_k\right)U}{n_k(s,a)}} + 37\frac{\Gamma HU}{n_k(s,a)},
\end{aligned} \tag{8}
$$

with the second inequality holding due to

$$
\mathbb{V}\left(\widehat{P}^k_{s,a},\widehat{V}_k\right) = \sum_{s' \in \mathcal{S}} \widehat{P}^k_{s,a,s'}\left(\widehat{V}_k(s') - \widehat{P}^k_{s,a}\widehat{V}_k\right)^2
$$

$$\overset{(i)}{\leq} \sum_{s' \in \mathcal{S}} \widehat{P}^k_{s,a,s'} \left(\widehat{V}_k(s') - P_{s,a} \widehat{V}_k \right)^2$$
$$\overset{(ii)}{\leq} \sum_{s' \in \mathcal{S}} \left(\frac{3}{2} P_{s,a,s'} + \frac{4 \mathbb{I}\left(P_{s,a,s'} > 0\right) U}{3 n_k(s,a)} \right) \left(\widehat{V}_k(s') - P_{s,a} \widehat{V}_k \right)^2$$
$$\leq \frac{3}{2} \mathbb{V}_{s,a}\left(\widehat{V}_k\right) + \frac{4 \Gamma H^2 U}{3 n_k(s,a)}.$$

Here, inequality (i) is because $\mathbb{E}[X]$ minimizes $f(\lambda) = \mathbb{E}[(X-\lambda)^2]$, and inequality (ii) holds under $\mathcal{E}_1$. Indeed, under $\mathcal{E}_1$, we have

$$\left|\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right| \leq \sqrt{\frac{2 P_{s,a,s'} \mathbb{I}\left(P_{s,a,s'} > 0\right) U}{n_k(s,a)}} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right) U}{3 n_k(s,a)}$$
$$\implies \widehat{P}^k_{s,a,s'} - P_{s,a,s'} \leq \frac{1}{2} P_{s,a,s'} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right) U}{n_k(s,a)} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right) U}{3 n_k(s,a)}$$
$$\implies \widehat{P}^k_{s,a,s'} \leq \frac{3}{2} P_{s,a,s'} + \frac{4 \mathbb{I}\left(P_{s,a,s'} > 0\right) U}{3 n_k(s,a)},$$

where the first implication holds because $\sqrt{ab} \leq \frac{a}{2} + \frac{b}{2}$ for $a, b \geq 0$.

Now, combining (6), (7), and (8) gives us that under $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\mathscr{T}_{\text{model}} \leq O\left(\sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(\sqrt{\frac{U \mathbb{V}_{s_t,a_t}\left(V^\star\right)}{n_k(s_t,a_t)}} + \sqrt{\frac{\Gamma U \mathbb{V}_{s_t,a_t}\left(\widehat{V}_k - V^\star\right)}{n_k(s_t,a_t)}} + \frac{\Gamma H U}{n_k(s_t,a_t)} + (1-\gamma)\varepsilon_k \right) \right).$$

The following lemma, which we prove in Appendix H.5, allows us to easily bound the above.z

Lemma 24 *We have the following bounds.*

1. *Define* $n_{\text{tot}}(s,a) := \sum_{t \in [T]} \mathbb{I}((s_t,a_t) = (s,a))$. *Then* $\sum_{t:(s,a)=(s_t,a_t)} \frac{1}{\sqrt{n_{k_t}(s,a)}} \leq 2\sqrt{2 n_{\text{tot}}(s,a)}$.
2. *For nonnegative numbers* $w_1, \ldots, w_T$, *we have* $\sum_{t=1}^{T} \sqrt{\frac{w_t}{n_{k_t}(s_t,a_t)}} \leq \sqrt{SAU \sum_{t=1}^{T} w_t}$.
3. $\sum_{t=1}^{T} \frac{1}{n_{k_t}(s_t,a_t)} \leq SAU$.
4. $\sum_{t=1}^{T} (1-\gamma)\varepsilon_{k_t} \leq SAU$.

First, we have by Part 1 of Lemma 24 that

$$\sum_{t=1}^{T} \sqrt{\frac{U \mathbb{V}_{s_t,a_t}(V^\star)}{n_{k_t}(s_t,a_t)}} = \sum_{s,a} \sum_{t:(s,a)=(s_t,a_t)} \sqrt{\frac{U \mathbb{V}_{s,a}(V^\star)}{n_{k_t}(s,a)}} \leq \sum_{s,a} 2\sqrt{2} \sqrt{U n_{\text{tot}}(s,a) \mathbb{V}_{s,a}(V^\star)}.$$

By Cauchy-Schwarz, it follows that

$$\sum_{t=1}^{T} \sqrt{\frac{U \mathbb{V}_{s_t,a_t}(V^\star)}{n_{k_t}(s_t,a_t)}} \leq 2\sqrt{2} \sqrt{SAU \sum_{s,a} n_{\text{tot}}(s,a) \mathbb{V}_{s,a}(V^\star)} = 2\sqrt{2} \sqrt{SAU \text{Var}^\star_\gamma}.$$

Next, setting $w_t = \mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t} - V^\star\right)$, Part 2 of Lemma 24 gives us that

$$\sum_{t=1}^{T}\sqrt{\frac{\Gamma U \mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t} - V^\star\right)}{n_{k_t}(s_t,a_t)}} \leq \sqrt{\Gamma SAU^2 \mathrm{Var}^{\mathrm{diff}}}.$$

Lastly, Parts 3 and 4 of Lemma 24 give us that

$$\sum_{t=1}^{T}\left(\frac{\Gamma HU}{n_{k_t}(s_t,a_t)} + (1-\gamma)\varepsilon_{k_t}\right) \leq 2\Gamma HSAU^2.$$

Combining these three bounds, we have that under $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\mathscr{T}_{\mathrm{model}} \leq O\left(\sqrt{SAU\mathrm{Var}^\star_\gamma} + \sqrt{\Gamma SAU^2\mathrm{Var}^{\mathrm{diff}}} + \Gamma HSAU^2\right).$$

It remains to show that $\mathcal{E}_1$ and $\mathcal{E}_2$ hold with the claimed high probability. Starting with $\mathcal{E}_1$, fix $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, and suppose $n_k(s,a) = n$ for some $n \geq 1$. Observe that $P^k_{s,a,s'} = \frac{1}{n}\sum_{i=1}^n Z_i$, where $Z_1, \ldots, Z_n$ are i.i.d. Bernoulli random variables with mean $P_{s,a,s'}$. Denoting Z to also be an independent Bernoulli random variable with mean $P_{s,a,s'}$, we have with probability at least $1-\delta'$,

$$\left|\widehat{P}^k_{s,a,s'} - P_{s,a,s'}\right| \leq \sqrt{\frac{2\mathrm{Var}(Z)\log(2/\delta')}{n}} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right)\log(2/\delta')}{3n} \leq \sqrt{\frac{2P_{s,a,s'}U}{n}} + \frac{\mathbb{I}\left(P_{s,a,s'} > 0\right)U}{3n},$$

where the first inequality is due to Lemma 11 and the observation that $P_{s,a,s'} = 0 \implies \widehat{P}^k_{s,a,s'} = 0$, and the second inequality holds because $\mathrm{Var}(Z) \leq \mathbb{E}[Z^2] = \mathbb{E}[Z] = P_{s,a,s'}$ and $\log(2/\delta') \leq U$. It follows from a union bound over possible s, s', a, n gives us that $\mathcal{E}_1$ holds with probability at least $1 - S^2AT\delta'$.

