Online Markov Decision Processes Under Bandit Feedback

Gergely Neu, András György, Csaba Szepesvári, András Antos

Online Markov Decision Processes under Bandit Feedback

Gergely Neu, András György, Csaba Szepesvári, András Antos

***Abstract*—We consider online learning in finite stochastic Markovian environments where in each time step a new reward function is chosen by an oblivious adversary. The goal of the learning agent is to compete with the best stationary policy in hindsight in terms of the total reward received. Specifically, in each time step the agent observes the current state and the reward associated with the last transition, however, the agent does not observe the rewards associated with other state-action pairs. The agent is assumed to know the transition probabilities. The state of the art result for this setting is an algorithm with an expected regret of $O(T^{2/3} \ln T)$. In this paper, assuming that stationary policies mix uniformly fast, we show that after T time steps, the expected regret of this algorithm (more precisely, a slightly modified version thereof) is $O\left(T^{1/2} \ln T\right)$, giving the first rigorously proven, essentially tight regret bound for the problem.**

I. Introduction

In this paper we consider online learning in finite stochastic Markovian environments where in each time step a new reward function may be chosen by an oblivious adversary. The interaction between the learner and the environment is shown in Figure 1. The environment is split into two parts: One part has a controlled Markovian dynamics, while another one has an unrestricted, uncontrolled (autonomous) dynamics. In each discrete time step t, the learning agent receives the state of the Markovian environment ($\mathbf{x}_t \in \mathcal{X}$) and some information ($y_{t-1} \in \mathcal{Y}$) about the previous state of the autonomous dynamics. The learner then makes a decision about the next action ($\mathbf{a}_t \in \mathcal{A}$), which is sent to the environment. In response, the environment makes a transition: the next state $\mathbf{x}_{t+1}$ of the Markovian part is drawn from a transition probability kernel $P(\cdot|\mathbf{x}_t, \mathbf{a}_t)$, while the other part makes a transition in an autonomous fashion. In the meanwhile, the agent incurs a reward $\mathbf{r}_t = r(\mathbf{x}_t, \mathbf{a}_t, y_t) \in [0, 1]$ that depends on the *complete* state of the environment and the chosen action; then the process continues with the next step. The goal of the learner is to collect as much reward as possible. The agent knows the transition probability kernel P and the reward function r, however, he does not know the sequence y_t in advance. We call this problem *online learning in Markov Decision Processes* (MDPs).

We take the viewpoint that the uncontrolled dynamics might be very complex and thus modeling it based on the available limited information might be hopeless. Equivalently, we assume that whatever can be modeled about the environment is modeled in the Markovian, controlled part. As a result, when evaluating the performance of the learner, the total reward of the learner will be compared to that of the *best stochastic stationary policy in hindsight* that assigns actions to the states of the Markovian part in a random manner. This stationary policy is thus selected as the policy that maximizes the total reward given the sequence of reward functions $r_t(\cdot, \cdot) \equiv r(\cdot, \cdot, y_t)$, $t = 1, 2, \ldots$.[1] Given a horizon $T > 0$, any policy π and initial distribution uniquely determines a distribution over the sequence space $(\mathcal{X} \times \mathcal{A})^T$. Noting that the expected total reward of π is then a linear function of the distribution of π and that the space of distributions is a convex polytope with vertices corresponding to distributions of deterministic policies, we see that there will always exist a deterministic policy that maximizes the total expected reward in T time steps. Hence, it is enough to consider deterministic policies only as a reference. To make the objective more precise, for a given stationary deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ let $(\mathbf{x}_t^\pi, \mathbf{a}_t^\pi)$ denote the state-action pair that would have been visited in time step t had one used policy π from the beginning of time (the initial state being fixed). Then, the goal can be expressed as keeping the *(expected) regret*,

$$\hat{L}_T = \max_\pi \mathbb{E}\left[\sum_{t=1}^T r_t(\mathbf{x}_t^\pi, \mathbf{a}_t^\pi)\right] - \mathbb{E}\left[\sum_{t=1}^T \mathbf{r}_t\right]$$

small, *regardless of the sequence of reward functions* $\{r_t\}_{t=1}^T$. In particular, a sublinear regret-growth, $\hat{L}_T = o(T)$ ($T \to \infty$) means that the *average* reward collected by the learning agent approaches that of the best policy in hindsight. Naturally, a smaller growth-rate is more desirable.[2]

The motivation to study this problem is manifold. One viewpoint

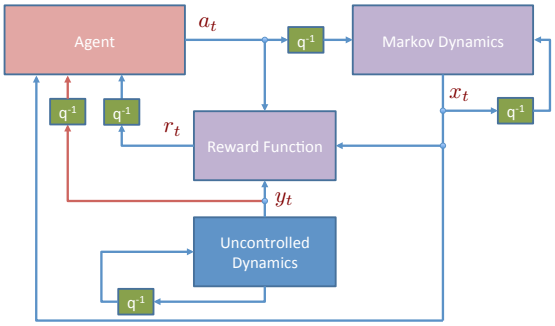

Fig. 1. The interaction between the learning agent and the environment. Here q^{-1} denotes a unit delay, that is, any information sent through such a box is received at the beginning of the next time step.

This research was supported in part by the National Development Agency of Hungary from the Research and Technological Innovation Fund (KTIA-OTKA CNK 77782), the Alberta Innovates Technology Futures, and the Natural Sciences and Engineering Research Council (NSERC) of Canada. Parts of this work have been published at the Twenty-Fourth Annual Conference on Neural Information Processing Systems (NIPS 2010) [16].

G. Neu is with the Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary, and with the Computer and Automation Research Institute of the Hungarian Academy of Sciences, Budapest, Hungary (email: neu.gergely@gmail.com).

