Narrowing the Gap between Adversarial and Stochastic MDPs via Policy Optimization

Daniil Tiapkin, Evgenii Chzhen, Gilles Stoltz

Narrowing the Gap between Adversarial and Stochastic MDPs via Policy Optimization

Daniil Tiapkin
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Centre de Mathématiques Appliquées — CNRS — École polytechnique
Institut Polytechnique de Paris, route de Saclay, 91128, Palaiseau
daniil.tiapkin@polytechnique.edu

Evgenii Chzhen **Gilles Stoltz**
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
{evgenii.chzhen, gilles.stoltz}@universite-paris-saclay.fr

Abstract

In this paper, we consider the problem of learning in adversarial Markov decision processes [MDPs] with an oblivious adversary in a full-information setting. The agent interacts with an environment during T episodes, each of which consists of H stages, and each episode is evaluated with respect to a reward function that will be revealed only at the end of the episode. We propose an algorithm, called APO-MVP, that achieves a regret bound of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\sqrt{SAT})$, where S and A are sizes of the state and action spaces, respectively. This result improves upon the best-known regret bound by a factor of $\sqrt{S}$, bridging the gap between adversarial and stochastic MDPs, and matching the minimax lower bound $\Omega(\sqrt{H^3SAT})$ as far as the dependencies in S, A, T are concerned. The proposed algorithm and analysis completely avoid the typical tool given by occupancy measures; instead, it performs policy optimization based only on dynamic programming and on a black-box online linear optimization strategy run over estimated advantage functions, making it easy to implement. The analysis leverages two recent techniques: policy optimization based on online linear optimization strategies (Jonckheere et al., 2023) and a refined martingale analysis of the impact on values of estimating transitions kernels (Zhang et al., 2023).

1 Introduction

We study adversarial Markov decision processes [MDPs], introduced by Even-Dar et al. (2009) and Yu et al. (2009), in an episodic setup with full monitoring. Unlike the standard setup, the reward function is not known to the learner beforehand and is revealed sequentially at the end of each episode.

To deal with this problem, many earlier works relied on online linear optimization [OLO] strategies (see the monograph by Cesa-Bianchi and Lugosi, 2006 for a survey) in the space of so-called occupancy measures (Zimin and Neu, 2013). These occupancy measures concern the state-action pairs within an episode induced by a given policy and transition kernel. This family of algorithms, known as O-REPS, has been extended to handle unknown transition kernels and bandit feedback by several studies (Rosenberg and Mansour, 2019a,b; Jin et al., 2020, 2021), using an exploration mechanism similar to UCRL2 (Auer et al., 2008). However, this type of exploration leads to an additional $\sqrt{S}$ factor in the regret, where S is the number of states, compared to the state of the art in the

Preprint. Under review.

non-adversarial case (Azar et al., 2017; Dann et al., 2017; Jin et al., 2018). Furthermore, O-REPS-based approaches require solving a high-dimensional convex program at each episode, resulting in a non-explicit policy update.

A recent line of work has focused on policy-optimization-based approaches for adversarial MDPs (Cai et al., 2020; Shani et al., 2020; Zanette et al., 2021; He et al., 2022; Lancewicki et al., 2022; Sherman et al., 2023; Zhong and Zhang, 2024). These algorithms use a more practical approach combining dynamic programming with optimization directly in the policy space, instead of working with occupancy measures. This approach actually is related to the well-known TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017) algorithms, heavily used by practitioners. However, to the best of our knowledge, policy-optimization-based approaches also suffer from an additional $\sqrt{S}$ factor in the regret bound when specialized to finite MDP settings.

To date, the question of whether dependency on the number of states can be matched between adversarial and stochastic cases remains open. We take the first step towards unifying these rates.

In this work, we use a black-box policy optimization approach, departing from the current state-of-the-art algorithms based on occupancy measures. This approach of policy optimization based on running online linear optimization strategies in a black-box way on estimated advantage functions was recently introduced by Jonckheere et al. (2023). The dynamic programming counterpart of our algorithm, as well as a part of the analysis, relies on the Monotonic Value Propagation [MVP] algorithm of Zhang et al. (2021, 2023), which allowed to achieve optimal regret bounds up to second-order terms. However, since we do not yet target the lower-order terms, we significantly simplify their approach and provide an arguably more transparent exposition thereof. All in all, our policy-optimization-based algorithm achieves a $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\sqrt{SAT})$ regret, where A is the number of actions and T is the number of episodes, and where we recall that H is the length of an episode and S is the number of states. This result improves on the previous regret bound of Rosenberg and Mansour (2019a) by a factor of $\sqrt{S}$, although it introduces an additional $\mathrm{poly}(H)$ factor. It also matches the minimax lower bound derived for the stochastic case (Jin et al., 2018; Domingues et al., 2021) in all parameters except H.

Therefore, we demonstrate that while *policy optimization is already known to be practical, it is also more sample-efficient in large state spaces compared to existing O-REPS-based methods.*

Contributions. This paper puts forward the following contributions, in the setting of adversarial episodic MDPs with full information: i) we introduce a algorithm called Adversarial Policy Optimization based on Monotonic Value Propagation (APO-MVP) that relies on a black-box online linear optimization solver and on dynamic programming, making it easier to implement in practice; ii) we demonstrate that the proposed algorithm is able to achieve a $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\sqrt{TSA})$ regret, improving on the previously best-known dependency on the number of states S and achieving the minimax lower bound $\Omega(\sqrt{H^3SAT})$ in all parameters, except H; iii) our analysis is modular and rather general, providing high flexibility and providing new tools for the study of adversarial MDPs with policy optimization.

Notation. For any positive integer N, we denote by $[N] \stackrel{\text{def}}{=} \{1, \dots, N\}$ and $[N]^* \stackrel{\text{def}}{=} \{0, 1, \dots, N\}$ the sets of the first positive and non-negative integers not greater than N, respectively. For $a, b \in \mathbb{R}$, we denote by $a \vee b$ and $a \wedge b$ the maximum and the minimum between a and b, respectively. For a finite set $\mathcal{E}$, we denote by $\Delta(\mathcal{E})$ the set of probability distributions over $\mathcal{E}$. We refer to natural logarithms by $\log$ and to logarithms in base 2 by $\log_2$. When we write $\widetilde{\mathcal{O}}(\,\cdot\,)$, we hide all absolute constants and polylog multiplicative terms.

2 Problem formulation

An H–episodic (obliviously) adversarial Markov decision process (MDP), where $H \geqslant 1$, is determined by a finite set of states $\mathcal{S}$, with cardinality S, a finite set of actions $\mathcal{A}$, with cardinality A, a sequence $\boldsymbol{P} = (P_h)_{h\in[H-1]}$ of Markov transition kernels $P_h \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and by a (potentially adversarially chosen) fixed-in-advance sequence $(\boldsymbol{r}_t)_{t\geqslant 1}$ of bounded time-inhomogeneous H–episodic reward functions. Each reward function is of the form $\boldsymbol{r}_t = (r_{t,h})_{h\in[H]}$, where $r_{t,h} \colon \mathcal{S}\times\mathcal{A} \to [0,1]$. For simplicity (and with no loss of generality, up to resorting to some doubling

trick), we assume that the number T of episodes is fixed and known. We set some initial state s_1 for each episode. At each episode t and at each stage h, based on past observations, the learner picks a stage policy $\pi_{t,h}\colon \mathcal{S} \to \Delta(\mathcal{A})$ to draw the action. The interaction with the environment is therefore governed by the following protocol. For each episode $t = 1, \dots, T$:

1. Reset state $s_{t,1} = s_1$;
2. Start new episode — for each stage $h = 1, \dots, H$:
   - Pick a policy $\pi_{t,h}$ and sample $a_{t,h} \sim \pi_{t,h}(\cdot \mid s_{t,h})$;
   - If $h \leqslant H-1$, move to the next state $s_{t,h+1} \sim P_h(\cdot \mid s_{t,h}, a_{t,h})$;
3. Observe the reward function $\boldsymbol{r}_t = (r_{t,h})_{h\in[H]}$.

We compare the performance of the policies $\boldsymbol{\pi}_t = (\pi_{t,h})_{h\in[H]}$ picked to the one achieved by resorting to a static policy $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$ in each episode, in terms of value functions. We define the value function of a policy $\boldsymbol{\pi}$, at episode $t \in [T]$, and started from step $h \in [H]$, as

$$V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{j=h}^{H} r_{t,j}(s_{t,j}, a_{t,j}) \;\middle|\; s_{t,h} = s\right];$$

we recall the environment (reward functions, transition kernels) in the notation for value functions and expectations, as environments will vary in the algorithm and analysis. The regret of the learner is defined as the difference between the accumulated value of the best static policy in hindsight and the gained value of the learner, that is,

$$R_T \stackrel{\text{def}}{=} \max_{\boldsymbol{\pi}} \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t,\boldsymbol{P}}(s_1)\right). \tag{1}$$

The goal of the learner is to design policies $(\boldsymbol{\pi}_t)_{t\in[T]}$ minimizing the above-defined regret.

Additional notation. For the analysis, we define Q–value functions and remind Bellman's equations. For any policy $\boldsymbol{\pi}$, we define the Q–value function at episode $t \in [T]$, and started from step $h \in [H]$, as

$$Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{j=h}^{H} r_{t,j}(s_{t,j}, a_{t,j}) \;\middle|\; s_{t,h} = s,\ a_{t,h} = a\right].$$

The advantage function is in turn defined as $A_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \stackrel{\text{def}}{=} Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) - V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s)$.

We use the usual convention that for a transition kernel $K\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, a policy $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$, and two functions $f\colon \mathcal{S} \to \mathbb{R}$ and $g\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$,

$$K\cdot f(s,a) \stackrel{\text{def}}{=} \sum_{s'\in\mathcal{S}} K(s' \mid s,a)\, f(s') \qquad \text{and} \qquad \pi\cdot g(s) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \pi(a \mid s)\, g(a,s)\,.$$

Then, Bellman's equations read for all episodes $t \in [T]$, steps $h \in [H-1]$, and policies $\boldsymbol{\pi}$, as

$$Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) = r_{t,h}(s,a) + P_h\cdot V_{h+1}^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \qquad \text{and} \qquad V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) = \pi_{t,h}\cdot Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s)\,,$$

while for $h = H$, one has $Q_H^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) = r_{t,H}(s,a)$ as well as $V_H^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) = \pi_{t,H}\cdot Q_H^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s)$.

3 Algorithm and main result

In this section, we first describe our algorithm, `APO-MVP`, which stands for Adversarial Policy Optimization based on Monotonic Value Propagation, and then state the performance bound obtained.