Next, for $\mathcal{E}_2$, fix $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and suppose $n_k(s,a) = n$ for some $n \geq 1$. Observe that $\widehat{P}^k_{s,a}V^\star = \frac{1}{n}\sum_{i=1}^n Z_i$, where $Z_1, \ldots, Z_n$ are i.i.d. multinoulli random variables that take the value $V^\star(s')$ with probability $P_{s,a,s'}$ for each s'. Letting Z be i.i.d. from the same distribution as the Z_i's, we have that $\mathbb{E}[Z] = P_{s,a}V^\star$ and

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = P_{s,a}(V^\star)^2 - (P_{s,a}V^\star)^2 = \mathbb{V}_{s,a}(V^\star).$$

Subsequently, Lemma 11 gives us that with probability at least $1-\delta'$,

$$\left|P^t_{s,a}V^\star - P_{s,a}V^\star\right| \leq \sqrt{\frac{2\mathbb{V}_{s,a}(V^\star)U}{n}} + \frac{HU}{3n},$$

so it follows from a union bound over possible s, a, n that $\mathcal{E}_2$ holds with probability at least $1 - SAT\delta'$. We conclude that $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs with probability at least $1 - 2S^2AT\delta'$, as desired. ■

H.3. Proof of Lemma 21

Proof Our goal is to bound $\mathrm{Var}^{\mathrm{diff}} = \sum_{t=1}^{T} \mathbb{V}_{s_t,a_t}\left(\widehat{V}_{k_t} - V^\star\right)$. We will proceed in a manner similar to the proof of Lemma 2, where we bounded $\mathrm{Var}^\star_\gamma$. For ease of notation, write $D_k := \widehat{V}_k - V^\star$. Then, set $\widetilde{D}_k := D_k - \min_s D_k(s)$ so that $\widetilde{D}_k \in [0,H]^S$ and $\mathrm{Var}^{\mathrm{diff}} = \sum_{t=1}^{T} \mathbb{V}_{s_t,a_t}\left(\widetilde{D}_{k_t}\right)$. We subsequently have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{V}_{s_t,a_t}\left(\widetilde{D}_{k_t}\right) &= \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(P_{s_t,a_t}\left(\widetilde{D}_k\right)^2 - \left(P_{s_t,a_t}\widetilde{D}_k\right)^2 \right) \\
&= \underbrace{\sum_{t=1}^{T} \left(P_{s_t,a_t}\left(\widetilde{D}_{k_t}\right)^2 - \left(\widetilde{D}_{k_t}(s_{t+1})\right)^2 \right)}_{=:\mathscr{T}_1^{\mathrm{diff}}} \\
&\quad + \underbrace{\sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(\left(\widetilde{D}_k(s_t)\right)^2 - \left(P_{s_t,a_t}\widetilde{D}_k\right)^2 \right)}_{=:\mathscr{T}_2^{\mathrm{diff}}} \\
&\quad + \underbrace{\sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(\left(\widetilde{D}_k(s_{t+1})\right)^2 - \left(\widetilde{D}_k(s_t)\right)^2 \right)}_{=:\mathscr{T}_3^{\mathrm{diff}}}.
\end{aligned}$$

With probability at least $1 - 3T\delta'$, we have

$$\mathscr{T}_1^{\mathrm{diff}} \le 2\sqrt{2}\sqrt{\sum_{t=1}^{T} \mathbb{V}_{s_t,a_t}\left(\left(\widetilde{D}_{k_t}\right)^2\right) U} + 4H^2 U \le 4\sqrt{2}H\sqrt{\mathrm{Var}^{\mathrm{diff}} U} + 4H^2 U,$$

where the first inequality is by Lemma 19 and the second inequality is by Lemma 14. We denote by $\mathcal{E}_{\mathrm{diff}}$ the event that the above inequality holds.

Next, we compute that

$$\begin{aligned}
\mathscr{T}_2^{\mathrm{diff}} &= \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \left(\left(\widetilde{D}_k(s_t)\right)^2 - \left(P_{s_t,a_t}\widetilde{D}_k\right)^2 \right) \\
&\le 2H \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \max\left\{ \widetilde{D}_k(s_t) - P_{s_t,a_t}\widetilde{D}_k, 0 \right\} \\
&= 2H \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \max\left\{ D_k(s_t) - P_{s_t,a_t}D_k, 0 \right\} \\
&= 2H \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \max\left\{ \widehat{V}_k(s_t) - P_{s_t,a_t}\widehat{V}_k - \left(V^\star(s_t) - P_{s_t,a_t}V^\star\right), 0 \right\},
\end{aligned}$$

with the inequality holding by $a^2 - b^2 \leq 2H \max\{a-b, 0\}$ for $a, b \in [0, H]$. Furthermore, we have

$$\begin{aligned} \widehat{V}_k(s_t) - P_{s_t,a_t}\widehat{V}_k &\leq r(s_t,a_t) + \gamma \widehat{P}^k_{s_t,a_t}\widehat{V}_k + \gamma b_k\left(s_t,a_t,\widehat{V}_k\right) - P_{s_t,a_t}\widehat{V}_k \\ &= r(s_t,a_t) - (1-\gamma)P_{s_t,a_t}\widehat{V}_k + \gamma\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + \gamma b_k\left(s_t,a_t,\widehat{V}_k\right) \quad (9) \end{aligned}$$

as well as

$$V^\star(s_t) - P_{s_t,a_t}V^\star \geq r(s_t,a_t) - (1-\gamma)P_{s_t,a_t}V^\star. \quad (10)$$

Continuing from above by plugging in (9) and (10), we have

$$\begin{aligned} \mathscr{T}_2^{\text{diff}} &\leq 2H\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1} \max\left\{\widehat{V}_k(s_t) - P_{s_t,a_t}\widehat{V}_k - \left(V^\star(s_t) - P_{s_t,a_t}V^\star\right), 0\right\} \\ &\leq 2H\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1} \max\left\{(1-\gamma)P_{s_t,a_t}\left(V^\star - \widehat{V}_k\right) + \gamma\left(\widehat{P}^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + \gamma b_k\left(s_t,a_t,\widehat{V}_k\right), 0\right\} \\ &\overset{(i)}{\leq} 2H\sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\left(P^k_{s_t,a_t} - P_{s_t,a_t}\right)\widehat{V}_k + b_k\left(s_t,a_t,\widehat{V}_k\right) + (1-\gamma)\varepsilon_k\right) \\ &= 2H\mathscr{T}_{\text{model}}. \end{aligned}$$

Note that inequality (i) holds under the high probability events of Lemma 18, in which case $(1-\gamma)P_{s_t,a_t}\left(V^\star - \widehat{V}_k\right) \leq (1-\gamma)\max_s\left(V^\star(s) - \widehat{V}_k(s)\right) \leq (1-\gamma)\varepsilon_k$.