A. György is with the Department of Computing Science, University of Alberta, Edmonton, Canada (email: gya@cs.bme.hu). During parts of this work he was with the Machine Learning Research Group of the Computer and Automation Research Institute of the Hungarian Academy of Sciences.

Cs. Szepesvári is with the Department of Computing Science, University of Alberta, Edmonton, Canada (email: szepesva@ualberta.ca).

A. Antos is with the Budapest University of Technology and Economics, Budapest, Hungary (email: antos@cs.bme.hu). During parts of this work he was with the Machine Learning Research Group of the Computer and Automation Research Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

[1] It is worth noting that the problem can be defined without referring to the uncontrolled, unmodelled dynamics by starting with an arbitrary sequence of reward functions $\{r_t\}$. That the two problems are equivalent follows because there is no restriction on the range of $\{y_t\}$ or its dynamics.

[2] Following previous works in the area, in this paper we only consider regret relative to a fixed stationary policy. However, as usual in online learning, our results and algorithms can also be extended to less restricted sets of reference policies, such as the class of sequences of stationary policies with a restricted number of switches. We discuss such extensions in Section IV-D.

is that a learning agent achieving sublinear regret growth shows *robustness* in the face of arbitrarily assigned rewards, thus, the model provides a useful generalization of learning and acting in Markov Decision Processes. Some examples where the need for such robustness arises naturally are discussed below. Another viewpoint is that this problem is a useful generalization of online learning problems studied in the machine learning literature (e.g., [5]). In particular, in this literature, the problems studied are so-called prediction problems that involve an (oblivious) environment that chooses a sequence of loss functions. The learner's predictions are elements in the common domain of these loss functions and the goal is to keep the regret small as compared with the best fixed prediction in hindsight. Identifying losses with negative rewards we may notice that this problem coincides exactly with our model with $|\mathcal{X}|= 1$, that is, our problem is indeed a generalization of this problem where the reward functions have memory represented by multiple states subject to the Markovian control.

Let us now consider some examples that fit the above model. Generally, since our approach assumes that the hard-to-model, uncontrolled part influences the rewards only, the examples concern cases where the reward is difficult to model. This is the case, for example, in various production- and resource-allocation problems, where the major source of difficulty is to model the prices that influence the rewards. Indeed, the prices in these problems tend to depend on external, generally unobserved factors and thus dynamics of the prices might be hard to model. Other examples include problems coming from computer science, such as the k-server problem, paging problems, or web-optimization (e.g., ad-allocation problems with delayed information) [see, e.g., 7, 22].

Previous results that concern online learning in MDPs (with known transition probability kernels) are summarized in Table I. In the current paper we study the problem with recurrent Markovian dynamics while assuming that the *only information received about the uncontrolled part is in the form of the actual reward* $\mathbf{r}_t$. In particular, in our model the agent does not receive y_t, while in most previous works it was assumed that y_t is observed [6, 7, 22]. Following the terminology used in the online learning literature [2], when y_t is available (equivalently, the agent receives the reward *function* $r_t : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ in every time step), we say that learning happens under *full information*, while in our case we say that learning happens under *bandit feedback* (note that Even-Dar et al. [7] suggested as an open problem to address the bandit situation studied here). In an earlier version of this paper [16], we provided an algorithm, MDP-EXP3, for learning in MDPs with recurrent dynamics under bandit feedback, and showed that it achieves a regret of order $\tilde{O}(T^{2/3})$.[3] In this paper we improve upon the analysis of [16] and prove an $\tilde{O}(T^{1/2})$-regret bound for the same algorithm. As it follows from a lower bound proven by Auer et al. [2] for bandit problems, apart from logarithmic and constant terms the rate obtained is unimprovable. The improvement compared to [16] is achieved by a more elaborate proof technique that builds on a (perhaps) novel observation that the so-called exponential weights technique (that our algorithm builds upon) changes its weights "slowly". As in previous works where "loopy" Markovian dynamics were considered, our main assumptions on the MDP transition probability kernel will be that stationary policies mix uniformly fast. In addition, we shall assume that the stationary distributions of these policies are bounded away from zero. These assumptions will be discussed later.

We also mention here that Yu and Mannor [20, 21] considered the related problem of online learning in MDPs where the transition probabilities may also change arbitrarily after each transition. This problem is significantly more difficult than the case where only the reward function is allowed to change. Accordingly, the algorithms proposed in these papers do not achieve sublinear regret. Unfortunately, these papers have also gaps in the proofs, as discussed in detail in [13].

Finally, we note in passing that the contextual bandit problem considered by Lazaric and Munos [12] can also be regarded as a simplified version of our online learning problem where the states are generated in an i.i.d. fashion (though we do not consider the problem of competing with the best policy in a restricted subset of stationary policies). For regret bounds concerning learning in purely stochastic *unknown* MDPs, see the work of Jaksch et al. [10] and the references therein. Learning in adversarial MDPs without loops was also considered by György et al. [8] for deterministic transitions under bandit feedback, and under full information but with *unknown* transition probability kernels in our recent paper [14].