3.1 Algorithm `APO-MVP`

Let us start with a high-level description of the proposed algorithm, and details will be provided below. Similarly to Rosenberg and Mansour (2019a), our algorithm proceeds in random epochs

$\mathcal{E}_e \subseteq [T]$ indexed by $e = 1, 2, \dots, m(T)$ of random lengths denoted by $E_1, \dots, E_{m(T)} \in [T]$, i.e., $E_e \overset{\text{def}}{=} |\mathcal{E}_e|$. At the beginning of each epoch e,

$$\text{estimates } \widehat{\boldsymbol{P}}^{(e)} = \Big(\widehat{P}_h^{(e)}\Big)_{h\in[H-1]} \qquad \text{and} \qquad \text{bonus functions } \boldsymbol{b}^{(e)} = \big(b_h^{(e)}\big)_{h\in[H]},$$

where $b_h^{(e)} \colon \mathcal{S} \times \mathcal{A} \to [0, H]$, are computed and will be used during the entire epoch e, as detailed in (3)–(4) and in Fact 2. Actually, $b_H^{(e)}$ will be identically null, but we consider it so that the bonus functions $\boldsymbol{b}^{(e)}$ may be added to reward functions $\boldsymbol{r}_t$.

Within-epoch statement. We now explain the updates and choices made at each episode $t \in [T]$. First, the policies $\boldsymbol{\pi}_t$ are picked, as indicated below. Then, denoting by e_t the epoch such that $t \in \mathcal{E}_{e_t}$, at the end of episode t, i.e., once $\boldsymbol{r}_t$ is revealed, we build optimistic estimates of the Q–value and value functions in a backward fashion, based on Bellman's equations: for $h = H$,

$$\widehat{Q}_{t,H}(s,a) \overset{\text{def}}{=} r_{t,H}(s,a) \qquad \text{and} \qquad \widehat{V}_{t,H}(s) \overset{\text{def}}{=} \pi_{t,H} \cdot \widehat{Q}_{t,H}(s)$$

and for $h \in [H-1]$,

$$\widehat{Q}_{t,h}(s,a) \overset{\text{def}}{=} r_{t,h}(s,a) + b_h^{(e_t)}(s,a) + \widehat{P}_h^{(e_t)} \cdot \widehat{V}_{t,h+1}(s,a) \qquad \text{and} \qquad \widehat{V}_{t,h}(s) \overset{\text{def}}{=} \pi_{t,h} \cdot \widehat{Q}_{t,h}(s)\,.$$

For all $h \in [H]$, estimated advantage functions are defined by $\widehat{A}_{t,h}(s,a) \overset{\text{def}}{=} \widehat{Q}_{t,h}(s,a) - \widehat{V}_{t,h}(s)$, and we denote $\widehat{A}_{t,h}(s,\,\cdot\,) \overset{\text{def}}{=} \big(\widehat{A}_{t,h}(s,a)\big)_{a\in\mathcal{A}}$.

The policies $\boldsymbol{\pi}_t = (\pi_{t,h})_{h\in[H]}$ are picked based on an online linear optimization [OLO] strategy $\varphi = (\varphi_t)_{t\geqslant 1}$, which is a sequence of functions $\varphi_t \colon (\mathbb{R}^{\mathcal{A}})^{t-1} \to \Delta(\mathcal{A})$ satisfying some performance guarantee stated in Definition 1. (The function φ_1 is constant.) We run SH such strategies in parallel as follows: for all $s \in \mathcal{S}$ and $h \in [H]$,

$$\pi_{t,h}(\,\cdot \mid s) = \varphi_t\Big(\Big(\widehat{A}_{\tau,h}(s,\,\cdot\,)\Big)_{\tau\in\mathcal{E}_{e_t}\cap[t-1]}\Big)\,. \tag{2}$$

Note that these choices indeed exploit information available at the beginning of episode t (at the end of episode $t-1$), and rely only on the estimated advantage functions of the current epoch. One may see φ as an adaptive version of PPO- or TRPO-like updates (Schulman et al., 2015, 2017). We will consider, for the sake of concreteness, the polynomial-potential- and exponential-potential-based strategies (see Examples 1 and 2 and references therein), but many other OLO strategies would work. Appendix A states closed-form expressions of the policies constructed with these strategies.

Remark 1 (Two technical remarks)**.** *The kernel estimate and the bonus functions are fixed within a given epoch, which is the main reason why we are able to provide a black-box treatment of the problem relying on any OLO strategy satisfying Definition 1.*

As the reward function takes values in $[0,1]$, the Q–value functions are bounded by H, and it is a common practice in the case of non-adversarial reward functions to clip the estimates to $[0,H]$ (see, e.g., Azar et al., 2017), which only helps. Unfortunately, our adversarial analysis related to the OLO part of the proof heavily relies on the so-called performance-difference lemma (Kakade and Langford, 2002), which does not hold once clipping is involved. Thus, we opt out from clipping, paying an additional H factor at the eventual regret bound of Theorem 1 but still improving the dependency on S. Successful incorporation of clipping could improve the regret by an H multiplicative factor.

Epoch switching. The epoch-switching conditions below were also considered and analyzed by Zhang et al. (2023). We introduce the following empirical counts, for all episodes $t \in [T]$, stages $h \in [H-1]$, state–action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$, and states $s' \in \mathcal{S}$:

$$n_{t,h}(s,a,s') \overset{\text{def}}{=} \sum_{\tau=1}^{t} \mathbb{I}\big\{(s_{\tau,h},, a_{\tau,h}, s_{\tau,h+1}) = (s,a,s')\big\} \quad \text{and} \quad n_{t,h}(s,a) \overset{\text{def}}{=} \sum_{s'\in\mathcal{S}} n_{t,h}(s,a,s')\,.$$

We start at epoch $e = 1$. When for some $(t,h,s,a) \in [T] \times [H-1] \times \mathcal{S} \times \mathcal{A}$, the count $n_{t,h}(s,a)$ equals $2^{\ell-1}$ for some integer $\ell \geqslant 1$, the next epoch is started at episode $t+1$.

Now, for each episode $t \in [T]$, stage $h \in [H-1]$, and state–action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we denote by

$$\mathrm{sw}_{t,h}(s,a) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } n_{t,h}(s,a) = 0\,, \\ \max\bigl\{\tau \in [t] : n_{\tau,h}(s,a) \text{ is of the form } 2^{\ell-1}, \text{ for } \ell \geqslant 1\bigr\} & \text{if } n_{t,h}(s,a) > 0\,, \end{cases}$$

the last episode when an epoch switch took place because, among others, of (s,a). We refer to the largest value of ℓ in the maximum defining $\mathrm{sw}_{t,h}(s,a)$ by

$$\ell_{t,h}(s,a) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } n_{t,h}(s,a) = 0\,, \\ \max\bigl\{\ell \geqslant 1 : n_{t,h}(s,a) \geqslant 2^{\ell-1}\bigr\} & \text{if } n_{t,h}(s,a) > 0\,. \end{cases}$$

The values $\ell_{t,h}(s,a)$ index local epochs for a given state–action pair (s,a), while the global epochs e_t are defined based on all local epochs. (More details may be found in Section 4.2.)

Fact 1. *By design, the functions* $\mathrm{sw}_{t-1,h}$ *and* $\ell_{t-1,h}$ *defined above (note the subscripts* $t-1$ *here) are identical for all episodes* $t \in \mathcal{E}_e$ *of a given epoch* e*.*

We may now define the estimated transition kernels $\widehat{\boldsymbol{P}}_t$ and bonus functions $\boldsymbol{b}_t$: for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $s' \in \mathcal{S}$, first for all $h \in [H-1]$,

$$\widehat{P}_{t,h}(s' \mid s,a) \overset{\text{def}}{=} \begin{cases} 1/S & \text{if } n_{\tau,h}(s,a) = 0\,, \\ \dfrac{n_{\tau,h}(s,a,s')}{n_{\tau,h}(s,a)} & \text{if } n_{\tau,h}(s,a) \geqslant 1\,, \end{cases} \quad \text{with } \tau = \mathrm{sw}_{t-1,h}(s,a)\,, \tag{3}$$

$$b_{t,h}(s,a) \overset{\text{def}}{=} \begin{cases} H & \text{if } \ell = 0\,, \\ \sqrt{\dfrac{2H^2 \log\bigl(2SATH\log_2(2T)/\delta\bigr)}{2^{\ell-1}}} \wedge H & \text{if } \ell \geqslant 1\,, \end{cases} \quad \text{with } \ell = \ell_{t-1,h}(s,a)\,; \tag{4}$$

we also set, by convention, $b_{t,H}(s,a) = 0$. In particular, $\widehat{P}_{t,h}(\cdot \mid s,a)$ corresponds to an empirical frequency vector based on $2^{\ell_{t-1,h}(s,a)-1}$ values when $\ell_{t-1,h}(s,a) \geqslant 1$.

Fact 2. *By Fact 1, the above-defined* $\widehat{\boldsymbol{P}}_t$ *and* $\mathbf{b}_t$ *are indeed identical over all episodes* $t \in \mathcal{E}_e$ *of a given epoch* e*.*

Summary. The strategy described above is summarized in an algorithm box in Appendix A.

3.2 Main result

We may now state our main result and discuss its relation to previously known bounds.

Theorem 1 (Main theorem)**.** *Algorithm* `APO-MVP`*, used, for instance, with the OLO strategies based on polynomial or exponential potential (see Examples 1 and 2), satisfies, with probability at least* $1-3\delta$*,*

$$\begin{aligned} R_T \leqslant\ & \sqrt{H^7SAT\log_2(2T)}\,\bigl(2\log_2(2T) + 8\sqrt{\log(A)}\bigr) \\ & + 3\sqrt{H^4SAT\log\bigl(2SATH\log_2(2T)/\delta\bigr)} + 2\sqrt{2H^5T\log_2(2T)\ln(2/\delta)} + 2H^3SA\,. \end{aligned}$$

Proof. The result follows the decomposition stated in the introduction of Section 4 together with Lemmas 1–5–6–7 located therein. □

Theorem 1 shows that the regret is $\widetilde{\mathcal{O}}(\sqrt{H^7SAT})$, matching the minimax lower bound $\Omega(\sqrt{H^3SAT})$ for stochastic MDPs in terms of dependencies on S, A, and T, up to logarithmic factors (Jin et al., 2018; Domingues et al., 2021). To the best of our knowledge, it is the first result that achieves the minimax optimal dependency on the number of states S in the adversarial setting.