Furthermore, we compute

$$\begin{aligned} \mathscr{T}_3^{\text{diff}} &= \sum_{k=1}^{m}\sum_{t=t_k}^{t_{k+1}-1}\left(\left(\widetilde{D}_k(s_{t+1})\right)^2 - \left(\widetilde{D}_k(s_t)\right)^2\right) \\ &= \sum_{k=1}^{m}\left(\left(\widetilde{D}_k(s_{t_{k+1}})\right)^2 - \left(\widetilde{D}_k(s_{t_k})\right)^2\right) \\ &\leq mH^2 \\ &\leq H^2SAU. \end{aligned}$$

Recombining terms, we have under $\mathcal{E}_{\text{diff}}$ and the high probability events of Lemma 18 that

$$\text{Var}^{\text{diff}} \leq \mathscr{T}_1^{\text{diff}} + \mathscr{T}_2^{\text{diff}} + \mathscr{T}_3^{\text{diff}} \leq O\left(H\sqrt{U}\sqrt{\text{Var}^{\text{diff}}} + H\mathscr{T}_{\text{model}} + H^2SAU\right),$$

which by Lemma 15 implies

$$\text{Var}^{\text{diff}} \leq O\left(H\mathscr{T}_{\text{model}} + H^2SAU\right).$$

■

H.4. Proof of Lemma 23

Proof Let $\mathcal{E}$ be the event that

$$\left|\left(\widehat{P}_{s,a}^k - P_{s,a}\right) V^\star\right| \leq 2\sqrt{\frac{\mathbb{V}\left(\widehat{P}_{s,a}^k, V^\star\right) U}{n_k(s,a)}} + \frac{14HU}{3n_k(s,a)} \qquad \forall (s,a,k) \in \mathcal{S} \times \mathcal{A} \times [m].$$

We restate an alternate, stronger version of Lemma 23, which we will prove instead. Afterwards, we will complete the proof of Lemma 23 by showing that $\mathcal{E}$ holds with the claimed high probability.

Lemma 25 *The following hold for all $k \in [m]$.*

1. *$\widehat{\mathcal{T}}_k$ satisfies the constant shift property: for any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $c \in \mathbb{R}$, we have $\widehat{\mathcal{T}}_k(Q + c\mathbf{1}) = \widehat{\mathcal{T}}_k Q + \gamma c\mathbf{1}$.*
2. *$\widehat{\mathcal{T}}_k$ is monotonic: for any $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that $MQ \geq MQ'$, we have $\widehat{\mathcal{T}}_k Q \geq \widehat{\mathcal{T}}_k Q'$.*
3. *$\widehat{\mathcal{T}}_k$ is a γ-contraction: for any $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, we have $\left\|\widehat{\mathcal{T}}_k Q - \widehat{\mathcal{T}}_k Q'\right\|_\infty \leq \gamma\|Q - Q'\|_\infty$.*
4. *Under the event $\mathcal{E}$, we also have $Q^\star \leq \widehat{\mathcal{T}}_k Q^\star$.*

We remark that while we call Part 2 the monotonicity property, it is actually stronger. Indeed, for $\widehat{\mathcal{T}}_k Q \geq \widehat{\mathcal{T}}_k Q'$ to hold, we only need $MQ \geq MQ'$, which is a weaker requirement than $Q \geq Q'$. This version will be useful in proving Part 3. Now, let $k \in [m]$ be arbitrary.

To show Part 1 of Lemma 25 (constant shift), let $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $c \in \mathbb{R}$ be arbitrary. Since $\mathrm{Clip}_H(M(Q + c\mathbf{1})) = \mathrm{Clip}_H(MQ + c\mathbf{1}) = \mathrm{Clip}_H(MQ) + c\mathbf{1}$, a straightforward computation gives us

$$\begin{aligned}
\left(\widehat{\mathcal{T}}_k(Q + c\mathbf{1})\right)(s,a) &= r(s,a) + \gamma\widehat{P}_{s,a}^k\left(\mathrm{Clip}_H(M(Q + c\mathbf{1}))\right) + \gamma b_k\left(s,a,\mathrm{Clip}_H(M(Q + c\mathbf{1}))\right) \\
&= r(s,a) + \gamma\widehat{P}_{s,a}^k\left(\mathrm{Clip}_H(MQ)\right) + \gamma\widehat{P}_{s,a}^k(c\mathbf{1}) + \gamma b_k\left(s,a,\mathrm{Clip}_H(MQ)\right) \\
&= \widehat{\mathcal{T}}_k Q + \gamma c\mathbf{1}.
\end{aligned}$$

To show Part 2 of Lemma 25 (monotonicity), let $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be such that $MQ \geq MQ'$. Setting $\alpha := \min_s(\mathrm{Clip}_H(MQ'))(s)$ and $\beta := \min_s((\mathrm{Clip}_H(MQ))(s) - (\mathrm{Clip}_H(MQ'))(s))$. Observe that $\beta > 0, \mathrm{Clip}_H(MQ') - \alpha\mathbf{1} \in [0,H]^{\mathcal{S}}, \mathrm{Clip}_H(MQ) - \alpha\mathbf{1} - \beta\mathbf{1} \in [0,2H]^{\mathcal{S}}$ and $\mathrm{Clip}_H(MQ) - \alpha\mathbf{1} - \beta\mathbf{1} \geq \mathrm{Clip}_H(MQ') - \alpha\mathbf{1}$. Using $f(p,v,n,u)$ as defined in Lemma 13, we have

$$\begin{aligned}
&\left(\widehat{\mathcal{T}}_k Q\right)(s,a) \\
&= r(s,a) + \gamma\widehat{P}_{s,a}^k\left(\mathrm{Clip}_H(MQ)\right) + \gamma b_k\left(s,a,\mathrm{Clip}_H(MQ)\right) \\
&= r(s,a) + \gamma\widehat{P}_{s,a}^k\left(\mathrm{Clip}_H(MQ) - \alpha\mathbf{1} - \beta\mathbf{1}\right) + \gamma b_k\left(s,a,\mathrm{Clip}_H(MQ) - \alpha\mathbf{1} - \beta\mathbf{1}\right) + \gamma\alpha + \gamma\beta \\
&= r(s,a) + \gamma f\left(\widehat{P}_{s,a}^k, \mathrm{Clip}_H(MQ) - \alpha\mathbf{1} - \beta\mathbf{1}, n_k(s,a), U\right) + \gamma\alpha + \gamma\beta \\
&\geq r(s,a) + \gamma f\left(\widehat{P}_{s,a}^k, \mathrm{Clip}_H(MQ') - \alpha\mathbf{1}, n_k(s,a), U\right) + \gamma\alpha + \gamma\beta \\
&= r(s,a) + \gamma\widehat{P}_{s,a}^k(\mathrm{Clip}_H(MQ')) + \gamma b_k\left(s,a,\mathrm{Clip}_H(MQ')\right) + \gamma\beta
\end{aligned}$$

$$= \left(\widehat{\mathcal{T}}_k Q'\right)(s,a) + \gamma\beta$$
$$\geq \left(\widehat{\mathcal{T}}_k Q'\right)(s,a),$$

where the first inequality is due to Lemma 13.

To show Part 3 of Lemma 25 (γ-contraction), let $Q, Q' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be arbitrary. Since $MQ \leq MQ' + \|MQ - MQ'\|_\infty \mathbf{1} = M\left(Q' + \|MQ - MQ'\|_\infty \mathbf{1}\right)$, we have by the monotonicity and constant shift properties of $\widehat{\mathcal{T}}_k$ that

$$\widehat{\mathcal{T}}_k Q \leq \widehat{\mathcal{T}}_k \left(Q' + \|MQ - MQ'\|_\infty \mathbf{1}\right) \leq \widehat{\mathcal{T}}_k Q' + \gamma \|MQ - MQ'\|_\infty \leq \widehat{\mathcal{T}}_k Q' + \gamma\|Q - Q'\|_\infty.$$

Rearranging gives us

$$\widehat{\mathcal{T}}_k Q - \widehat{\mathcal{T}}_k Q' \leq \gamma\|Q - Q'\|_\infty.$$

Reversing the roles of Q and Q' in the above, we also have

$$\widehat{\mathcal{T}}_k Q' - \widehat{\mathcal{T}}_k Q \leq \gamma\|Q - Q'\|_\infty,$$

and combining this with the above, we conclude that

$$\left\|\widehat{\mathcal{T}}_k Q - \widehat{\mathcal{T}}_k Q'\right\|_\infty \leq \gamma\|Q - Q'\|_\infty.$$

Finally, we show Part 4 of Lemma 25. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}
Q^\star(s,a) &= r(s,a) + \gamma P_{s,a} V^\star \\
&= r(s,a) + \gamma \widehat{P}^k_{s,a} V^\star + \gamma\left(P_{s,a} - \widehat{P}^k_{s,a}\right) V^\star \\
&\leq r(s,a) + \gamma \widehat{P}^k_{s,a} \mathrm{Clip}_H(MQ^\star) + \gamma b_k\left(s,a,V^\star\right) \\
&= \widehat{\mathcal{T}}_k Q^\star,
\end{aligned}$$

where the inequality holds under $\mathcal{E}$.