The rest of the paper is organized as follows: The problem is laid out in Section II, which is followed by a section that makes our assumptions precise (Section III). The algorithm and the main result are given and discussed in Section IV, with the proofs presented in Section V.

paper	algorithm	**feedback**	**loops**	**regret bound**
Even-Dar et al. [6, 7]	MDP-E	full information	yes	$\tilde{O}(T^{1/2})$
Yu et al. [22]	LAZY-FPL[1]	full information	yes	$\tilde{O}(T^{3/4+\epsilon})$, $\epsilon > 0$
Yu et al. [22]	Q-FPL[2]	bandit	yes	$o(T)$
Neu et al. [13]		bandit	no	$O(T^{1/2})$
Neu et al. [16]	MDP-EXP3	bandit	yes	$\tilde{O}(T^{2/3})$
this paper	MDP-EXP3	bandit	yes	$\tilde{O}(T^{1/2})$

TABLE I
SUMMARY OF PREVIOUS RESULTS. PREVIOUS WORKS CONCERNED PROBLEMS WITH EITHER FULL-INFORMATION OR BANDIT FEEDBACK, PROBLEMS WHEN THE MDP DYNAMICS MAY OR MAY NOT HAVE LOOPS (TO BE MORE PRECISE, IN NEU ET AL. [13] WE CONSIDERED EPISODIC MDPS WITH RESTARTS). FOR EACH PAPER, THE ORDER OF THE OBTAINED REGRET BOUND IN TERMS OF THE TIME HORIZON T IS GIVEN.

[1] The Lazy-FPL algorithm has smaller computational complexity than MDP-E.
[2] The stochastic regret of Q-FPL was shown to be sublinear almost surely (not only in expectation).

II. NOTATION AND PROBLEM DEFINITION

The purpose of this section is to provide the formal definition of our problem and to set the goals. We start with some preliminaries, in particular by reviewing the language we use in connection to Markov Decision Processes (MDPs). This will be followed by the definition of the online learning problem. We assume that the reader is familiar with the concepts necessary to study MDPs, our purpose here is to introduce the notation only. For more background about MDPs, consult Puterman [17].

We define a finite Markov Decision Process (MDP) M by a finite state space $\mathcal{X}$, a finite action set $\mathcal{A}$, a transition probability kernel $P : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$, and a reward function $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$. At time $t \in \{1, 2, \dots\}$, based on the sequence of past states, observed rewards, and actions, $(\mathbf{x}_1, \mathbf{a}_1, r(\mathbf{x}_1, \mathbf{a}_1), \mathbf{x}_2, \dots, \mathbf{x}_{t-1}, \mathbf{a}_{t-1}, r(\mathbf{x}_{t-1}, \mathbf{a}_{t-1}), \mathbf{x}_t) \in (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^{t-1} \times \mathcal{X}$, an agent acting in the MDP M chooses an action $\mathbf{a}_t \in \mathcal{A}$ to be executed.[4] As a result, the process moves to state $\mathbf{x}_{t+1} \in \mathcal{X}$ with probability

[3] Here, $\tilde{O}(g(s))$ denotes the class of functions $f : \mathbb{N} \to \mathbb{R}^+$ satisfying $\sup_{s\in\mathbb{N}} \frac{f(s)}{g(s)\ln^\alpha(g(s))} < \infty$ for some $\alpha \geq 0$.

[4] Throughout the paper we will use boldface letters to denote random variables.

$P(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t)$ and the agent incurs the reward $r(\mathbf{x}_t, \mathbf{a}_t)$. We note in passing that at the price of increased notational load, but with essentially no change to the contents, we could consider the case where the set of actions available at time step t is restricted to a non-empty subset $\mathcal{A}(\mathbf{x}_t)$ of all actions, where the set-system, $(\mathcal{A}(x))_{x\in\mathcal{X}}$, is known to the agent. However, for simplicity, in the rest of the paper we stick to the case $\mathcal{A}(x) = \mathcal{A}$. In an MDP the goal of the agent is to maximize the long-term reward. In particular, in the so-called *average-reward problem*, the goal of the agent is to maximize the long-run average reward. In what follows, the symbols $x, x', \ldots$ will be reserved to denote a state in $\mathcal{X}$, while a, a', b will be reserved to denote an action in $\mathcal{A}$. In expressions involving sums over $\mathcal{X}$, the domain of $x, x', \ldots$ will be suppressed to avoid clutter. The same holds for sums involving actions.

Before defining the learning problem, let us introduce some more notation. We use $\|v\|_p$ to denote the L_p-norm of a function or a vector. In particular, for $p = \infty$ the supremum norm of a function $v : S \to \mathbb{R}$ is defined as $\|v\|_\infty = \sup_{s\in S} |v(s)|$, and for $1 \le p < \infty$ and for any vector $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$, $\|u\|_p = \left(\sum_{i=1}^d |u_i|^p\right)^{1/p}$. We use $e_1, \ldots, e_d$ to denote the *row* vectors of the canonical basis of the Euclidean space $\mathbb{R}^d$. Since we will identify $\mathcal{X}$ with the integers $\{1, \ldots, |\mathcal{X}|\}$, we will also use the notation e_x for $x \in \mathcal{X}$. We will use $\ln$ to denote the natural logarithm function.

A. Online learning in MDPs

In this paper we consider a so-called *online learning problem* when the reward function is allowed to change arbitrarily in every time step. That is, instead of a single reward function r, a sequence of reward functions $\{r_t\}$ is given. This sequence is assumed to be fixed ahead of time, and, for simplicity, we assume that $r_t(x, a) \in [0, 1]$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $t \in \{1, 2, \ldots\}$. No other assumptions are made about this sequence.

The learning agent is assumed to know the transition probabilities P, but is not given the sequence $\{r_t\}$. The protocol of interaction with the environment is unchanged: At time step t the agent selects an action $\mathbf{a}_t$ based on the information available to it, which is sent to the environment. In response, the reward $r_t(\mathbf{x}_t, \mathbf{a}_t)$ and the next state $\mathbf{x}_{t+1}$ are communicated to the agent. The initial state $\mathbf{x}_1$ is generated from a fixed distribution P_1, which may or may not be known.