Comparison to Rosenberg and Mansour (2019a). Algorithm `UC-O-REPS` by Rosenberg and Mansour (2019a) achieves $\widetilde{\mathcal{O}}(\sqrt{H^4S^2AT})$ regret bound in our setting and

with our notation (taking $L = H$ and $|\mathcal{X}| = HS$ since a state-space layer $\mathcal{X}$ may be represented as H independent copies of $\mathcal{S}$). In particular, our result improves upon the previous best known bound in the regime of large state spaces $S \geqslant H^3$. We suspect that—perhaps through successful incorporation of clipping, see Remark 1—the regret bound could be improved to $\widetilde{\mathcal{O}}(\sqrt{H^5SAT})$. We plan to investigate this in future works, and it remains an open problem to fully match the minimax lower bound $\Omega(\sqrt{H^3SAT})$.

Finally, our analysis in Section 4.3 relies significantly on the fact that the adversary is oblivious, while `UC-O-REPS` can handle fully adversarial setups. However, due to the exploration mechanism used, this algorithm is not able to take advantage of the oblivious adversary and would still pay the same $\sqrt{S}$ factor.

Comparison to Cai et al. (2020). Our algorithm shares similarities with the online proximal policy optimization [`OPPO`] approach of Cai et al. (2020) and Shani et al. (2020), which also uses dynamic programming and policy optimization through online mirror descent. However, our approach incorporates the doubling trick to stabilize value updates, enabling us to: i) employ any online linear optimization strategy in a black-box manner without unnecessary adaptations; and ii) improve the dependency on S by a multiplicative factor $\sqrt{S}$, by leveraging the analysis of Zhang et al. (2023).

4 Proof sketch for Theorem 1

We decompose the regret into four terms that are treated separately. Denoting by $\boldsymbol{\pi}^\star$ as the policy that achieves the maximum in (1), we decompose the regret, following ideas of Auer et al. (2008) and Azar et al. (2017), as:

$$
\begin{aligned}
R_T = & \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) \right) && \overset{\text{def}}{=} (\mathbf{A}) \\
& + \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) \right) && \overset{\text{def}}{=} (\mathbf{B}) \\
& + \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \boldsymbol{P}}(s_1) \right) && \overset{\text{def}}{=} (\mathbf{C}) \\
& + \sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t, \boldsymbol{b}_t, \boldsymbol{P}}(s_1) \,, && \overset{\text{def}}{=} (\mathbf{D})
\end{aligned}
$$

where we used the linearity of the value functions: $V_1^{\boldsymbol{\pi}, \boldsymbol{g}+\boldsymbol{g}', \boldsymbol{Q}} \equiv V_1^{\boldsymbol{\pi}, \boldsymbol{g}, \boldsymbol{Q}} + V_1^{\boldsymbol{\pi}, \boldsymbol{g}', \boldsymbol{Q}}$. We provide a high-level overview of the techniques used for each term.

For $(\mathbf{A})$ we leverage the careful choice of the bonuses $\boldsymbol{b}_t$ to show that on a properly chosen high-probability event, $(\mathbf{A})$ is non-positive.

For $(\mathbf{B})$, we crucially use that within each epoch the considered transition kernels are constant (see Fact 2), so that we may resort to the adversarial-learning technique by Jonckheere et al. (2023), which consists of running SH independent OLO strategies.

Term $(\mathbf{C})$ is the most involved part of the analysis from the probabilistic standpoint: we resort to the machinery developed by Zhang et al. (2023), which relies greatly on a doubling trick that we mimicked in the definition of the `APO-MVP` algorithm.

Finally, $(\mathbf{D})$ is the least involved term, it can be controlled by some lines of elementary calculations.

In what follows, each section provides additional details of the analysis per term, in the order: $(\mathbf{B})$ – additional technical concepts – $(\mathbf{A})$ – $(\mathbf{D})$ – $(\mathbf{C})$.

4.1 Term $(\mathbf{B})$: OLO analysis

The goal of this section is to prove the following result, which also holds for other OLO strategies satisfying the performance guarantee of Definition 1.

Lemma 1. *Among others, the OLO strategies based on polynomial or exponential potentials (see Examples 1 and 2) satisfy*

$$(\mathbf{B}) \leqslant 8\sqrt{H^7 SAT \log_2(2T) \log(A)}\,.$$

Before proving this result, let us briefly recall what online linear optimization [OLO] consists of; see the monograph by Cesa-Bianchi and Lugosi (2006) for a more detailed exposition. We take some generic notation for now but will later connect OLO to constructions of policies; in particular, we consider for now reward vectors of length $K \geqslant 2$, but will later replace $[K]$ by the action space $\mathcal{A}$.

Online linear optimization. At each round $t \geqslant 1$ and based on the past, a learning strategy $\varphi = (\varphi_t)_{t\geqslant 1}$ picks a convex combination $\boldsymbol{w}_t = (w_{t,1}, \dots, w_{t,K}) \in \Delta\big([K]\big)$ while an opponent player picks, possibly at random, a vector $\boldsymbol{g}_t = (g_{t,1}, \dots, g_{t,K})$ of signed rewards. Both $\boldsymbol{w}_t$ and $\boldsymbol{g}_t$ are revealed at the end of the round. By "based on the past", we mean, for the learning strategy, that $\boldsymbol{w}_t = \varphi_t\big((\boldsymbol{g}_\tau)_{\tau\leqslant t-1}\big)$. The initial vector $\boldsymbol{w}_1$ is constant.

Definition 1. *A learning strategy φ controls the regret in the adversarial setting with rewards bounded by $M > 0$ if there exists a sequence $(B_{T,K})_{T\geqslant 1}$ with $B_{T,K}/T \to 0$ such that, against all opponent players sequentially picking reward vectors $\boldsymbol{g}_t$ in $[-M, M]^K$, for all $T \geqslant 1$,*

$$\max_{k\in[K]} \sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T} \sum_{j\in[K]} w_{t,j}\, g_{t,j} \leqslant M\, B_{T,K}\,.$$

The optimal orders of magnitude of $B_{T,K}$ are $\sqrt{T\ln K}$. In Definition 1, the strategy may know M and rely on its value. Also, the strategy should work for any optimization horizon T (see the final "for all $T \geqslant 1$" in the definition above): this is because the lengths E_e of the global epochs $\mathcal{E}_e$ are not known in advance. There exist several strategies meeting the requirements of Definition 1; we provide two examples below.

Example 1. *The potential-based strategies by Cesa-Bianchi and Lugosi (2003) are defined based on a non-decreasing function $\Phi\colon \mathbb{R} \to [0, +\infty)$. They resort to $w_{1,k} = 1/K$ and for $t \geqslant 2$,*

$$w_{t,k} = \frac{v_{t,k}}{\sum_{j\in[K]} v_{t,j}}\,, \qquad \text{where} \qquad v_{t,k} = \Phi\left(\sum_{\tau=1}^{t-1} g_{\tau,k} - \sum_{\tau=1}^{t-1}\sum_{j\in[K]} w_{\tau,j} g_{\tau,j}\right). \tag{5}$$

For the polynomial potential $\Phi\colon x \mapsto \big(\max\{x, 0\}\big)^{2\ln K}$, Cesa-Bianchi and Lugosi (2003, Section 2) show that the strategy satisfies the performance guarantee of Definition 1 with $B_{T,K} = \sqrt{6T\ln K}$.

Example 2. *Auer et al. (2002) studied the use of exponential potential with time-varying learning rates $\eta_t = (1/M)\sqrt{(\ln K)/t}$, i.e., using $\Phi_t(x) = \exp(\eta_t x)$ in (5) to define the weights at round t. This strategy satisfies the performance guarantee of Definition 1 with $B_{T,K} = \sqrt{T\ln K}$.*

There exist adaptive versions of the two previous strategies: `ML-Poly` in Gaillard et al. (2014), `AdaHedge` in Erven et al. (2011), de Rooij et al. (2014), Orabona and Pál (2015).

Appendix A states closed-form expressions of the policies (2) constructed with the strategies of Examples 1 and 2, as well as `AdaHedge`.

Connection between OLO and the construction of policies. Jonckheere et al. (2023) prove the following. Let $r'_{t,h}\colon \mathcal{S}\times\mathcal{A} \to [0, M]$ be a sequence of reward functions. Define a sequence of policies $(\boldsymbol{\pi}'_t)_{t\geqslant 1}$ as: for each $t \geqslant 1$, for each $s \in \mathcal{S}$, for each $h \in [H]$,

$$\pi'_{t,h}(\,\cdot\,|s) = \varphi_t\left(\left(A_h^{\boldsymbol{\pi}'_\tau, \boldsymbol{r}'_\tau, \boldsymbol{P}'}(s,\,\cdot\,)\right)_{\tau\leqslant t-1}\right), \quad \text{where} \quad A_h^{\boldsymbol{\pi}'_\tau, \boldsymbol{r}'_\tau, \boldsymbol{P}'}(s,\,\cdot\,) = \left(A_h^{\boldsymbol{\pi}'_\tau, \boldsymbol{r}'_\tau, \boldsymbol{P}'}(s,a)\right)_{a\in\mathcal{A}}.$$

Lemma 2 (Jonckheere et al., 2023)**.** *If the learning strategy satisfies the conditions of Definition 1, then the sequence of policies defined right above is such that, for all fixed policies $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$, for all $T \geqslant 1$,*

$$\sum_{t=1}^{T}\left(V_1^{\boldsymbol{\pi}, \boldsymbol{r}'_t, \boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t, \boldsymbol{r}'_t, \boldsymbol{P}'}(s_1)\right) \leqslant MH^2\, B_{T,A}\,.$$

For the sake of completeness, the proof of Lemma 2 is provided in Appendix B. We are now ready to prove Lemma 1.

Proof of Lemma 1. We apply Lemma 2 in each global epoch $\mathcal{E}_e$, with $\boldsymbol{P}' = \widehat{\boldsymbol{P}}_t$ (see Fact 2) and $\boldsymbol{r}'_t = \boldsymbol{r}_t + \boldsymbol{b}_t$ for all $t \in \mathcal{E}_e$. Since $r_{t,h} \in [0,1]$ and $b_{t,h} \in [0,H]$, we can pick $M = 1 + H \leqslant 2H$. Decomposing term $(\mathbf{B})$ into a summation over the global epochs and using the bound of Lemma 2 for each of them, we deduce that, for both strategies of Examples 1 and 2,

$$(\mathbf{B}) \leqslant 2H^3 \sum_{e=1}^{m(T)} B_{E_e,A} \leqslant 2H^3 \sum_{e=1}^{m(T)} \sqrt{6E_e \log(A)} \leqslant 8H^3 \sqrt{Tm(T)\log(A)}\,, \tag{6}$$

where we applied Jensen's inequality to the root. Lemma 3 below then yields the claimed result. □

4.2 Additional technical concepts

To deal with the remaining terms $(\mathbf{A})$ – $(\mathbf{D})$ – $(\mathbf{C})$, we will not need anymore to pay attention to global epochs $\mathcal{E}_{e_t}$, only local epochs $\ell_{t,h}(s,a)$ will be of interest.