It remains to show that $\mathcal{E}$ occurs with high probability. Fix $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and suppose $n_k(s,a) = n$ for some $n \geq 2$. Note that the $n_k(s,a) = 1$ case trivially always holds.

Observe that $P^k_{s,a} V^\star = \frac{1}{n}\sum_{i=1}^n Z_i$, where $Z_1, \ldots, Z_n$ are i.i.d. multinoulli random variables that take the value $V^\star(s')$ with probability $P_{s,a,s'}$ for each s'. Letting Z be i.i.d. from the same distribution as the Z_i's, we have that $\mathbb{E}[Z] = P_{s,a}V^\star$ and in the notation of Lemma 12,

$$\widehat{\mathrm{Var}}_n = \mathbb{V}\left(\widehat{P}^k_{s,a}, V^\star\right).$$

Subsequently, Lemma 12 gives us that with probability at least $1-\delta'$,

$$\left|\widehat{P}^k_{s,a} V^\star - P_{s,a} V^\star\right| \leq \sqrt{\frac{2\mathbb{V}\left(\widehat{P}^k_{s,a}, V^\star\right)\log(2/\delta')}{n-1}} + \frac{7H\log(2/\delta')}{3(n-1)} \leq 2\sqrt{\frac{\mathbb{V}\left(\widehat{P}^k_{s,a}, V^\star\right)U}{n}} + \frac{14HU}{3n},$$

where the second inequality follows from the fact that $\frac{1}{x-1} \leq \frac{2}{x}$ for $x \geq 2$, as well as the fact that $\log(2/\delta') \leq U$.

Taking a union bound over all possible s, a, n thus gives us that $\mathcal{E}$ holds with probability at least $1 - SAT\delta'$. ∎

H.5. Proof of Lemma 24

Proof We start by proving Part 1, that $\sum_{t:(s,a)=(s_t,a_t)} \frac{1}{\sqrt{n_{k_t}(s,a)}} \leq 2\sqrt{2n_{\text{tot}}(s,a)}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This follows almost immediately from the doubling trick:

$$\sum_{t:(s,a)=(s_t,a_t)} \frac{1}{\sqrt{n_{k_t}(s,a)}} \leq \sqrt{2} \sum_{i=1}^{n_{\text{tot}}(s,a)} \frac{1}{\sqrt{i}} \leq 2\sqrt{2n_{\text{tot}}(s,a)}.$$

Next we prove Part 3, that $\sum_{t=1}^{T} \frac{1}{n_{k_t}(s_t,a_t)} \leq SAU$. Observe that

$$\sum_{t=1}^{T} \frac{1}{n_{k_t}(s_t,a_t)} = \sum_{t=1}^{T} \sum_{s,a} \frac{\mathbb{I}\left((s,a) = (s_t,a_t)\right)}{n_{k_t}(s,a)} = \sum_{s,a} \sum_{t=1}^{T} \frac{\mathbb{I}\left((s,a) = (s_t,a_t)\right)}{n_{k_t}(s,a)}.$$

Now, fix $(s,a) \in \mathcal{S} \times \mathcal{A}$. We have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\mathbb{I}\left((s,a) = (s_t,a_t)\right)}{n_{k_t}(s,a)} &= \sum_{t=1}^{T} \sum_{i=0}^{\lfloor \log_2 T \rfloor} \frac{\mathbb{I}\left((s,a) = (s_t,a_t), 2^i \leq n_{k_t}(s,a) \leq 2^{i+1}-1\right)}{2^i} \\
&= \sum_{i=0}^{\lfloor \log_2 T \rfloor} \sum_{t=1}^{T} \frac{\mathbb{I}\left((s,a) = (s_t,a_t), 2^i \leq n_{k_t}(s,a) \leq 2^{i+1}-1\right)}{2^i} \\
&\leq \sum_{i=0}^{\lfloor \log_2 T \rfloor} \frac{2^i}{2^i} \\
&= \lfloor \log_2 T \rfloor + 1 \leq U,
\end{aligned}$$

where the first inequality holds due to the doubling trick. Namely,

$$\sum_{t=1}^{T} \mathbb{I}\left((s_t,a_t) = (s,a), 2^i \leq n_{k_t}(s,a) \leq 2^{i+1}-1\right) \leq 2^i.$$

Summing over all possible (s,a) gives us the desired bound.

We proceed to proving Part 2. Let $w_1, \ldots, w_T \in \mathbb{R}$ be nonnegative numbers.

$$\begin{aligned}
\sum_{t=1}^{T} \sqrt{\frac{w_t}{n_{k_t}(s_t,a_t)}} &\leq \sqrt{\sum_{t=1}^{T} \frac{1}{n_{k_t}(s_t,a_t)}} \sqrt{\sum_{t=1}^{T} w_t} \\
&\leq \sqrt{SAU \sum_{t=1}^{T} w_t}.
\end{aligned}$$

The first inequality is by Cauchy-Schwarz, and the second inequality is by Part 3 above.

Lastly, we prove Part 4. We have

$$\sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} (1-\gamma)\varepsilon_k = \sum_{k=1}^{m} \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{t_k} = \sum_{k=1}^{m} \frac{N_k}{t_k}.$$

Now, for each $k \in [m]$, it holds that $t_k = 1 + \sum_{\ell=1}^{k-1} N_\ell$. Continuing, we subsequently have

$$\begin{aligned}
\sum_{k=1}^{m} \frac{N_k}{t_k} &= \sum_{k=1}^{m} \frac{N_k}{1 + \sum_{\ell=1}^{k-1} N_\ell} \\
&\leq \sum_{k=1}^{m} \log\left(1 + \frac{N_k}{1 + \sum_{\ell=1}^{k-1} N_\ell}\right) \\
&= \sum_{k=1}^{m} \log\left(\frac{1 + \sum_{\ell=1}^{k} N_\ell}{1 + \sum_{\ell=1}^{k-1} N_\ell}\right) \\
&= \sum_{k=1}^{m} \left(\log\left(1 + \sum_{\ell=1}^{k} N_\ell\right) - \log\left(1 + \sum_{\ell=1}^{k-1} N_\ell\right)\right) \\
&= \log\left(1 + \sum_{\ell=1}^{m} N_\ell\right) - \log(1) \\
&= \log(1 + T) \\
&\leq U.
\end{aligned}$$

■

Appendix I. Proof of Corollary 7

Here we prove Corollary 7.