Let the *expected total reward* collected by the agent up to time T be denoted by

$$\widehat{R}_T = \mathbb{E}\left[\sum_{t=1}^T r_t(\mathbf{x}_t, \mathbf{a}_t)\right].$$

As before, the goal of the agent is to make this sum as large as possible. In classical approaches to learning one would assume some kind of regularity of r_t and then derive bounds on how much reward the learning agent loses as compared to the agent that knew about the regularity of the rewards and who acted optimally from the beginning of time. The loss or *regret*, measured in terms of the difference of total expected rewards of the two agents, quantifies the learner's efficiency. In this paper, following the recent trend in the machine learning literature [5], while keeping the regret criterion, we will avoid making any assumption on how the reward sequence is generated, and take a worst-case viewpoint. The potential benefit is that the results will be more generally applicable and the algorithms will enjoy added robustness, while, generalizing from results available for supervised learning [4, 11, 18], the algorithms can also be shown to avoid being too pessimistic.

The concept of regret in our case is defined as follows: We shall consider algorithms which are competitive with stochastic stationary policies. Fix a (stochastic) stationary policy $\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$ and let $\{(\mathbf{x}'_t, \mathbf{a}'_t)\}$ be the trajectory that results from following policy π from $\mathbf{x}'_1 \sim P_1$ (in particular, $\mathbf{a}'_t \sim \pi(\cdot|\mathbf{x}'_t) \stackrel{\text{def}}{=} \pi(\mathbf{x}'_t, \cdot)$). The *expected total reward of* π over the first T time steps is defined as

$$R_T^\pi = \mathbb{E}\left[\sum_{t=1}^T r_t(\mathbf{x}'_t, \mathbf{a}'_t)\right].$$

Now, the *(expected) regret* (or *expected relative loss*) of the learning agent relative to the class of stationary policies is defined as

$$\hat{L}_T = \sup_\pi R_T^\pi - \widehat{R}_T,$$

where the supremum is taken over all stochastic stationary policies in M. Note that the policy maximizing the total expected reward is chosen in hindsight, that is, based on the knowledge of the reward functions $r_1, \ldots, r_T$. Thus, the regret measures how well the learning agent is able to generalize from its moment to moment knowledge of the rewards to the sequence $r_1, \ldots, r_T$. If the regret of an agent grows sublinearly with T then it can be said to act as well as the best (stochastic stationary) policy in the long run (i.e., the average expected reward of the agent in the limit is equal to that of the best policy). In this paper our main result will show that there exists an algorithm such that if that algorithm is followed by the learning agent, then the learning agent's regret will be bounded by $C\sqrt{T}\ln T$, where $C > 0$ is a constant that depends on the transition probability kernel, but is independent of the sequence of rewards $\{r_t\}$.

III. Assumptions on the transition probability kernel

Before describing our assumptions, a few more definitions are needed: First of all, for brevity, in what follows we will call stochastic stationary policies just policies. Further, without loss of generality, we shall identify the states with the first $|\mathcal{X}|$ integers and assume that $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$. Now, take a policy π and define the Markov kernel $P^\pi(x'|x) = \sum_a \pi(a|x)P(x'|x, a)$. The identification of $\mathcal{X}$ with the first $|\mathcal{X}|$ integers makes it possible to view P^π as a matrix: $(P^\pi)_{x,x'} = P^\pi(x'|x)$. In what follows, we will also take this view when convenient.

In general, distributions will also be treated as *row* vectors. Hence, for a distribution μ over $\mathcal{X}$, μP^π is the distribution over $\mathcal{X}$ that results from using policy π for one step after a state is sampled from μ (i.e., the "next-state distribution" under π). Finally, a stationary distribution of a policy π is a distribution μ_{st} that satisfies $\mu_{\text{st}} P^\pi = \mu_{\text{st}}$.

In what follows we assume that every (stochastic stationary) policy π has a well-defined unique stationary distribution μ_{st}^π. This ensures that the average reward underlying any stationary policy is a well-defined single real number. It is well-known that in this case the convergence to the stationary distribution is exponentially fast. Following Even-Dar et al. [7], we consider the following stronger, "uniform mixing condition" (which implies the existence of the unique stationary distributions):

Assumption A1: There exists a number $\tau \ge 0$ such that for any policy π and any pair of distributions μ and μ' over $\mathcal{X}$,

$$\left\|(\mu - \mu')P^\pi\right\|_1 \le e^{-1/\tau} \left\|\mu - \mu'\right\|_1. \tag{1}$$

As Even-Dar et al. [7], we call the smallest τ satisfying this assumption the *mixing time* of the transition probability kernel P. Together with the existence and uniqueness of the stationary policy, the next assumption ensures that every state is visited eventually no matter what policy is chosen:

Assumption A2: The stationary distributions are uniformly bounded away from zero:

$$\inf_{\pi,x} \mu_{\text{st}}^\pi(x) \ge \beta > 0$$

for some $\beta \in \mathbb{R}$.

Note that $e^{-1/\tau}$ is the supremum over all policy π of the Markov-Dobrushin coefficient of ergodicity, defined as $m_{P^\pi} = \sup_{\mu \neq \mu'} \frac{\|(\mu-\mu')P^\pi\|_1}{\|\mu-\mu'\|_1}$ for the transition probability kernel P^π, see, e.g., [9]. It is also known that $m_{P^\pi} = 1 - \min_{x,x' \in \mathcal{X}} \sum_{y \in \mathcal{X}} \min\{P^\pi(y|x), P^\pi(y|x')\}$ [9]. Since m_{P^π} is a continuous function of π and the set of policies is compact, there is a policy π' with $m_{P^{\pi'}} = \sup_\pi m_{P^\pi}$. These facts imply that Assumption A1 is satisfied, that is, $\sup_\pi m_{P^\pi} < 1$, if and only if for every π, $m_{P^\pi} < 1$, that is, P^π is a *scrambling* matrix (P^π is a scrambling matrix if any two rows of P^π share some column in which they both have a positive element). Furthermore, if P^π is a scrambling matrix for any deterministic policy π then it is also a scrambling matrix for any stochastic policy. Thus, to guarantee Assumption A1 it is enough to verify mixing for deterministic policies only. The assumptions will be further discussed in Section IV-D.