We review two concepts which have been successfully used by Zhang et al. (2023) to derive minimax optimal regret bounds in the case of stochastic MDPs.

The first concept: epoch-switching conditions and profiles. The functions indicating local epochs $\ell_{t,h}(s,a)$ were called a profile by Zhang et al. (2023); they take bounded values:

$$\ell_{t,h} \colon \mathcal{S} \times \mathcal{A} \to \big[\lceil \log_2(T) \rceil\big]^*, \qquad \text{and let} \qquad \boldsymbol{\ell}_t = (\ell_{t,h})_{h \in [H-1]} \tag{7}$$

with the agreement that $\ell_{0,h}(\,\cdot\,,\,\cdot\,) \equiv 0$ for all $h \in [H-1]$. We also introduce $\boldsymbol{\ell}_{<t} = (\boldsymbol{\ell}_\tau)_{0 \leqslant \tau \leqslant t-1}$. Using the above-defined profiles, we note that the global epoch e_t of a given episode $t \in [T]$ may be obtained as a function of $\boldsymbol{\ell}_{<t}$, namely,

$$e_t = \Psi\big(\boldsymbol{\ell}_{<t}\big) \stackrel{\text{def}}{=} \sum_{\tau=1}^{t-1} \min\left\{ \sum_{(s,a,h)} \big(\ell_{\tau,h}(s,a) - \ell_{\tau-1,h}(s,a)\big),\ 1 \right\}. \tag{8}$$

Indeed, if the counter of no triplet (s,a,h) has reached a value of the form 2^r for some integer r by passing from episode $\tau - 1$ to τ, then the summation in the minimum is zero, meaning that the episodes τ and $\tau + 1$ belong to the same global epoch. On the contrary, if the counter of at least one (s,a,h) reached such a value, then this sum is at least 1 (there can be more than one triplets satisfying this), meaning that τ and $\tau + 1$ belong to different global epochs. Thus, thanks to the minimum, the above quantity counts the number of (global) epoch switches from $\tau = 1$ to $\tau = t$. In other words, the global epoch e_t is uniquely determined by the preceding profiles.

Since there are $SA(H-1)$ different triplets (s,a,h) and each such triplet is associated with at most $\lceil \log_2(T) \rceil$ doubling conditions, we obtain the following bound.

Lemma 3. *There are at most $m(T) \leqslant SAH \log_2(2T)$ global epochs.*

The second concept: optional skipping for estimated transition kernels. The trick detailed here is standard in the bandit and reinforcement-learning literature. The original reference is Theorem 5.2 of Doob (1953, Chapter III, p. 145); one can also check (Chow and Teicher, 1988, Section 5.3) for a more recent reference. A pedagogical exposition of the trick and of its uses in the bandit literature may be found in (Garivier et al., 2022, Section 4.1), which we adapt to the setting of reinforcement learning.

For each triplet $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ and each integer $j \geqslant 1$, we denote by

$$N_{h,s,a,j} \stackrel{\text{def}}{=} \inf\big\{ t \geqslant 1 : n_{t,h}(s,a) = j \big\}$$

(with the convention that the infimum of an empty set equals $+\infty$) the predictable stopping time whether and when (s,a) occurs for the j–th time. We are interested in the distribution of the states $s_{t,h+1}$ drawn at rounds t when $(s_{t,h}, a_{t,h}) = (s,a)$; these rounds are given by the stopping times $N_{h,s,a,j}$ introduced above. It turns out that these states are i.i.d. with distribution $P_h(\,\cdot \mid s,a)$. We also have independence across sequences of states. All these results are formally stated in the following lemma: to do so, one needs to set the values of the number of times $n_{t,h}(s,a)$ each triplet (h,s,a) was encountered till a given round.

Lemma 4 (Doob's optional skipping)**.** *Fix $t \geqslant 1$ and consider sequences of integers $J_{h,s,a} \geqslant 1$ and the intersection of events*

$$\mathcal{C} = \bigcap_{h\in[H-1]} \bigcap_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big\{n_{t,h}(s,a) = J_{h,s,a}\big\}\,.$$

It holds that

$$\text{on } \mathcal{C}, \qquad \text{each of the sequences} \quad (\widetilde{s}_{h,s,a,j})_{j\in[J_{h,s,a}]} \overset{\text{def}}{=} \big(s_{N_{h,s,a,j},h+1}\big)_{j\in[J_{h,s,a}]}$$

is formed by i.i.d. variables, with common distribution $P_h(\,\cdot \mid s,a)$. In addition, these sequences are independent from each other as (h,s,a) vary in $[H-1]\times\mathcal{S}\times\mathcal{A}$.

One of our applications of Lemma 4 will be the following, to handle term $(\mathbf{A})$. The proof consists of noting first that on $\big\{\ell_{t-1,h}(s,a) = \ell\big\}$, the distribution $\widehat{P}_{t,h}(\,\cdot \mid s,a)$ corresponds to the empirical measure of the i.i.d. variables $\widetilde{s}_{h,s,a,j}$ with $1 \leqslant j \leqslant 2^{\ell-1}$, and second, by dropping the indicator function.

Notation-wise, we will be using $\widetilde{s}_{h,s,a,j}$ (as in Lemma 4) for random variables generated by the MDP interactions and $\sigma_{h,s,a,j}$ (as in Corollary 1) for random variables independent from everything else and that are representations of the former.

Corollary 1. *Fix $h \in [H-1]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$ and let $(\sigma_{h,s,a,j})_{j\geqslant 1}$ be a sequence of i.i.d. variables with distribution $P_h(\,\cdot \mid s,a)$. For all functions $\psi\colon \mathbb{R} \to [0,+\infty)$, all functions $g\colon \mathcal{S}\to\mathbb{R}$, and all integers $\ell \geqslant 1$,*

$$\mathbb{E}\Big[\psi\Big(\widehat{P}_{t,h}\cdot g(s,a)\Big)\,\mathbb{I}\big\{\ell_{t-1,h}(s,a)=\ell\big\}\Big] \leqslant \mathbb{E}\left[\psi\left(\frac{1}{2^{\ell-1}}\sum_{j=1}^{2^{\ell-1}} g(\sigma_{h,s,a,j})\right)\right].$$

4.3 Term $(\mathbf{A})$: Optimism

Term $(\mathbf{A})$ is handled thanks to a result already present in the analysis of the `UCBVI` algorithm (Azar et al., 2017, Lemma 18), relying on an induction, and thanks to applications of Hoeffding's inequalities together with optional skipping. Appendix C provides the (straightforward) details of the proof of the following lemma.

Lemma 5. *With probability at least $1-\delta$, for all $t\in[T]$ and all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$,*

$$Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \leqslant Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a) \qquad \text{and} \qquad V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s) \leqslant V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s)\,.$$

In particular, with probability at least $1-\delta$, we have $(\mathbf{A}) \leqslant 0$.

4.4 Term $(\mathbf{D})$: Bonus summation

Without the doubling trick, the exploration bonuses summed up along the trajectory can be classically bounded by a $O(\sqrt{T})$ term. The doubling trick introduces only minor changes to this classical step. Appendix D provides the (straightforward) details of the proof of the following lemma, based on the Hoeffding–Azuma inequality together with simple controls of the form, for all $h\in[H-1]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_{t,h}(s,a)}}\,\mathbb{I}\big\{(s_{t,h},,a_{t,h})=(s,a)\big\}\,\mathbb{I}\big\{n_{t,h}(s,a)\geqslant 2\big\} = \sum_{n=2}^{n_{T,h}(s,a)} \frac{1}{\sqrt{n}} \leqslant 2\sqrt{n_{T,h}(s,a)}\,.$$

Lemma 6. *With probability at least $1-\delta$, we have*

$$(\mathbf{D}) \leqslant 3\sqrt{H^4SAT\log(2SATH\log_2(2T)/\delta)} + H^2SA\,.$$

4.5 Term $(\mathbf{C})$: Concentration

Let us start by formally stating the result, whose detailed proof may be found in Appendix E; below, we only sketch that proof. The analysis is essentially borrowed from Zhang et al. (2023) with minor

technical modifications but a much simplified exposition (as we do not target optimized bounds yet). Also, we explain in Remark 2 of Appendix E that the dependency in H of the leading term in the upper bound of Lemma 7 could be improved to $\sqrt{H^5}$ with some more efforts, but that there is no point in doing so, given the bound of Lemma 1, which also scales with H as $\sqrt{H^7}$.

Lemma 7. *With probability at least $1-\delta$, it holds that*

$$(\mathbf{C}) \leqslant 2\sqrt{H^7 S A T \big(\log_2(2T)\big)^3} + 2\sqrt{2H^5\, T \log_2(2T) \ln(2/\delta)} + SAH^3\,.$$

Proof sketch. An application of the performance-difference lemma in case of different transition kernels (see, e.g., Russo, 2019, Lemma 3) together with the Hoeffding–Azuma inequality first shows that with probability at least $1-\delta/2$,

$$\begin{aligned}(\mathbf{C}) &= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=1}^{H-1} \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_{t,h}, a_{t,h}) \,\middle|\, \boldsymbol{\pi}_t,\, \boldsymbol{b}_t,\, \widehat{\boldsymbol{P}}_t\right] \\ &\leqslant \sum_{h=1}^{H-1} \underbrace{\sum_{t=1}^{T} \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_{t,h}, a_{t,h})}_{= \xi_{T,h}} + \sqrt{2H^5 T \ln(2/\delta)}\,.\end{aligned}$$

We bound the quantities $\xi_{T,h}$ for each fixed $h \in [H-1]$. We apply optional skipping in a careful way on the event $\mathcal{C}_{\ell,j,h,s,a,t}$ when $(s,a) \in \mathcal{S} \times \mathcal{A}$ is played for the $(2^{\ell-1}+j)$–th time in stage h at episode t:

$$\begin{aligned}&\text{on } \mathcal{C}_{\ell,j,h,s,a,t}, && \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_{t,h}, a_{t,h}) \\ &&\text{behaves like} \quad & \frac{1}{2^{\ell-1}} \sum_{j \in [2^{\ell-1}]} \big(\widetilde{V}_{s,a,h+1}(\sigma_{h,s,a,j}) - P_h \widetilde{V}_{s,a,h+1}(s,a)\big)\,,\end{aligned}$$

for some random variable $\widetilde{V}_{s,a,h+1}$, where the $\sigma_{h,s,a,j}$ are i.i.d. according to $P_h(\cdot \mid s,a)$ and are independent from $\widetilde{V}_{s,a,h+1}$.