Proof

After applying Theorem 5 with span bound $H = \sqrt{\frac{T}{S^3 A}}$ and combining with Lemma 2 and the fact that $\left\|V_\gamma^\star\right\|_{\text{sp}} \leq 2\left\|h^\star\right\|_{\text{sp}}$ (Wei et al., 2020) to bound the variance parameter (and taking a union bound to get a failure probability of at most 2δ that these do not both hold) all that remains is to confirm that the resulting regret bound of

$$\text{Regret}(T) \leq \widetilde{O}\left(\sqrt{\left(\left\|h^\star\right\|_{\text{sp}} T + \left\|h^\star\right\|_{\text{sp}}^2\right) SA} + \sqrt{SAT}\right) \tag{11}$$

whenever $T \geq \left\|h^\star\right\|_{\text{sp}}^2 S^3 A$ implies the claim that $\text{Regret}(T) \leq \widetilde{O}\left(\sqrt{(\left\|h^\star\right\|_{\text{sp}} + 1)SAT}\right)$ for $T \geq \left\|h^\star\right\|_{\text{sp}}^2 S^3 A$.

Observe that $T \geq \left\|h^\star\right\|_{\text{sp}}^2 S^3 A \geq \left\|h^\star\right\|_{\text{sp}}^2$ implies that $\left\|h^\star\right\|_{\text{sp}} \sqrt{SA} \leq \sqrt{SAT}$, so (11) simplifies to a bound of

$$\text{Regret}(T) \leq \widetilde{O}\left(\sqrt{\left\|h^\star\right\|_{\text{sp}} SAT} + \sqrt{SAT}\right) = \widetilde{O}\left(\sqrt{\left(\left\|h^\star\right\|_{\text{sp}} + 1\right) SAT}\right)$$

whenever $T \geq \left\|h^\star\right\|_{\text{sp}}^2 S^3 A$, as required.

Note that the second regret bound stated in Corollary 7 follows by considering cases. Indeed, for $T \leq \left\|h^\star\right\|_{\text{sp}}^2 S^3 A$, the regret can be bounded by T, while for $T \geq \left\|h^\star\right\|_{\text{sp}}^2 S^3 A$ we have just

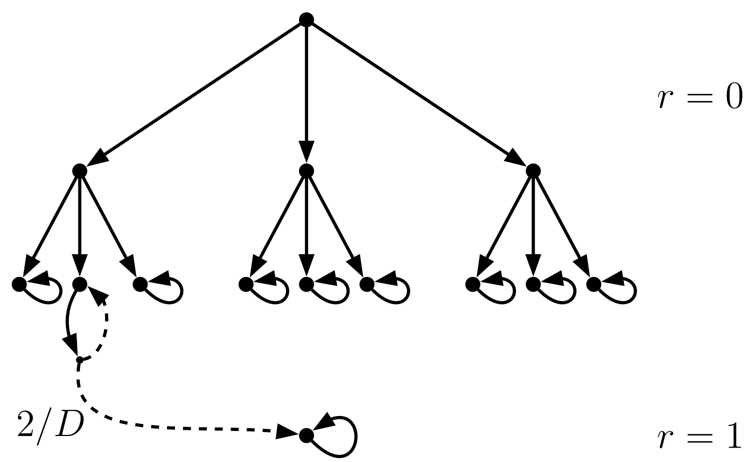

Figure 2: An example of a hard MDP construction for $S = 14$ and $A = 3$. To avoid clutter, we omit an additional deterministic self-loop at each leaf state. We also omit the deterministic actions which transit from the leaf states to the root and from the good state to the root, as these actions only serve to keep the diameter bounded by D.

shown that the regret is bounded by $\widetilde{O}\left(\sqrt{(\|h^\star\|_{\text{sp}}+1)SAT}\right)$. So for all T, the regret is bounded by $\widetilde{O}\left(\sqrt{(\|h^\star\|_{\text{sp}}+1)SAT} + \|h^\star\|_{\text{sp}}^2 S^3 A\right)$. ■

Appendix J. Proof of Theorem 9

Proof Let $S \geq 2$ and $A \geq 2$ be integers, and let $D \geq 4\lceil \log_A S \rceil$. Suppose that $T \leq \frac{1}{32} DSA$.

Let $S' = \lceil \frac{S-1}{2} \rceil$ and $A' = A - 1$, and observe that $T \leq \frac{1}{8} DS'A'$ and $2\lceil \log_A S \rceil \leq \frac{D}{2}$ (we will need these facts later). We construct a family of hard MDPs $\{P_{(s,a)} \mid (s,a) \in [S'] \times [A']\}$ such that each MDP has S states, A actions, and diameter at most D (see Figure 2). First, we use an A-ary tree structure to connect states $1, \ldots, S-1$ by assigning A actions with deterministic transitions at each non-leaf state. In particular, each action goes to a distinct child node. If a non-leaf node has fewer than A children, the remaining actions are deterministic self-loops. Note that we can (and do) choose a tree with depth at most $\lceil \log_A S \rceil$ and at least S' leaf nodes. Next, we let state S be the "good" state, where all actions are deterministic self-loops with reward 1, except for one action which deterministically returns to the root of the tree. At every other node the reward for any action is 0, so the learner's goal is to reach state S.

To reach state S, the learner must search for the correct action in the correct leaf node, which is different in each MDP instance. In the MDP $P_{(s,a)}$, the correct state-action pair is (s, a). That is, at all leaf nodes except state s, A' actions are deterministic self-loops, and the remaining action is a deterministic transition back to the root. State s is identical except that action a transits to the good state with probability $\frac{2}{D}$, and stays in its current state with probability $1 - \frac{2}{D}$. Note that the diameter of this MDP is $2\lceil \log_A S \rceil + \frac{D}{2} \leq D$ as required.

Now fix any horizon-T algorithm. For any $\theta \in [S'] \times [A']$, write $\mathbb{E}_\theta$ to denote the expectation induced by the algorithm when the underlying MDP is P_θ, and write $\mathbb{P}_a$ for the corresponding probability. For each (s, a), let the random variable $N_{(s,a)}$ be the number of times action a is taken in state s, and let N be the total number of visits to the bad states with reward 0. We observe that $\mathrm{Regret}(T) \geq N$, so our goal is to lower bound $\mathbb{E}_\theta[N]$ for some θ.

Let $\mathcal{E}$ denote the event that the learner does **not** observe a transition to the good state. Further let $\theta \in [S'] \times [A']$. We start with the simple observation that since,

$$T \geq \mathbb{E}_\theta[N \mid \mathcal{E}] \geq \sum_{(s,a)\in[S']\times[A']} \mathbb{E}_\theta[N_{(s,a)} \mid \mathcal{E}],$$

it follows that for some (s', a') we have $\mathbb{E}_\theta[N_{(s',a')} \mid \mathcal{E}] \leq \frac{T}{S'A'} \leq \frac{D}{8}$. We then claim that

$$\mathbb{E}_{(s',a')}[N_{(s',a')}] \leq \mathbb{E}_{(s',a')}[N_{(s',a')} \mid \mathcal{E}] = \mathbb{E}_\theta[N_{(s',a')} \mid \mathcal{E}] \leq \frac{D}{8}.$$

The first inequality holds because we can assume WLOG that the algorithm will stay in the good state upon transiting there. The second inequality holds because the algorithm will behave exactly the same on any of the hard MDPs under the event $\mathcal{E}$. So, denoting $\theta' := (s', a')$, we have shown that $\mathbb{E}_{\theta'}[N_{\theta'}] \leq \frac{D}{8}$.