IV. Learning in online MDPs under bandit feedback

In this section we shall first introduce some additional, standard MDP concepts that we will need. That these concepts are well-defined follows from our assumptions on P and from standard results to be found, for example, in the book by Puterman [17]. After the definitions, we specify our algorithm. The section is finished by the statement of our main result concerning the performance of the proposed algorithm.

A. Preliminaries

Fix an arbitrary policy π and $t \geq 1$. Let $\{(\mathbf{x}'_s, \mathbf{a}'_s)\}$ be a random trajectory generated by π and the transition probability kernel P and an arbitrary everywhere positive initial distribution over the states. We will use q_t^π to denote the *action-value function* underlying π and the immediate reward r_t, while we will use v_t^π to denote the corresponding *(state) value function*.[5] That is, for $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$q_t^\pi(x,a) = \mathbb{E}\left[\sum_{s=1}^\infty \left(r_t(\mathbf{x}'_s, \mathbf{a}'_s) - \rho_t^\pi\right) \,\middle|\, \mathbf{x}'_1 = x, \mathbf{a}'_1 = a\right],$$
$$v_t^\pi(x) = \mathbb{E}\left[\sum_{s=1}^\infty \left(r_t(\mathbf{x}'_s, \mathbf{a}'_s) - \rho_t^\pi\right) \,\middle|\, \mathbf{x}'_1 = x\right],$$

where ρ_t^π is the *average reward per stage* corresponding to π:

$$\rho_t^\pi = \lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^S \mathbb{E}[r_t(\mathbf{x}'_s, \mathbf{a}'_s)].$$

The average reward per stage can be expressed as

$$\rho_t^\pi = \sum_x \mu_{\mathrm{st}}^\pi(x) \sum_a \pi(a|x) r_t(x,a),$$

where μ_{st}^π is the stationary distribution underlying policy π. Under our assumptions stated in the previous section, up to a shift by a constant function, the value functions q_t^π, v_t^π are the unique solutions to the *Bellman equations*

$$\begin{aligned} q_t^\pi(x,a) &= r_t(x,a) - \rho_t^\pi + \sum_{x'} P(x'|x,a) v_t^\pi(x'), \\ v_t^\pi(x) &= \sum_a \pi(a|x) q_t^\pi(x,a), \end{aligned} \tag{2}$$

which hold simultaneously for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ (Corollary 8.2.7 of [17]). We will use q_t^* to denote the *optimal action-value function*, that is, the action-value function underlying a policy that maximizes the average-reward in the MDP specified by (P, r_t). We will also need these concepts for an arbitrary reward function $r\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. In such a case, we will use v^π, q^π, and ρ^π to denote the respective value function, action-value function, and average reward of a policy π.

Now, consider the trajectory $\{(\mathbf{x}_t, \mathbf{a}_t)\}$ followed by a learning agent with $\mathbf{x}_1 \sim P_1$. For any $t \geq 1$, define

$$\mathbf{u}_t = (\mathbf{x}_1, \mathbf{a}_1, r_1(\mathbf{x}_1, \mathbf{a}_1), \ \ldots, \ \mathbf{x}_t, \mathbf{a}_t, r_t(\mathbf{x}_t, \mathbf{a}_t)) \tag{3}$$

and introduce the policy followed in time step t, $\boldsymbol{\pi}_t(a|x) = \mathbb{P}[\mathbf{a}_t = a | \mathbf{u}_{t-1}, \mathbf{x}_t = x]$, where $\mathbf{u}_0$ and, more generally $\mathbf{u}_s$ for all $s \leq 0$ is defined to be the empty sequence. Note that $\boldsymbol{\pi}_t$ is computed based on past information and is therefore random. We introduce the following notation:

$$\mathbf{q}_t = q_t^{\boldsymbol{\pi}_t}, \qquad \mathbf{v}_t = v_t^{\boldsymbol{\pi}_t}, \qquad \boldsymbol{\rho}_t = \rho_t^{\boldsymbol{\pi}_t}.$$

With this, we see that the following equations hold simultaneously for all $(x,a) \in \mathcal{X} \times \mathcal{A}$:

$$\begin{aligned} \mathbf{q}_t(x,a) &= r_t(x,a) - \boldsymbol{\rho}_t + \sum_{x'} P(x'|x,a) \mathbf{v}_t(x'), \\ \mathbf{v}_t(x) &= \sum_a \boldsymbol{\pi}_t(a|x) \mathbf{q}_t(x,a). \end{aligned} \tag{4}$$

B. The algorithm

Our algorithm, MDP-Exp3, shown as Algorithm 1, is inspired by that of Even-Dar et al. [7], while also borrowing ideas from the Exp3 algorithm (exponential weights algorithm for exploration and exploitation) of Auer et al. [2]. The main idea of the algorithm is to construct estimates $\{\hat{\mathbf{q}}_t\}$ of the action-value functions $\{\mathbf{q}_t\}$, which are then used to determine the action-selection probabilities $\boldsymbol{\pi}_t(\cdot|x)$ in each state x in each time step t. In particular, the probability of selecting action a in state x at time step t is computed as the mixture of the uniform distribution (which encourages exploring actions irrespective of what the algorithm has learned about the action-values) and a Gibbs distribution, the mixture parameter being $\gamma > 0$. Given a state x, the Gibbs distribution defines the probability of choosing action a at time step t to be proportional to $\exp(\eta \sum_{s=N}^{t-N} \hat{\mathbf{q}}_s(x,a))$.[6]

Algorithm 1 MDP-Exp3: an algorithm for online learning in MDPs

Set $N \geq 1$, $\mathbf{w}_1(x,a) = \mathbf{w}_2(x,a) = \cdots = \mathbf{w}_{2N-1}(x,a) = 1$, $\gamma \in (0,1)$, $\eta \in (0,\gamma]$.