The argument also extends between pairs (s,a) so that a careful application of the Hoeffding-Azuma inequality (this is the delicate part of the proof), together with the consideration of all values for ℓ and j, then shows that, with probability at least $1-\delta/2$,

$$\xi_{T,h} \leqslant \sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \sqrt{2H^4 \frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{I}\big\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\big\} \ln \frac{1}{\delta'}}\,,$$

where δ' equals $\delta/2$ divided by the number of times we applied the union bound over ℓ, j, H, and in the course of optional skipping; we bound this number of times by $4H(T+1)^{1+SAH\lceil \log_2(T) \rceil}$.

We conclude the proof by two consecutive applications of Jensen's inequality for the root:

$$\begin{aligned}&\sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \sqrt{\frac{2H^4}{2^{\ell-1}} \sum_{(s,a)} \mathbb{I}\big\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\big\} \ln \frac{1}{\delta'}} \\ &\qquad \leqslant \sqrt{2H^4 \lceil \log_2 T \rceil \underbrace{\sum_{(s,a)} \sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \mathbb{I}\big\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\big\}}_{\leqslant \sum_{(s,a)} n_{T,h}(s,a) = T} \ln \frac{1}{\delta'}}\,,\end{aligned}$$

together with some algebra. ∎

5 Conclusion and limitations

In this work, we proposed an algorithm called APO-MVP that extends algorithm MVP of Zhang et al. (2023) and its analysis to the case of adversarial reward functions, thanks, in particular, to a black-box adversarial aggregation mechanism due to Jonckheere et al. (2023) that takes care of the adversarial nature of reward functions. Algorithm APO-MVP is easy to implement in practice as it relies on OLO learning strategies in the policy space combined with dynamic programming; it does not at all rely on so-called occupancy measures. Furthermore, it achieves a better regret bound compared to previous approaches based on the occupancy measures, reducing their regret bounds by a $\sqrt{S}$ multiplicative factor and narrowing the gap between the adversarial and stochastic regret bounds, which are both shown to be of order $\sqrt{SAT}$ up to logarithmic factors, as far as dependencies on S, A, and T are concerned.

We believe that this work opens many interesting follow-up questions. The two main open questions are inevitably linked with the main limitations of this paper and are discussed below.

Limitations. The main limitation is rather a high dependency of our regret bound on the length H of the episodes, of order $\sqrt{H^7}$. Improving this dependency while maintaining a regret of order $\sqrt{SAT}$ up to logarithmic terms is one of the remaining open questions. Also, as it is common in the literature, we have only considered full monitoring cases so far; extending our approach to bandit monitoring seems to be non-trivial. It is still unknown if a $\sqrt{SAT}$ order of magnitude for the regret is possible in the adversarial case with bandit feedback.

References

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR, 2017.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51:239–261, 2003.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Y. Chow and H. Teicher. *Probability Theory*. Springer, 1988.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

J.L. Doob. *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons, 1953.

Tim Erven, Wouter M Koolen, Steven Rooij, and Peter Grünwald. Adaptive hedge. *Advances in Neural Information Processing Systems*, 24, 2011.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory (COLT'2014)*, volume PMLR:35, pages 176–196, 2014.

A. Garivier, H. Hadiji, P. Ménard, and G. Stoltz. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *Journal of Machine Learning Research*, 23(179):1–66, 2022.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 4259–4280. PMLR, 2022.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34: 20491–20502, 2021.

Matthieu Jonckheere, Chiara Mignacco, and Gilles Stoltz. Symphony of experts: orchestration with adversarial insights in reinforcement learning, 2023. Preprint, arXiv:2310.16473.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7281–7289, 2022.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and Dávid Pál. Scale-free algorithms for online linear optimization. In Kamalika Chaudhuri, Claudio Gentile, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 287–301, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24486-0.

A. Rosenberg and Y. Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, volume PMLR:97, 2019a.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32, 2019b.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.

Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021.

Zihan Zhang, Yuxin Chen, Jason D. Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning, 2023.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

A Detailed algorithm description

In this section, we first provide a an algorithmic description of the strategy introduced in Section 3.1, and we then write closed-form expressions of the policy constructions 2 based on the strategies of Examples 1 and 2, as well as AdaHedge.

Algorithm 1: Adversarial Policy Optimization based on Monotonic Value Propagation (`APO-MVP`)

Data: Number of rounds T, number of states, actions and horizon S, A, H, confidence level δ, online linear optimization strategy $\varphi = (\varphi_t)_{t \geqslant 1}$
Result: Sequence of policies $\boldsymbol{\pi}^t = (\pi_{t,h})_{h \in [H]}$, for $t \in [T]$

Initialize kernels $\widehat{P}_h(s'|s,a) = 1/S$ for all $(s', s, a, h) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times [H-1]$;
Initialize counters $n_h(s,a,s') = n_h(s,a) = 0$ for all $(s', s, a, h) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times [H-1]$;
Initialize histories $\mathcal{H}_{h,s} = \emptyset$ for all $h \in [H]$ and $s \in \mathcal{S}$;
Initialize $\pi_{1,h}(\cdot \mid s) = \varphi_1(\emptyset)$ for all $h \in [H]$ and $s \in \mathcal{S}$;
Select initial state $s_1 \in \mathcal{S}$;
for *rounds* $t = 1, \ldots, T$ **do**
  `/* Interaction */`
  Set $s_{t,1} = s_1$;
  for $h = 1, \ldots, H$ **do**
    Play action $a_{t,h} \sim \pi_{t,h}(\cdot \mid s_{t,h})$;
    Receive next state $s_{t,h+1} \sim P_h(\cdot \mid s_{t,h}, a_{t,h})$;
    Update counters $n_h(s_{t,h}, a_{t,h}) \mathrel{+}= 1$ and $n_h(s_{t,h}, a_{t,h}, s_{t,h+1}) \mathrel{+}= 1$;
    `/* Trigger, update the model */`
    if $n_h(s_{t,h}, a_{t,h}) = 2^{\ell-1}$ *for some* $\ell \geqslant 1$ **then**
      $\widehat{P}_h(s'|s_{t,h}, a_{t,h}) = \dfrac{n_h(s_{t,h}, a_{t,h}, s')}{2^{\ell-1}}$ for all $s' \in \mathcal{S}$;
      $b_h(s_{t,h}, a_{t,h}) = \sqrt{\dfrac{2H^2 \log(2SATH \log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H$;
      Activate `trigger`;
    end
  end
  Receive a reward function $\boldsymbol{r}_t = (r_{t,h})_{h \in [H]}$;
  if *`trigger`* **then**
    Drop all histories, i.e., set $\mathcal{H}_{h,s} = \emptyset$ for all $h \in [H]$ and $s \in \mathcal{S}$;
    Set $\pi_{t+1,h}(\cdot \mid s) = \varphi_1(\emptyset)$ for all $h \in [H]$ and $s \in \mathcal{S}$;
    Deactivate `trigger`;
  else
    `/* Compute first the advantage functions via Bellman's equations */`
    Let $Q_H(s,a) = r_{t,H}(s,a)$ and $V_h(s) = \pi_{t,H} \cdot Q_H(s)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$;
    for $h = H-1, H-2, \ldots, 1$ **do**
      Let $Q_h(s,a) = r_{t,h}(s,a) + b_h(s,a) + \widehat{P}_h \cdot V_{h+1}(s,a)$ and $V_h(s) = \pi_{t,h} \cdot Q_h(s)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$;
    end
    Let $A_h(s,a) = Q_h(s,a) - V_h(s)$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    Add $\big(A_h(s,a)\big)_{a \in \mathcal{A}}$ to the history $\mathcal{H}_{h,s}$ for each $(h,s) \in [H] \times \mathcal{S}$;
    `/* Next, obtain` $\boldsymbol{\pi}_{t+1}$ `via the learning strategy */`
    Let $\pi_{t+1,h}(\cdot \mid s) = \varphi(\mathcal{H}_{h,s})$ for all $(h,s) \in [H] \times \mathcal{S}$;
  end
end

Closed-form expressions of the policy constructions. We first recall the statement (2) for the construction of policies: for all $t \geqslant 1$, all $h \in [H]$, and $s \in \mathcal{S}$,

$$\pi_{t,h}(\cdot \mid s) = \varphi_t\Big(\big(\widehat{A}_{\tau,h}(s, \cdot)\big)_{\tau \in \mathcal{E}_{e_t} \cap [t-1]}\Big).$$

We now illustrate this definition with the strategies of Examples 1 and 2, as well as with `AdaHedge`. A key observation to do so will be that, by definition of advantage functions and since $\widehat{A}_{\tau,h}(s,\,\cdot\,)$ is based on the policy $\pi_{\tau,h}$,

$$\forall \tau \in [T],\ \ \forall s \in \mathcal{S}, \qquad \sum_{a \in \mathcal{A}} \pi_{\tau,h}(a \mid s)\, \widehat{A}_{\tau,h}(s,a) = 0\,.$$

Polynomial potential (Example 1). We denote by $(x)_+ = \max\{x,\,0\}$ the non-negative part of $x \in \mathbb{R}$. We have $\varphi_1 \equiv (1/A,\,\ldots,\,1/A)$ and for $t \geqslant 2$, whenever $\mathcal{E}_{e_t} \cap [t-1]$ contains at least one element,

$$\begin{aligned}
\pi_{t,h}(a \mid s) &= \frac{\left(\displaystyle\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a) - \sum_{a'' \in \mathcal{A}} \pi_{\tau,h}(a'' \mid s)\, \widehat{A}_{\tau,h}(s,a'')\right)_+^{2\ln A}}{\displaystyle\sum_{a' \in \mathcal{A}} \left(\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a') - \sum_{a'' \in \mathcal{A}} \pi_{\tau,h}(a'' \mid s)\, \widehat{A}_{\tau,h}(s,a'')\right)_+^{2\ln A}} \\
&= \frac{\left(\displaystyle\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a)\right)_+^{2\ln A}}{\displaystyle\sum_{a' \in \mathcal{A}} \left(\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a')\right)_+^{2\ln A}}\,.
\end{aligned}$$

Exponential potential (Example 2). Similarly to above, we have $\varphi_1 \equiv (1/A,\,\ldots,\,1/A)$ and for $t \geqslant 2$, whenever $\mathcal{E}_{e_t} \cap [t-1]$ contains at least one element,

$$\pi_{t,h}(a \mid s) = \frac{\exp\left(\eta_t \displaystyle\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a)\right)}{\displaystyle\sum_{a' \in \mathcal{A}} \exp\left(\eta_t \sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a')\right)} \qquad \text{where} \qquad \eta_t = \frac{1}{H+1}\sqrt{\frac{\ln A}{\left|\mathcal{E}_{e_t} \cap [t-1]\right|}}$$

are time-varying learning rates, based on the cardinality $\left|\mathcal{E}_{e_t} \cap [t-1]\right|$ of $\mathcal{E}_{e_t} \cap [t-1]$.