Next, we compute that

$$\mathbb{P}_{\theta'}(\mathcal{E}) \geq \mathbb{P}_{\theta'}\left(\mathcal{E} \mid N_{\theta'} < D/4\right) \mathbb{P}_{\theta'}\left(N_{\theta'} < D/4\right) \geq \left(1 - \frac{2}{D}\right)^{D/4} \frac{1}{2} \geq \frac{1}{4}$$

where the last inequality is due to Bernoulli's inequality (Lemma 16), and the second inequality is due to

$$\mathbb{P}_{\theta'}(N_{\theta'} \geq D/4) \leq \frac{\mathbb{E}_{\theta'}[N_{\theta'}]}{D/4} \leq \frac{1}{2}.$$

Since

$$\mathbb{E}_{\theta'}[N] \geq \mathbb{E}_{\theta'}[N \mid \mathcal{E}]\mathbb{P}_{\theta'}(\mathcal{E}) \geq \frac{T}{4},$$

we conclude that $\mathbb{E}_{\theta'}[\mathrm{Regret}(T)] \geq \frac{T}{4}$. ■

Appendix K. Proof of Theorem 8

In this section we prove the lower bound Theorem 8. First we show the following intermediate result.

Theorem 26 *There exist universal constants $c_1 > 1$ and $0 < c_2 < 1$ such that the following holds. Fix integers $T \geq 1, S \geq 2, A \geq 2$, and fix $B \geq \max\{c_1, 2\lceil\log_A S\rceil\}$. Let `Alg` be any horizon-T algorithm. Then there exist two communicating MDPs P_1, P_2 such that:*

1. *P_1 and P_2 both have S states and A actions.*
2. *$\left\|h^\star_{P_1}\right\|_{\mathrm{sp}} = B$ and $\left\|h^\star_{P_2}\right\|_{\mathrm{sp}} = \frac{1}{2}$.*

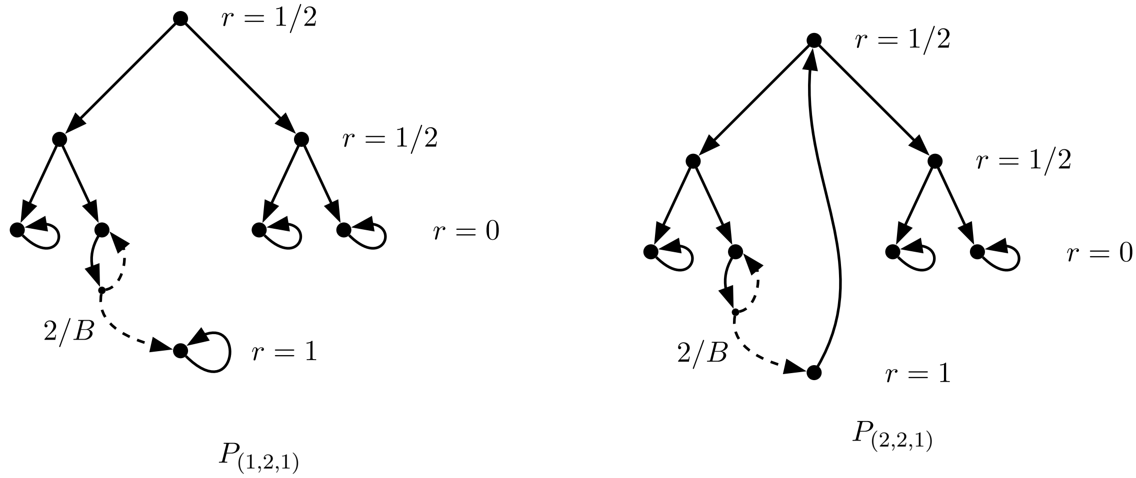

Figure 3: An example of the MDPs used in the proof of Theorem 26. If the transition associated with a state-action pair is deterministic, it is denoted with a solid arrow. If it is stochastic, it is represented as a solid line splitting into multiple dashed arrows to different states, each annotated with the associated probability of that transition. The MDPs are parameterized by $B > 1$. Some actions, such as those which transit from leaf states back to the root state, are omitted.

3. If $\mathbb{E}_{P_1}^{\mathcal{A}lg}[\mathrm{Regret}(T)] < T/4$*, then* $\mathbb{E}_{P_2}^{\mathcal{A}lg}[\mathrm{Regret}(T)] \geq c_2 BSA$*.*

Proof Let $T \geq 1, S \geq 2, A \geq 2$ be integers, and let $B \geq \max\{50, \lceil \log_A S \rceil\}$. Further define $S' = \lceil \frac{S-1}{2} \rceil$ and $A' = A - 1$. We will construct a family of MDPs $\{P_{(i,s,a)} \mid (i, s, a) \in [2] \times [S'] \times [A']\}$, where each $P_{(i,s,a)}$ will have a tree construction similar to that used in the proof of Theorem 9 in Section J (see Figure 3). Specifically, we again use an A-ary tree structure to connect states $1, \dots, S-1$ by assigning A actions with deterministic transitions at each non-leaf state, with each action going to a distinct child node. If a non-leaf node has fewer than A children, the remaining actions are deterministic self-loops. We use a tree with depth at most $\lceil \log_A S \rceil \leq B/2$ and at least S' leaf nodes.

For the MDPs $P_{(1,s,a)}$ and $P_{(2,s,a)}$, in all leaf states other than s, actions $1, \dots, A'$ are deterministic self-loops, and action A is a deterministic transition back to the root. State s is identical except that action a transits to state S outside the tree with probability $2/B$. In state S, actions $2, \dots, A$ are all deterministic transitions back to the root. The reward is 1/2 for any action in any non-leaf state or any deterministic transition from a tree state back to the root. The reward for action 1 in state S is 1, and the reward at any other state-action pair is 0.

The only difference between $P_{(1,s,a)}$ and $P_{(2,s,a)}$ is that in $P_{(1,s,a)}$, action 1 is a deterministic self-loop, while in $P_{(2,s,a)}$, action 1 is a deterministic transition back to the root. Subsequently, it is straightforward to see that the optimal strategy in $P_{(1,s,a)}$ is to take action (s, a) until a transition to state S is observed, then repeatedly take action 1. On the other hand, in $P_{(2,s,a)}$ it is not worth trying to reach state S, so the optimal strategy is to stay in the tree and only take reward 1/2 actions. For a learner to distinguish between the two MDPs, it must reach state S and take action 1, but that will take many time steps when the correct (s,a) is not known beforehand. One can easily verify that the span of the optimal bias function is at most B for all $P_{(1,s,a)}$ and $\frac{1}{2}$ for all $P_{(2,s,a)}$, as required.

We now consider any horizon-T algorithm `Alg`. WLOG we can assume the following:

1. `Alg` is deterministic, meaning that given a sequence of past states and actions, the next action is computed by a deterministic function. This assumption can be justified via a standard argument that any randomized strategy is equivalent to some random choice from the set of all deterministic strategies (Auer et al. (2002), Auer et al. (2008)).
2. Once a transition to state S is observed, `Alg` acts optimally. If this were not the case, the expected regret would only increase.

Since we assume that `Alg` is deterministic, under the event that no transition to state S occurs `Alg`, will always observe the deterministic sequence of state-action pairs $(s^{(1)},a^{(1)}),\dots,(s^{(T)},a^{(T)})$. For any (s,a), we denote by $t_k(s,a)$ the index of the kth occurrence of (s,a) in this sequence, and we denote by $n_t(s,a)$ the number of times (s,a) occurs through index t in this sequence.