For $t = 1, 2, \ldots$ repeat:

1) Set
$$\boldsymbol{\pi}_t(a|x) = (1-\gamma)\frac{\mathbf{w}_t(x,a)}{\sum_b \mathbf{w}_t(x,b)} + \frac{\gamma}{|\mathcal{A}|}$$
for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.
2) Draw an action $\mathbf{a}_t \sim \boldsymbol{\pi}_t(\cdot|\mathbf{x}_t)$.
3) Receive reward $r_t(\mathbf{x}_t, \mathbf{a}_t)$ and observe $\mathbf{x}_{t+1}$.
4) If $t \geq N$
   a) Compute $\boldsymbol{\mu}_t^N$ for all $x \in \mathcal{X}$ using (8).
   b) Construct estimates $\hat{\mathbf{r}}_t$ using (6) and compute $\hat{\mathbf{q}}_t$ using (5).
   c) Set $\mathbf{w}_{t+N}(x,a) = \mathbf{w}_{t+N-1}(x,a) e^{\eta \hat{\mathbf{q}}_t(x,a)}$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.

[5] Most sources would call these functions *differential* action- and state-value functions. We omit this adjective for brevity.

[6] In the algorithm the Gibbs action-selection probabilities are computed in an incremental fashion with the help of the "weights" $\mathbf{w}_t(x,a)$. Note that a numerically stable implementation would calculate the action-selection probabilities based on the relative value differences, $\sum_{s=N}^{t-N} \hat{\mathbf{q}}_s(x,\cdot) - \max_{a \in A} \sum_{s=N}^{t-N} \hat{\mathbf{q}}_s(x,a)$. These relative value differences can also be updated incrementally. The form shown in Algorithm 1 is preferred for mathematical clarity.

Here, $\eta > 0$, $N > 0$ are further parameters of the algorithm. Note that for the single-state setting with $N = 1$, MDP-EXP3 is equivalent to the EXP3 algorithm of Auer et al. [2].

It is interesting to discuss how the Gibbs policy (i.e., $\frac{\mathbf{w}_t(x,\cdot)}{\sum_b \mathbf{w}_t(x,b)}$) is related to what is known as the Boltzmann-exploration policy in the reinforcement learning literature [e.g., 19]. Remember that given a state x, the Boltzmann-exploration policy would select action a at time step t with probability proportional to $\exp(\eta\hat{\mathbf{q}}^*_{t-1}(x,a))$ for some estimate $\hat{\mathbf{q}}^*_{t-1}$ of the optimal action-value function in the MDP $(P, \hat{\mathbf{r}}_{t-1})$, where $\{\hat{\mathbf{r}}_t\}$ is the sequence of estimated reward functions. Thus, we can see a couple of differences between the Boltzmann exploration and our Gibbs policy. The first difference is that the Gibbs policy in our algorithm uses the cumulated sum of the estimates of action-values, while the Boltzmann policy uses only the last estimate. By depending on the sum, the Gibbs policy will rely less on the last estimate. This reduces how fast the policies can change, making the learning "smoother". Another difference is that in our Gibbs policy the sum of previous action-values runs only up to step $t-N$ instead of using the sum that runs up to the last step $t-1$. The reasons for doing this will be explained below. Finally, the Gibbs policy uses the action-value function estimates (in the MDPs $\{(P, \hat{\mathbf{r}}_s)\}$) of the policies $\{\boldsymbol{\pi}_s\}$ selected by the algorithm, as opposed to using an estimate of the optimal action-value function. This makes our algorithm closer in spirit to (modified) policy iteration than to value iteration and is again expected to reduce the variance of the learning process.

The reason the Gibbs policy does not use the last N estimates is to allow the construction of a reasonable estimate $\hat{\mathbf{q}}_t$ of the action-value function $\mathbf{q}_t$. If r_t was available, one could compute $\mathbf{q}_t$ based on r_t (cf. (4)) and the sum could then run up to $t-1$, resulting in the algorithm of Even-Dar et al. [7]. Since in our problem r_t is not available, we estimate it using an importance sampling estimator $\hat{\mathbf{r}}_t$ below (from now on, $t \geq N$). Given this $\hat{\mathbf{r}}_t$, the estimate $\hat{\mathbf{q}}_t$ of the action-value function $\mathbf{q}_t$ is defined as the action-value function underlying policy $\boldsymbol{\pi}_t$ in the average-reward MDP given by the transition probability kernel P and reward function $\hat{\mathbf{r}}_t$. Thus, $\hat{\mathbf{q}}_t$, up to a shift by a constant function, can be computed as the solution to the Bellman equations corresponding to $(P, \hat{\mathbf{r}}_t)$ (cf. (4)):

$$\begin{aligned}
\hat{\mathbf{q}}_t(x,a) &= \hat{\mathbf{r}}_t(x,a) - \hat{\boldsymbol{\rho}}_t + \sum_{x'} P(x'|x,a)\hat{\mathbf{v}}_t(x')\,, \\
\hat{\mathbf{v}}_t(x) &= \sum_{a'} \boldsymbol{\pi}_t(a'|x)\hat{\mathbf{q}}_t(x,a')\,, \\
\hat{\boldsymbol{\rho}}_t &= \sum_{x',a'} \mu^{\boldsymbol{\pi}_t}_{\mathrm{st}}(x')\boldsymbol{\pi}_t(a'|x')\hat{\mathbf{r}}_t(x',a')\,,
\end{aligned} \tag{5}$$

which hold simultaneously for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. Since $\boldsymbol{\pi}_t$ is invariant to constant shifts of $\hat{\mathbf{q}}_t$, any of the solutions of these equations leads to the same sequence of policies. *Hence, in what follows, without loss of generality we assume that the algorithm uses $\hat{\mathbf{q}}_t$, i.e., the value function of $\boldsymbol{\pi}_t$ in the average-reward MDP defined by $(P, \hat{\mathbf{r}}_t)$.*