Adaptive versions of exponential-potential-based strategies. The literature proposed many ways of setting the learning rates for exponential potentials based on past information—a series of work initiated by Auer et al. (2002), whose learning rates were used in the paragraph above. One may cite, among (many) others, Cesa-Bianchi et al. (2007), Erven et al. (2011), de Rooij et al. (2014), Orabona and Pál (2015); sometimes, the resulting strategy is called `AdaHedge`. For instance, Orabona (2019, Section 7.6) summarizes this literature by the following learning rates:

$$\eta_t = \frac{\max\{4,\, 2^{-1/4}\sqrt{\log A}\}}{\sqrt{\displaystyle\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \max_{a' \in \mathcal{A}} \big(\widehat{A}_{\tau,h}(s,a')\big)^2}}\,.$$

These updates correspond to an OLO strategy satisfying the bound of Definition 1 with a performance bound $B_{T,K} = 4\sqrt{T \log K}$.

B Term (B)

It only remains to prove Lemma 2, which we restate below. For the sake of completeness, we copy the proof by Jonckheere et al. (2023).

Lemma 2 (Jonckheere et al., 2023)**.** *If the learning strategy satisfies the conditions of Definition 1, then the sequence of policies defined right above is such that, for all fixed policies $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$, for all $T \geqslant 1$,*

$$\sum_{t=1}^{T}\Big(V_1^{\boldsymbol{\pi},\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1)\Big) \leqslant MH^2\, B_{T,A}\,.$$

Proof. As the reward function takes values in $[0, M]$, we have that $\big|A_{\tau,h}^{\boldsymbol{\pi}'_\tau}(s,a)\big| \leqslant M(H-h+1)$. By the definition of advantage functions (for the equality to 0) and by Definition 1 (for the upper bound), we have, for all $s \in \mathcal{S}$,

$$\max_{a\in\mathcal{A}} \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a) - \sum_{t=1}^{T} \underbrace{\sum_{a\in\mathcal{A}} \pi'_{t,h}(a|s)\, A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)}_{=0} \leqslant M(H-h+1)\, B_{T,A}\,. \tag{9}$$

Now, the so-called performance difference lemma (see, e.g., Kakade and Langford (2002) for the result in the discounted setting) shows that

$$V_1^{\boldsymbol{\pi},\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s) \sum_{a\in\mathcal{A}} \pi_h(a|s)\, A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)$$

where $\mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}$ is the distribution of $s_{t,h}$ induced in the h–th episode by $\boldsymbol{\pi}$ given the state transitions $\boldsymbol{P}'$ and the initial state s_1. Summing this equality over t and rearranging, we get

$$\begin{aligned}
\sum_{t=1}^{T}\Big(V_1^{\boldsymbol{\pi},\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1)\Big) &= \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s) \sum_{a\in\mathcal{A}} \pi_h(a|s) \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a) \\
&\leqslant \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s) \underbrace{\max_{a\in\mathcal{A}} \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)}_{\leqslant M(H-h+1)\, B_{T,A}} \\
&\leqslant MH^2\, B_{T,A}\,,
\end{aligned}$$

where we substituted (9). Here, we crucially used that the convex combination with weights $\mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s)$ is independent of t and only depends on the fixed benchmark policy $\boldsymbol{\pi}$, on the state transitions $\boldsymbol{P}'$, and on the initial states s_1 (identical for all t). □

C Term $(\mathbf{A})$

We start with the following consequence of Hoeffding's inequality.

Lemma 8. *For each $t \in [T]$, for each $h \in [H-1]$, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, for each $\ell \in \big[\lceil \log_2(T) \rceil\big]$,*

$$\mathbb{P}\left\{ \left| \frac{1}{2^{\ell-1}} \sum_{j=1}^{2^{\ell-1}} V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(\sigma_{h,s,a,j}) - P_h \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| \right.$$
$$\left. > \sqrt{\frac{2H^2 \log(2SATH\log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H \right\} \leqslant \frac{\delta}{SATH\log_2(2T)}\,,$$

where $(\sigma_{h,s,a,j})_{1 \leqslant j \leqslant 2^{\ell-1}}$ is a sequence of i.i.d. variables with distribution $P_h(\,\cdot \mid s,a)$.

Proof. The policy $\boldsymbol{\pi}^\star$ is fixed, as it only depends on the $\boldsymbol{r}_t$ and $\boldsymbol{P}$, which are all fixed beforehand. The function $g = V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}$ is therefore a fixed deterministic function. The expectation of $g(\sigma_{h,s,a,j})$ is indeed, given our notation, $P_h g$. By the boundedness of rewards in $[0,1]$, and thus the boundedness of values in the range $[0,H]$, we may therefore apply Hoeffding's inequality: we do so for each $t \in [T]$, each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H-1]$, and each $\ell \in \big[\lceil \log_2(T) \rceil\big]$, and get that for all $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\left| \frac{1}{2^{\ell-1}} \sum_{j=1}^{2^{\ell-1}} V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(\sigma_{h,s,a,j}) - P_h \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| \leqslant \sqrt{\frac{2H^2 \log(2/\delta')}{2^{\ell-1}}}\,.$$

The proof is concluded by keeping in mind that the left-hand side necessarily belongs to $[0,H]$ by boundedness of values in $[0,H]$. □

We are now ready to prove Lemma 5, which we restate below; we do so by mimicking the proof of Azar et al. (2017, Lemma 18).

Lemma 5. *With probability at least $1-\delta$, for all $t \in [T]$ and all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,*

$$Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \leqslant Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) \qquad \text{and} \qquad V_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s) \leqslant V_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s)\,.$$

In particular, with probability at least $1-\delta$, we have $(\mathbf{A}) \leqslant 0$.

Proof. We proceed by backward induction, for each given $t \in [T]$; more precisely, we consider, for $h \in [H]$, the induction hypothesis

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \leqslant Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) \qquad (\mathcal{H}_h)$$
$$\text{and} \qquad V_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s) \leqslant V_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s)\,.$$

For $h = H$, we note that for all (s,a),

$$Q_H^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) = r_{H,t}(s,a) = r_{t,H}(s,a) + b_{t,H}(s,a) = Q_H^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a)\,,$$

so that $(\mathcal{H}_H)$ is trivially satisfied. For $h \in [H-1]$, by Bellman equations,

$$Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) - Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) = b_{t,h}(s,a) + \widehat{P}_{t,h} \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) - P_h \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a)\,,$$

where by the induction hypothesis $(\mathcal{H}_{h+1})$, we have $V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}^t}(s') \geqslant V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s')$ for any s'. Thus,

$$Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) - Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \geqslant b_{t,h}(s,a) + \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a)\,, \qquad (10)$$

and a similar inequality for values, as the latter are obtained as convex combinations of Q–values, where the convex weights are determined solely by the common policy $\boldsymbol{\pi}^\star$ used. Therefore, $(\mathcal{H}_h)$ holds at least on the event

$$\mathcal{G}_{t,h} \stackrel{\text{def}}{=} \left\{ \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \left| \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| \leqslant b_{t,h}(s,a) \right\}.$$

All in all, the inequalities required in the statement of the lemma thus hold on the intersection of the events $\mathcal{G}_{t,h}$ over $t \in [T]$ and $h \in [H-1]$.

To conclude the proof, it suffices to show that this intersection is of probability at least $1-\delta$. By considering the complements and by a union bound, it suffices to show that any event

$$\overline{\mathcal{G}}_{t,h,s,a} \stackrel{\text{def}}{=} \left\{ \left| (\widehat{P}_{t,h} - P_h) \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| > b_{t,h}(s,a) \right\}$$

is of probability at most $\delta/(SATH)$. We partition the probability space based on the value of $\ell_{t-1,h}(s,a)$, resort to optional skipping and Corollary 1, with the deterministic function $g = V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}$ (see the proof of Lemma 8), to get the first inequality below, and to Lemma 8 for the second inequality below. We also use the definition (4) of $b_{t,h}$:

$$\begin{aligned}
\mathbb{P}\big(\overline{\mathcal{G}}_{t,h,s,a}\big) &= \sum_{\ell=0}^{\lceil \log_2(T) \rceil} \mathbb{P}\Big(\overline{\mathcal{G}}_{t,h,s,a} \cap \big\{\ell_{t-1,h}(s,a) = \ell\big\}\Big) \\
&= \sum_{\ell=1}^{\lceil \log_2(T) \rceil} \mathbb{P}\Bigg\{ \left| (\widehat{P}_{t,h} - P_h) \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| > \sqrt{\frac{2H^2 \log(2SATH \log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H \\
&\qquad\qquad \text{and } \ell_{t-1,h}(s,a) = \ell \Bigg\} \\
&\leqslant \sum_{\ell=1}^{\lceil \log_2(T) \rceil} \mathbb{P}\left\{ \left| \frac{1}{2^{\ell-1}} \sum_{j=1}^{2^{\ell-1}} V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(\sigma_{h,s,a,j}) - P_h \cdot V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| \right. \\
&\qquad\qquad \left. > \sqrt{\frac{2H^2 \log(2SATH \log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H \right\} \\
&\leqslant \lceil \log_2(T) \rceil \, \frac{\delta}{SATH \log_2(2T)} \leqslant \frac{\delta}{SATH} \,,
\end{aligned}$$

where we used the fact that $b_{t,h}(s,a) = H$ is a trivial upper bound on the difference of values at hand in the case $\ell = 0$, which is why the element $\ell = 0$ gets dropped in the summation in the second equality. □

D Term $(\mathbf{D})$

We first restate and then prove Lemma 6.