For $\theta \in [2]\times[S']\times[A']$, let $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ denote the probability and expectation, respectively, induced by `Alg` when the underlying MDP is P_θ. For each (s,a), let $N_{(s,a)}$ denote the number of times the learner takes action a in state s. Further write $N_{\text{leave}} = \sum_{(s,a)\in[S']\times[A']} N_{(s,a)}$, and let N_{stay} be the number of times the learner takes a reward $\frac{1}{2}$ action. Additionally, let $\mathcal{E}$ denote the event that the algorithm observes a transition to state S.

When the underlying MDP is some $P_{(1,s,a)}$ the optimal gain is 1, and hence the algorithm adds 1 to the regret any time it tries to reach state S (i.e. some state-action in $[S']\times[A']$ and adds 1/2 to the regret any time it takes a reward 1/2 action within the tree. In particular,

$$\operatorname{Regret}(T, P_{(1,s,a)}, \texttt{Alg}) \geq T - \frac{1}{2}N_{\text{stay}} - N_{(S,1)}.$$

Similarly, when the underlying MDP is some $P_{(2,s,a)}$, the optimal gain is 1/2, and hence the algorithm adds 1/2 to the regret any time it tries to reach state S. It may subtract 1/2 from the regret in the event that it does $(S,1)$. Consequently,

$$\operatorname{Regret}(T, P_{(2,s,a)}, \texttt{Alg}) \geq \frac{1}{2}N_{\text{leave}} - \frac{1}{2}N_{(S,1)} \geq \frac{1}{2}N_{\text{leave}} - \frac{1}{2},$$

with the second inequality due to the fact that the algorithm will try $(S,1)$ at most once in $P_{(2,s,a)}$.

Now, let $(s,a)\in[S']\times[A']$. Assuming that $\mathbb{E}_{(1,s,a)}[\operatorname{Regret}(T)] < \frac{T}{4}$, we will work to show a lower bound on $n_T(s,a)$. Under our assumption, we have

$$\begin{aligned}
\frac{T}{4} &> T - \frac{1}{2}\mathbb{E}_{(1,s,a)}\left[N_{\text{stay}}\right] - \mathbb{E}_{(1,s,a)}\left[N_{(S,1)}\right] \\
&\implies \frac{1}{2}\mathbb{E}_{(1,s,a)}[N_{\text{stay}}] + \mathbb{E}_{(1,s,a)}[N_{(S,1)}] > \frac{3T}{4} \\
&\implies \mathbb{E}_{(1,s,a)}[N_{(S,1)}] > \frac{T}{4},
\end{aligned}$$

with the second implication following from the trivial fact that $N_{\text{stay}} \leq T$. Furthermore, since $N_{(S,1)} \leq T\mathbb{I}(\mathcal{E})$, we have

$$\frac{T}{4} < \mathbb{E}_{(1,s,a)}[N_{(S,1)}] \leq T\mathbb{P}_{(1,s,a)}(\mathcal{E}) \implies \mathbb{P}_{(1,s,a)}(\mathcal{E}) > \frac{1}{4}.$$

We have shown that there is a constant probability of observing a transition to state S, and we will use this fact to show that $N_{(s,a)}$ is large with constant probability. Towards this end, we compute

$$\begin{aligned}\mathbb{P}_{(1,s,a)}\left(\mathcal{E}^c\right) &\geq \mathbb{P}_{(1,s,a)}\left(\mathcal{E}^c \mid N_{(s,a)} < \frac{B}{10}\right)\mathbb{P}_{(1,s,a)}\left(N_{(s,a)} < \frac{B}{10}\right)\\ &\geq \left(1-\frac{2}{B}\right)^{\frac{B}{10}}\mathbb{P}_{(1,s,a)}\left(N_{(s,a)} < \frac{B}{10}\right)\\ &\geq \frac{4}{5}\mathbb{P}_{(1,s,a)}\left(N_{(s,a)} < \frac{B}{10}\right),\end{aligned}$$

with the last inequality holding due to Bernoulli's inequality (Lemma 16). Hence,

$$\frac{1}{4} < \mathbb{P}_{(1,s,a)}(\mathcal{E}) = 1 - \mathbb{P}_{(1,s,a)}\left(\mathcal{E}^c\right) \leq 1 - \frac{4}{5}\mathbb{P}_{(1,s,a)}\left(N_{(s,a)} < \frac{B}{10}\right),$$

which implies that

$$\mathbb{P}_{(1,s,a)}\left(N_{(s,a)} < \frac{B}{10}\right) < \frac{3}{4}\cdot\frac{5}{4} = \frac{15}{16}.$$

Therefore, $\mathbb{P}_{(1,s,a)}(N_{(s,a)} \geq \frac{B}{10}) \geq \frac{1}{16}$, which implies that $n_T(s,a) \geq \frac{B}{10}$ (otherwise $\mathbb{P}_{(1,s,a)}(N_{(s,a)} \geq \frac{B}{10})$ would be 0). Since (s,a) was arbitrary, we have shown that $n_T(s,a) \geq \frac{B}{10}$ for all $(s,a) \in [S'] \times [A']$.

Our next step is to derive a lower bound on $\mathbb{E}_{(1,s,a)}[N_{\text{leave}}]$ for some (s,a). Observe that for any (s,a), we have

$$\mathbb{E}_{(1,s,a)}[N_{\text{leave}}] = \sum_{t=1}^{T} t\,\mathbb{P}_{(1,s,a)}(N_{\text{leave}} = t).$$

Furthermore, for $t < T$, $N_{\text{leave}} = t$ occurs precisely when $(s_t, a_t) = (s,a)$, all previous occurrences of (s,a) do not result in a transition to S, and this occurrence of (s,a) does result in a transition to S. $N_{\text{leave}} = T$ is similar except that it does not require a transition to S. In other words, we have

$$\mathbb{E}_{(1,s,a)}[N_{\text{leave}}] \geq \sum_{t=1}^{T} t\,\mathbb{I}\Big((s_t,a_t) = (s,a)\Big)\left(1-\frac{2}{B}\right)^{n_t(s,a)-1}\frac{2}{B}.$$

Writing $t_k(s,a)$ to be the kth time (s,a) occurs, we can lower bound this sum by

$$\sum_{k=1}^{\lfloor B/8 \rfloor} t_k(s,a)\left(1-\frac{2}{B}\right)^{k-1}\frac{2}{B}.$$

The following technical lemma, the proof of which we postpone, shows that we can bound this sum for at least one (s,a). Intuitively, the sum will be large enough when the $t_k(s,a)$ are large, and because all (s,a) occur many times, there is at least one (s,a) whose indices are sufficiently large.

Lemma 27 *Let $B > 0$ and $c \in (0,1)$ such that $\lceil cB \rceil \geq 2$. Let M be a positive integer. Let $z_1, z_2, \ldots$ be a sequence of integers such that each $i \in \{1,\ldots,M\}$ occurs at least cB times. Let $t_k(i)$ be the index of the kth occurrence of value i. Then*

$$\max_{i\in\{1,\ldots,M\}} \sum_{k=1}^{\lfloor cB \rfloor} t_k(i)\left(1-\frac{2}{B}\right)^{k-1}\frac{2}{B} \geq \frac{c^2(1-2c)}{2}BM.$$

An application of Lemma 27 with $c = 1/10$ and $M = S'A'$ gives us that there exists some (s', a') satisfying

$$\mathbb{E}_{(1,s',a')}[N_{\text{leave}}] \geq \frac{1}{250} BS'A'.$$

Finally, it is not hard to see that $\mathbb{E}_{(1,s',a')}[N_{\text{leave}}] = \mathbb{E}_{(2,s',a')}[N_{\text{leave}}]$, so

$$\mathbb{E}_{(2,s',a')}[\text{Regret}(T, P_{(2,s,a)}, \texttt{Alg})] \geq \mathbb{E}_{(2,s',a')}\left[\frac{1}{2}N_{\text{leave}} - \frac{1}{2}\right] \geq \frac{BS'A'}{250} - \frac{1}{2} \geq \frac{BSA}{2000} - \frac{1}{2} \geq \frac{BSA}{4000},$$

where the final inequality is due to the fact that $B \geq 500 \implies \frac{BSA}{4000} \geq \frac{1}{2}$. We conclude that the requirements of the theorem hold with $c_1 = 500, c_2 = 1/4000, P_1 = P_{(1,s',a')}$, and $P_2 = P_{(2,s',a')}$. ■

We now prove Lemma 27.