To define the estimator $\hat{\mathbf{r}}_t$ define $\boldsymbol{\mu}^N_t(x)$ as the probability of visiting state x at time step t, conditioned on the history $\mathbf{u}_{t-N}$ up to time step $t-N$, including $\mathbf{x}_{t-N}$ and $\mathbf{a}_{t-N}$ (cf. (3) for the definition of $\{\mathbf{u}_t\}$):

$$\boldsymbol{\mu}^N_t(x) \stackrel{\mathrm{def}}{=} \mathbb{P}\left[\mathbf{x}_t = x \,|\, \mathbf{u}_{t-N}\right], \qquad x \in \mathcal{X}.$$

Then, the estimate of r_t is constructed using

$$\hat{\mathbf{r}}_t(x,a) = \begin{cases} \frac{r_t(x,a)}{\boldsymbol{\pi}_t(a|x)\boldsymbol{\mu}^N_t(x)}, & \text{if } (x,a) = (\mathbf{x}_t, \mathbf{a}_t)\,; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The importance sampling estimator (6) is well-defined only if for $x = \mathbf{x}_t$,

$$\boldsymbol{\mu}^N_t(x) > 0 \tag{7}$$

holds almost surely (by construction $\boldsymbol{\pi}_t(\cdot|\mathbf{x}_t) > \gamma/|\mathcal{A}| > 0$). To see the intuitive reason of why (7) holds, it is instructive to look into how the distribution $\boldsymbol{\mu}^N_t$ can be computed.

When $t = N$, it should be clear from the definition of $\boldsymbol{\mu}^N_t$ that, viewing $\boldsymbol{\mu}^N_t$ as a row vector, $\boldsymbol{\mu}^N_N = P_1(P^{\boldsymbol{\pi}_1})^{N-1}$. Now let $t > N$. Denote by P^a the transition probability matrix of the policy that selects action a in every state and recall that e_x denotes the x^{th} unit row vector of the canonical basis of the $|\mathcal{X}|$-dimensional Euclidean space. We may write

$$\boldsymbol{\mu}^N_t = e_{\mathbf{x}_{t-N}} P^{\mathbf{a}_{t-N}} P^{\boldsymbol{\pi}_{t-N+1}} \cdots P^{\boldsymbol{\pi}_{t-1}}\,, \qquad t > N. \tag{8}$$

This holds because for any $t \geq N$, $\boldsymbol{\pi}_t$ is entirely determined by the history $\mathbf{u}_{t-N}$, while for $t > N$ the history $\mathbf{u}_{t-N}$ also includes (and thus determines) $\mathbf{x}_{t-N}, \mathbf{a}_{t-N}$. Using the notation $\mathbf{z} \in \sigma(\mathbf{u}_{t-N})$ to denote that the random variable $\mathbf{z}$ is measurable with respect to the sigma-algebra generated by the history $\mathbf{u}_{t-N}$, the above fact can be stated as

$$\begin{aligned}
\mathbf{x}_{t-N}, \mathbf{a}_{t-N} &\in \sigma(\mathbf{u}_{t-N}) \;\text{ for } t > N, \\
\boldsymbol{\pi}_t &\in \sigma(\mathbf{u}_{t-N}) \;\text{ for } t \geq N.
\end{aligned} \tag{9}$$

Consequently, we also have that $\boldsymbol{\pi}_{t-1}, \ldots, \boldsymbol{\pi}_{t-N+1} \in \sigma(\mathbf{u}_{t-N})$ and therefore (8) follows from the law of total probability. Note also that

$$\mathbb{P}\left[\mathbf{a}_t = a | \mathbf{x}_t = x, \mathbf{u}_{t-N}\right] = \mathbb{P}\left[\mathbf{a}_t = a | \mathbf{x}_t = x, \mathbf{u}_{t-1}\right] = \boldsymbol{\pi}_t(a|x), \tag{10}$$

where the last equality follows from the definition of $\boldsymbol{\pi}_t$ and $\mathbf{a}_t$.

The algorithm as presented needs to know P_1 to compute $\boldsymbol{\mu}^N_t$ at step $t = N$. When P_1 is unknown, instead of starting the computation of the weights at time step $t = N$, we can start the computation at time step $t = N+1$ (i.e., change $t \geq N$ of step 4 to $t \geq N+1$). Clearly, in the worst-case, the regret can only increase by a constant amount (the magnitude of the largest reward) as a result of this change.