Lemma 6. *With probability at least $1-\delta$, we have*

$$(\mathbf{D}) \leqslant 3\sqrt{H^4SAT\log(2SATH\log_2(2T)/\delta)} + H^2SA\,.$$

Proof. Since $b_{t,h}(s,a) \in [0,H]$, the Hoeffding–Azuma inequality implies that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t,\boldsymbol{b}_t,\boldsymbol{P}}(s_1) - \sum_{t=1}^{T}\sum_{h\in[H]} b_{t,h}(s_{t,h},a_{t,h}) \leqslant \sqrt{\frac{(H^2)^2\,T\log(1/\delta)}{2}}\,; \tag{11}$$

we crucially use here that the policies $\boldsymbol{\pi}_t$ only depend on information gathered during previous episodes $\tau \leqslant t-1$ and that the stochastic environment $\boldsymbol{P}$ considered in the definition of $(\mathbf{D})$ is the true underlying environment.

We fix $h \in [H-1]$ (recall that $b_{t,H} \equiv 0$) and a pair $(s',a') \in \mathcal{S}\times\mathcal{A}$, and show that

$$\sum_{t=1}^{T} b_{t,h}(s_{t,h},a_{t,h})\,\mathbb{I}\big\{(s_{t,h},,a_{t,h}) = (s',a')\big\} \leqslant H + \sqrt{2H^2\log(2SATH\log_2(2T)/\delta)\,n_{T,h}(s',a')}\,. \tag{12}$$

Indeed, there can only be at most one t such $(s_{t,h},,a_{t,h}) = (s',a')$ and $n_{t,h}(s',a') = 1$; for this t, we use the upper bound $b_{t,h}(s_{t,h},a_{t,h}) \leqslant H$. For t such that $n_{t,h}(s',a') \geqslant 2$, we have, by the definitions in Section 3.1, that $n_{t,h}(s',a') \geqslant n_{t-1,h}(s',a') \geqslant 2^{\ell_{t-1,h}(s',a')-1}$. Therefore, substituting this inequality in the definition (4) of $b_{t,h}$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} b_{t,h}(s_{t,h},a_{t,h})\,\mathbb{I}\big\{(s_{t,h},,a_{t,h}) = (s',a')\big\} \\
&\leqslant H + \sum_{t=1}^{T}\sqrt{\frac{2H^2\log(2SATH\log_2(2T)/\delta)}{n_{t,h}(s',a')}}\,\mathbb{I}\big\{(s_{t,h},,a_{t,h}) = (s',a')\big\}\,\mathbb{I}\big\{n_{t,h}(s',a') \geqslant 2\big\}\,,
\end{aligned}$$

where, using that the counters $n_{t,h}(s',a')$ vary (by $+1$) if and only if $(s_{t,h},,a_{t,h}) = (s',a')$, we also have

$$\sum_{t=1}^{T}\frac{1}{\sqrt{n_{t,h}(s',a')}}\,\mathbb{I}\big\{(s_{t,h},,a_{t,h}) = (s',a')\big\}\,\mathbb{I}\big\{n_{t,h}(s',a') \geqslant 2\big\} = \sum_{n=2}^{n_{T,h}(s',a')}\frac{1}{\sqrt{n}} \leqslant 2\sqrt{n_{T,h}(s',a')}\,.$$

We conclude the proof by noting first that for each $(s',a') \in \mathcal{S}\times\mathcal{A}$, by concavity of the root,

$$\sum_{(s',a')}\sqrt{n_{T,h}(s',a')} \leqslant \sqrt{SAT}\,,$$

so that summing (12) over $h \in [H-1]$ and $(s',a') \in \mathcal{S}\times\mathcal{A}$ yields

$$\sum_{t=1}^{T}\sum_{h\in[H]} b_{t,h}(s_{t,h},a_{t,h}) \leqslant H^2SA + 2H\sqrt{2H^2SAT\log(2SATH\log_2(2T)/\delta)}\,.$$

We combine this inequality with (11) and note that

$$\sqrt{\frac{H^4\,T\log(1/\delta)}{2}} \leqslant \sqrt{H^4SAT\log(2SATH\log_2(2T)/\delta)}$$

to get the claimed bound. ∎

E Term (C)

This section is devoted to the analysis of the term (C), which is the most involved part of the proof. We leverage the recently developed techniques of Zhang et al. (2023). We start by restating the claimed bound.

Lemma 7. *With probability at least $1-\delta$, it holds that*

$$(\mathbf{C}) \leqslant 2\sqrt{H^7SAT\big(\log_2(2T)\big)^3} + 2\sqrt{2H^5\,T\log_2(2T)\ln(2/\delta)} + SAH^3\,.$$

The proof starts with an application of the performance-difference lemma in case of different transition kernels (see, e.g., Russo, 2019, Lemma 3):

$$V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\boldsymbol{P}}(s_1) = \mathbb{E}_{\boldsymbol{\pi}_t,\boldsymbol{P}}\left[\sum_{h=1}^{H-1}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s'_h,a'_h)\right],$$

where the piece of notation $\mathbb{E}_{\boldsymbol{\pi}_t,\boldsymbol{P}}$ indicates (as in Section 2) that the expectation is taken over trajectories $(s'_1,a'_1,\,\ldots,\,s'_H,a'_H)$ started at $s'_1=s_1$ and induced by the policies $\boldsymbol{\pi}_t$ and the transition kernels $\boldsymbol{P}$. Actually, one such trajectory is exactly $(s_{t,1},a_{t,1},\,\ldots,\,s_{t,H},a_{t,H})$ and we could rewrite the considered expectation as a conditional expectation:

$$\mathbb{E}_{\boldsymbol{\pi}_t,\boldsymbol{P}}\left[\sum_{h=1}^{H-1}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s'_h,a'_h)\right]$$
$$= \mathbb{E}\left[\sum_{h=1}^{H-1}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h})\,\middle|\,\boldsymbol{\pi}_t,\,\boldsymbol{b}_t,\,\widehat{\boldsymbol{P}}_t\right].$$

Next, we apply the Hoeffding–Azuma inequality, by resorting to a lexicographic order on pairs (t,h) and by noting that the random variables at hand satisfy

$$\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h}) \in [-H^2,H^2]\,;$$

indeed, the sums $r_{t,h}+b_{t,h}$ lies in $[0,H+1]$ and the value functions are weighted sums of at most $H-1$ such terms. We obtain that with probability at least $1-\delta/2$,

$$\begin{aligned}(\mathbf{C}) &= \sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\boldsymbol{P}}(s_1)\\ &\leqslant \sum_{t=1}^{T}\sum_{h=1}^{H-1}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h}) + \sqrt{2H^5T\ln(2/\delta)}\,. \qquad (13)\end{aligned}$$

We fix $h\in[H-1]$ and use the decomposition

$$\begin{aligned}&\sum_{t=1}^{T}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h})\\ &\leqslant SAH^2 + \sum_{t=1}^{T}\sum_{(s,a)}\sum_{\ell=1}^{\lceil\log_2 T\rceil}\sum_{j=1}^{2^{\ell-1}}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a)\,\mathbb{I}\big\{(s_{t,h},a_{t,h})=(s,a)\big\} \qquad (14)\\ &\qquad\qquad \times\mathbb{I}\big\{n_{t,h}(s,a)=2^{\ell-1}+j\big\}\,;\end{aligned}$$

the term SAH^2 comes from the fact that for each pair (s,a), there is at most once round t when $(s_{t,h},a_{t,h})=(s,a)$ and $n_{t,h}(s,a)=1$. We prove below the following lemma.

Lemma 9. *For each pair (ℓ,j), where $\ell\in\big[\lceil\log_2 T\rceil\big]$ and $j\in[2^{\ell-1}]$, with probability at least $1-\delta/(4TH)$,*

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{(s,a)}\big(\widehat{P}_{t,h}-P_h\big)\cdot V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a)\,\mathbb{I}\big\{(s_{t,h},a_{t,h})=(s,a)\big\}\,\mathbb{I}\big\{n_{t,h}(s,a)=2^{\ell-1}+j\big\}\\ &\qquad\leqslant\sqrt{2H^4\,\frac{1}{2^{\ell-1}}\sum_{(s,a)}\mathbb{I}\big\{n_{T,h}(s,a)\geqslant 2^{\ell-1}+j\big\}\ln\frac{4H(T+1)^{1+SAH\lceil\log_2(T)\rceil}}{\delta}}\,.\end{aligned}$$

Remark 2. *The result of Lemma 9 could be extended to taking also a sum over all $h \in [H-1]$ instead of considering a fixed $h \in [H-1]$, which would result in a factor $\sqrt{2H^5}$ instead of $\sqrt{2H^4}$ in the upper bound. Below, we will rather sum the bound of Lemma 9 over $h \in [H-1]$, which will result in a $\sqrt{2H^6}$ factor in the final upper bound. While this is sub-optimal, we do so for the sake of simplicity and because the control of term $(\mathbf{B})$ leads anyway to a factor $\sqrt{H^7}$ in the final regret bound.*

We conclude the proof of Lemma 7 based on Lemma 9. There are at most $2T$ different pairs (ℓ, j) considered, so that all events considered in Lemma 9 hold simultaneously with probability at least $1 - \delta/(2H)$. In addition, a first application of Jensen's inequality guarantees that for each $1 \leqslant \ell \leqslant \lceil \log_2 T \rceil$,

$$\sum_{j=1}^{2^{\ell-1}} \sqrt{\frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{I}\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\}} \leqslant \sqrt{\sum_{(s,a)} \sum_{j=1}^{2^{\ell-1}} \mathbb{I}\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\}}\,,$$

and a second application yields

$$\begin{aligned}
&\sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \sqrt{\frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{I}\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\}} \\
&\qquad \leqslant \sqrt{\lceil \log_2 T \rceil \sum_{(s,a)} \underbrace{\sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \mathbb{I}\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\}}_{= n_{T,h}(s,a)-1}} \leqslant \sqrt{T \lceil \log_2 T \rceil}\,.
\end{aligned}$$

Therefore, substituting the bound above (together with Lemma 9) into (14), we proved so far that with probability at least $1 - \delta/2$,

$$\begin{aligned}
&\sum_{t=1}^{T} (\widehat{P}_{t,h} - P_h) \cdot V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_{t,h}, a_{t,h}) \\
&\qquad \leqslant SAH^2 + \sqrt{2H^4 T \lceil \log_2 T \rceil \ln \frac{4H(T+1)^{1+SAH\lceil \log_2(T) \rceil}}{\delta}} \\
&\qquad \leqslant SAH^2 + 2\sqrt{H^5 SAT \big(\log_2(2T)\big)^3} + \sqrt{2H^4 T \lceil \log_2 T \rceil \ln(1/\delta)}\,.
\end{aligned}$$

Summing this bound over $h \in [H-1]$ and combining the outcome with (13) leads to

$$(\mathbf{C}) \leqslant SAH^3 + 2\sqrt{H^7 SAT \big(\log_2(2T)\big)^3} + \sqrt{2H^4 T \lceil \log_2 T \rceil \ln(1/\delta)} + \sqrt{2H^5 T \ln(2/\delta)}\,,$$

and thus to the upper bound claimed in Lemma 7.