Proof Fix arbitrary $B > 0$ and $c \in (0, 1)$ satisfying $\lceil cB \rceil \geq 2$, let $M \in \mathbb{Z}_{\geq 1}$, and let $z_1, z_2, \ldots$ be a sequence of integers such that $\sum_{j=1}^{\infty} \mathbb{I}(z_j = i) \geq cB$ for all $i \in \{1, \ldots, M\}$. For ease of presentation, write $B' := \lceil cB \rceil$ and $w_k := \left(1 - \frac{2}{B}\right)^{k-1} \frac{2}{B}$ so that our goal is to bound

$$\max_{i \in \{1,\ldots,M\}} \sum_{k=1}^{B'} t_k(i) w_k.$$

First, we use that the max is greater the average, so that

$$\max_{i \in \{1,\ldots,M\}} \sum_{k=1}^{B'} t_k(i) w_k \geq \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{B'} t_k(i) w_k.$$

Observing that $\{t_k(i)\} = \{1, \ldots, MB'\}$, we then reindex the sum and write

$$\frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{B'} t_k(i) w_k = \frac{1}{M} \sum_{t=1}^{MB'} t\, w_{n(t)},$$

where $n(t)$ is defined as the number of times that z_t appears through index t. Furthermore, the rearrangement inequality (Lemma 17) gives us that

$$\frac{1}{M} \sum_{t=1}^{MB'} t\, w_{n(t)} \geq \frac{1}{M} \sum_{t=1}^{MB'} t\, w_{\lceil t/M \rceil} \geq \frac{1}{M} \sum_{j=1}^{B'} M((j-1)M + 1) w_j \geq M \sum_{j=1}^{B'} (j-1) w_j.$$

It remains to analyze $M \sum_j (j-1) w_j$. Note that for any $j \in \{1, \ldots, B'\}$, Bernoulli's inequality (Lemma 16) gives us

$$w_j = \left(1 - \frac{2}{B}\right)^{j-1} \frac{2}{B} \geq \left(1 - \frac{2}{B}\right)^{B'-1} \frac{2}{B} \geq \left(1 - \frac{2(B'-1)}{B}\right) \frac{2}{B} \geq \frac{2 - 4c}{B}.$$

Consequently,

$$M \sum_{j=1}^{B'} (j-1) w_j \geq \frac{M(2-4c)}{B} \sum_{j=1}^{B'-1} j = \frac{M(2-4c)}{B} \frac{B'(B'-1)}{2} \geq M \frac{1}{B} \frac{2-4c}{2} cB \frac{cB}{2} \geq \frac{c^2(1-2c)}{2} BM.$$

■

Since we have fully established Theorem 26, we can turn to proving Theorem 8.

Proof Let $S \geq 2$ and $A \geq 2$ be integers. Fix some $\alpha \in [1, 2)$, and let $T > SA(c_4\beta_T)^{\frac{4}{2-\alpha}}$, where $c_4 \geq 1$ is a universal constant to be defined later. Suppose that some horizon-T algorithm $\texttt{Alg}$ has for all MDPs P,

$$\mathbb{E}[\text{Regret}(T, P, \texttt{Alg})] \leq \sqrt{\beta_T \left\|h_P^\star\right\|_{\text{sp}} SAT} + \beta_T SA \left\|h_P^\star\right\|_{\text{sp}}^\alpha .$$

We want to use Theorem 26 to show a contradiction. Hence we need a choice of B such that

$$\sqrt{\beta_T BSAT} + \beta_T SAB^\alpha < T/4 \tag{12}$$

$$\sqrt{\beta_T SAT/2} + \beta_T SA/2^\alpha < c_2 BSA. \tag{13}$$

We will set B in terms of T, S, A to be as small as possible such that the second inequality holds, and then show that the first inequality holds under the assumed conditions.

Assuming that $T > \beta_T SA$, we can then derive that

$$\sqrt{\beta_T SAT/2} + \beta_T SA/2^\alpha < 2\sqrt{\beta_T SAT/2},$$

so we can satisfy (13) by setting $c_2 BSA = 2\sqrt{\beta_T SAT/2} \iff B = \sqrt{\frac{c_3\beta_T T}{SA}}$, where $c_3 := \frac{2}{c_2\sqrt{2}}$. Now checking that our choice of B admits (12), we calculate the equivalence

$$\begin{aligned}
&\sqrt{\beta_T BSAT} + \beta_T SAB^\alpha < T/4 \\
\iff &\left(c_3^2\beta_T^3 T^3 SA\right)^{\frac{1}{4}} + \beta_T SA\left(\sqrt{\frac{c_3\beta_T T}{SA}}\right)^\alpha < T/4.
\end{aligned}$$

Hence a sufficient condition for (12) is that both of the following are true:

$$\left(c_3^2\beta_T^3 T^3 SA\right)^{\frac{1}{4}} < T/8 \tag{14}$$

$$\beta_T SA\left(\sqrt{\frac{c_3\beta_T T}{SA}}\right)^\alpha < T/8. \tag{15}$$

The condition (14) is equivalent to

$$\begin{aligned}
&\left(c_3^2\beta_T^3 T^3 SA\right)^{\frac{1}{4}} < T/8 \\
\iff &c_3^2\beta_T^3 T^3 SA < T^4/8^4 \\
\iff &T > 8^4 c_3^2\beta_T^3 SA.
\end{aligned}$$

Defining $c_4 := 8c_3$, for condition (15) we compute

$$\begin{aligned}
&\beta_T SA\left(\sqrt{\frac{c_3\beta_T T}{SA}}\right)^\alpha < T/8 \\
\iff &T^{\frac{2-\alpha}{2}} > 8c_3^{\frac{\alpha}{2}}\beta_T^{\frac{2+\alpha}{2}}(SA)^{\frac{2-\alpha}{2}}
\end{aligned}$$

$$\iff T > SA8c_3^{\frac{\alpha}{2-\alpha}}\beta_T^{\frac{2+\alpha}{2-\alpha}}$$
$$\Longleftarrow T > SA(c_4\beta_T)^{\frac{4}{2-\alpha}}$$

where the final implication is due to $\alpha < 2$ and $c_3 > 1$.

So the desired contradiction holds as long as $T > \beta_T SA$, $T > 8^4 c_3^2 \beta_T^3 SA$, and $T > SA(c_4\beta_T)^{\frac{4}{2-\alpha}}$, but all of these conditions are clearly implied by $T > SA(c_4\beta_T)^{\frac{4}{2-\alpha}}$, since $\alpha \geq 1$ implies that $SA(c_4\beta_T)^{\frac{4}{2-\alpha}} \geq SA(8c_3\beta_T)^4$. ■