It therefore only remains to prove Lemma 9.

Proof. We denote by

$$\tau_{\ell,j,h}(s,a) \stackrel{\text{def}}{=} \begin{cases} t & \text{if } (s_{t,h}, a_{t,h}) = (s,a) \text{ and } n_{t,h}(s,a) = 2^{\ell-1}+j\,, \\ +\infty & \text{if } n_{T,h}(s,a) \leqslant 2^{\ell-1}+j-1\,, \end{cases}$$

the stopping time whether and when (s,a) was reached in stage h for the $(2^{\ell-1}+j)$–th time, with the convention $\tau_{\ell,j,h}(s,a) = +\infty$ if (s,a) was reached fewer times than that. To apply optional skipping, we will partition the underlying probability space according to the values of all the $\ell_{t',h'}(s',a')$ as t', h', s', a' vary and of the $\tau_{\ell,j,h}(s',a')$ as s', a' only vary.

Part 1: Hoeffding–Azuma inequality. We fix consistent sequences $k_{t',h'}(s',a') \in [T]$ and $\kappa_{\ell,j,h}(s',a')$ of values for the $\ell_{t',h'}(s',a')$ and the $\tau_{\ell,j,h}(s',a')$; in particular, $k_{\kappa_{\ell,j,h}(s,a)-1,h}(s,a) = \ell$. The notation in the display below is heavy but the high-level idea is simple to grasp: only rounds

$t = \kappa_{\ell,j,h}(s,a)$ matter, and we know to which global epoch each of these rounds belongs and, in particular, we know which averages are in the components $\widehat{P}_{t,h}(\,\cdot \mid s',a')$ of $\widehat{P}_{t,h}$.

We rewrite the quantity at hand on the event associated with the sequences fixed:

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{(s,a)} \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a)\, \mathbb{I}\big\{(s_{t,h}, a_{t,h}) = (s,a)\big\}\, \mathbb{I}\big\{n_{t,h}(s,a) = 2^{\ell-1} + j\big\} \\
&\quad \times \prod_{(s',a') \in \mathcal{S}\times\mathcal{A}} \left(\mathbb{I}\big\{\tau_{\ell,j,h}(s',a') = \kappa_{\ell,j,h}(s',a')\big\} \prod_{t'=1}^{T} \prod_{h' \in [H-1]} \mathbb{I}\big\{\ell_{t',h'}(s',a') = k_{t',h'}(s',a')\big\} \right) \\
\leqslant & \sum_{(s,a):\kappa_{\ell,j,h}(s,a) \leqslant T} \big(\widehat{P}_{\kappa_{\ell,j,h}(s,a),h} - P_h\big) \cdot \widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}(s,a) \\
&\qquad \times \prod_{(s',a') \in \mathcal{S}\times\mathcal{A}} \prod_{h' \in [H-1]} \mathbb{I}\big\{\ell_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a') = k_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a')\big\}\,,
\end{aligned} \tag{15}$$

where we used the short-hand notation $\widehat{V}_{t,h+1} \stackrel{\text{def}}{=} V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}$.

We are now ready to apply optional skipping—a concept recalled in Section 4.2. On the events

$$\mathcal{C}' \stackrel{\text{def}}{=} \bigcap_{(s',a') \in \mathcal{S}\times\mathcal{A}} \bigcap_{h' \in [H-1]} \big\{\ell_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a') = k_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a')\big\}$$

considered, the empirical averages $\widehat{P}_{\kappa_{\ell,j,h}(s,a),h'}(\,\cdot \mid s',a')$ have the same distributions as the empirical frequency vectors associated with the i.i.d. random variables

$$\sigma_{h',s',a',j}\,, \qquad j \in \big[2^{k_{\kappa_{\ell,j,h}(s,a)-1,h'}-1}\big]\,,$$

and are independent from each other as h', s', a' vary. In particular, $\widehat{P}_{\kappa_{\ell,j,h}(s,a),h}(\,\cdot \mid s,a)$ is distributed as the empirical frequency vector of $2^{\ell-1}$ i.i.d. random variables $\sigma_{h,s,a,j}$, with $j \in [2^{\ell-1}]$. In addition, Bellman's equations (see the beginning of Section 3.1) show that on the events $\mathcal{C}'$ considered to apply optional skipping, $\widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}$ only depends on the $\pi_{\kappa_{\ell,j,h}(s,a),h'}$ with $h' \geqslant h+1$, on the $\widehat{P}_{\kappa_{\ell,j,h}(s,a),h'}(\,\cdot \mid s',a')$ with $h' \geqslant h+1$, and on state-action pairs relative to stages $h' \geqslant h+1$. Given the form of the adversarial learning strategy used, we conclude that on the events $\mathcal{C}'$ considered to apply optional skipping, all the $\widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}$, as s,a vary, only depend on state-action pairs of stages $h' \geqslant h+1$ and are therefore independent from all the $\widehat{P}_{\kappa_{\ell,j,h}(s',a'),h}$, as s',a' vary.

Put differently, optional skipping entails here that for all $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{ \sum_{(s,a):\kappa_{\ell,j,h}(s,a) \leqslant T} \big(\widehat{P}_{\kappa_{\ell,j,h}(s,a),h} - P_h\big) \cdot \widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}(s,a) > \varepsilon \right\} \cap \mathcal{C}' \right) \\
&\leqslant \mathbb{P}\left\{ \sum_{(s,a):\kappa_{\ell,j,h}(s,a) \leqslant T} \frac{1}{2^{\ell-1}} \sum_{j \in [2^{\ell-1}]} \big(\widetilde{V}_{s,a,h+1}(\sigma_{h,s,a,j}) - P_h \widetilde{V}_{s,a,h+1}(s,a)\big) > \varepsilon \right\},
\end{aligned} \tag{16}$$

for some $\widetilde{V}_{s,a,h+1}$ independent from all the $\sigma_{h,s,a,j}$ as s,a,j vary (and h is fixed). We recall that the $\sigma_{h,s,a,j}$ are independent from each other as s,a,j vary (and h is fixed).

By the independencies noted above, and by boundedness of the values functions in the interval $[0,(H-h)(H+1)] \subseteq [0,H^2]$, the Hoeffding–Azuma inequality guarantees that for all $\delta' \in (0,1)$,

$$\begin{aligned}
\mathbb{P}\Bigg\{ &\sum_{(s,a):\kappa_{\ell,j,h}(s,a) \leqslant T} \frac{1}{2^{\ell-1}} \sum_{j \in [2^{\ell-1}]} \big(\widetilde{V}_{s,a,h+1}(\sigma_{h,s,a,j}) - P_h \widetilde{V}_{s,a,h+1}(s,a)\big) \\
&> \sqrt{\frac{1}{2 \times 2^{\ell-1}} \sum_{(s,a):\kappa_{\ell,j,h}(s,a) \leqslant T} H^4 \ln \frac{1}{\delta'}} \Bigg\} = \delta'\,.
\end{aligned} \tag{17}$$

We summarize what we proved so far. Denoting by

$$\Delta_{T,\ell,j} \stackrel{\text{def}}{=} \sum_{t=1}^{T} \sum_{(s,a)} \big(\widehat{P}_{t,h} - P_h\big) \cdot V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a)\, \mathbb{I}\big\{(s_{t,h}, a_{t,h}) = (s,a)\big\}\, \mathbb{I}\big\{n_{t,h}(s,a) = 2^{\ell-1} + j\big\}$$

the target quantity, and by

$$\mathcal{C} \stackrel{\text{def}}{=} \bigcap_{(s',a') \in \mathcal{S} \times \mathcal{A}} \left(\big\{\tau_{\ell,j,h}(s',a') = \kappa_{\ell,j,h}(s',a')\big\} \bigcap_{t' = \in [T]} \bigcap_{h' \in [H-1]} \big\{\ell_{t',h'}(s',a') = k_{t',h'}(s',a')\big\} \right)$$

the event associated with the values fixed, the bounds (15)–(16)–(17) show that for all $\delta' \in (0,1)$,

$$\begin{aligned} &\mathbb{P}\left(\left\{ \Delta_{T,\ell,j} > \sqrt{2H^4 \frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{I}\big\{n_{T,h}(s,a) \geqslant 2^{\ell-1} + j\big\} \ln \frac{1}{\delta'}} \right\} \cap \mathcal{C} \right) \\ &= \mathbb{P}\left(\left\{ \Delta_{T,\ell,j} > \sqrt{\frac{1}{2^\ell} \sum_{(s,a) : \kappa_{\ell,j,h}(s,a) \leqslant T} H^2 \ln \frac{1}{\delta'}} \right\} \cap \mathcal{C} \right) \leqslant \delta' \,. \end{aligned} \tag{18}$$

Part 2: Union bound and counting the sequences. The proof is concluded by counting how many different sets $\mathcal{C}$ may be obtained. We do so in a rough way, that will be sufficient for our purposes. First, we need to count the profile values (7). There are $T(H-1)$ functions $\ell_{t,h} : \mathcal{S} \times \mathcal{A} \to \big[\lceil \log_2(T) \rceil\big]^*$, satisfying some monotonicity constraints, as well as some other constraints which we ignore. The monotonicity constraints imply that for each (s,a) and $h \in [H-1]$, it is sufficient to determine the at most $\lceil \log_2(T) \rceil$ time steps t among $[T]$ when $\ell_{t,h}$ increases by 1. Thus, there are at most

$$(T+1)^{\lceil \log_2(T) \rceil}$$

possible sequences of values for the $\ell_{t,h}(s,a)$ as t varies and h, s, a are fixed. All in all, the profile part in the number of different sets $\mathcal{C}$ is smaller than

$$\big((T+1)^{\lceil \log_2(T) \rceil}\big)^{SA(H-1)} \,.$$

For stopping times, we need to determine, for each (s,a), a single value, in a set included in $[T] \cup \{+\infty\}$. We neglect other constraints and see that there are therefore at most $(T+1)^{SA}$ such choices.

As a conclusion, there are at most

$$M = (T+1)^{SAH\lceil \log_2(T) \rceil}$$

different possible values for the sets $\mathcal{C}$. The proof of Lemma 9 is concluded by a union bound over the events (18), with $\delta' = \delta/(4HTM)$. □