An essential step of the proof of our main result is to show that inequality (7) indeed holds, that is, $\boldsymbol{\mu}^N_t(x)$ is bounded away from zero. In fact, we will show that this inequality holds almost surely[7] for all $x \in \mathcal{X}$ provided that N is large enough, which explains why the sum in the definition of the Gibbs policy runs from time N. This will be done by first showing that the policies $\boldsymbol{\pi}_t$ (especially, during the last $N-1$ steps) change "sufficiently slowly" (this is where it becomes useful that the Gibbs policy is defined using a sum of previous action values). Consequently, $\boldsymbol{\pi}_{t-N+1}, \ldots, \boldsymbol{\pi}_{t-1}$ will all be "quite close" to the policy of the last time step. Then, the expression on the right-hand side of (8) can be seen to be close to the $N-1$-step state distribution of $\boldsymbol{\pi}_t$ when starting from $(\mathbf{x}_{t-N}, \mathbf{a}_{t-N})$, which, if N is large enough, will be shown to be close to the stationary distribution of $\boldsymbol{\pi}_t$ thanks to Assumption A1. Since by Assumption A2, $\min_{x \in \mathcal{X}} \mu^{\boldsymbol{\pi}_t}_{\mathrm{st}}(x) \geq \beta > 0$ then, by choosing the algorithm's parameters appropriately, we can show that $\boldsymbol{\mu}^N_t(x) \geq \beta/2 > 0$ holds for all $x \in \mathcal{X}$, that is, inequality (7) follows. This is shown in Lemma 13.

It remains to be seen that the estimate $\hat{\mathbf{r}}_t$ is meaningful. In this regard, we claim that

$$\mathbb{E}\left[\hat{\mathbf{r}}_t(x,a) \,|\, \mathbf{u}_{t-N}\right] = r_t(x,a) \tag{11}$$

[7] In what follows, for the sake of brevity, unless otherwise stated, we will omit the modifier "almost surely" from probabilistic statements. It is worth to mention that the finiteness of $\mathcal{X}$ and $\mathcal{A}$ allows several statements concerning conditional expectations to hold always, instead of almost surely.

holds for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. First note that

$$\mathbb{E}\left[\hat{\mathbf{r}}_t(x,a)|\mathbf{u}_{t-N}\right] = \frac{r_t(x,a)}{\boldsymbol{\pi}_t(a|x)\boldsymbol{\mu}_t^N(x)}\mathbb{E}\left[\mathbb{I}_{\{(x,a)=(\mathbf{x}_t,\mathbf{a}_t)\}}|\mathbf{u}_{t-N}\right],$$

where we have exploited that $\boldsymbol{\pi}_t, \boldsymbol{\mu}_t^N \in \sigma(\mathbf{u}_{t-N})$. Now,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{I}_{\{(x,a)=(\mathbf{x}_t,\mathbf{a}_t)\}} \,|\, \mathbf{u}_{t-N}\right] \\ &\qquad = \mathbb{P}\left[\mathbf{a}_t = a \,|\, \mathbf{x}_t = x, \mathbf{u}_{t-N}\right]\mathbb{P}\left[\mathbf{x}_t = x \,|\, \mathbf{u}_{t-N}\right].\end{aligned}$$

By definition, $\mathbb{P}\left[\mathbf{x}_t = x \,|\, \mathbf{u}_{t-N}\right] = \boldsymbol{\mu}_t^N(x)$ and by (10), $\mathbb{P}\left[\mathbf{a}_t = a \,|\, \mathbf{x}_t = x, \mathbf{u}_{t-N}\right] = \boldsymbol{\pi}_t(a|x)$. Putting together the equalities obtained, we get (11).

By linearity of expectation and since $\boldsymbol{\pi}_t, \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t} \in \sigma(\mathbf{u}_{t-N})$, it then follows from (5) and (11) that $\mathbb{E}[\hat{\boldsymbol{\rho}}_t|\mathbf{u}_{t-N}] = \boldsymbol{\rho}_t$, and, hence, by the linearity of the Bellman equations and by our assumption that $\hat{\mathbf{q}}_t$ is the value function underlying the MDP $(P, \hat{\mathbf{r}}_t)$ and policy $\boldsymbol{\pi}_t$, we have, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{q}}_t(x,a)|\mathbf{u}_{t-N}] &= \mathbf{q}_t(x,a), \\ \mathbb{E}[\hat{\mathbf{v}}_t(x)|\mathbf{u}_{t-N}] &= \mathbf{v}_t(x).\end{aligned} \tag{12}$$

As a consequence, we also have, for all $(x,a) \in \mathcal{X} \times \mathcal{A}, t \geq N$,

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\rho}}_t] &= \mathbb{E}\left[\boldsymbol{\rho}_t\right], \\ \mathbb{E}[\hat{\mathbf{q}}_t(x,a)] &= \mathbb{E}\left[\mathbf{q}_t(x,a)\right], \\ \mathbb{E}[\hat{\mathbf{v}}_t(x)] &= \mathbb{E}\left[\mathbf{v}_t(x)\right].\end{aligned} \tag{13}$$

Let us finally comment on the computational complexity of our algorithm. Due to the delay in updating the policies based on the weights, the algorithm needs to store N policies (or weights, leading to the policies). Thus, the memory requirement of MDP-EXP3 scales with $N|\mathcal{A}||\mathcal{X}|$ (in the real-number model). The computational complexity of the algorithm is dominated by the cost of computing $\hat{\mathbf{r}}_t$ and, in particular, by the cost of computing $\boldsymbol{\mu}_t^N$, plus the cost of solving the Bellman equations (5). The cost of this is $O\left(|\mathcal{X}|^2(N + |\mathcal{X}|+|\mathcal{A}|)\right)$ in the worst case, for each time step, however, it can be much smaller for specific practical cases such as when the number of possible next-states is limited.

C. Main result

Our main result is the following bound concerning the performance of MDP-EXP3.

Theorem 1 (Regret under bandit feedback): Let the transition probability kernel P satisfy Assumptions A1 and A2. Let $T > 0$ and let $N = 1 + \lceil \tau \ln T \rceil$, and $h(y) = 2y \ln y$ for $y > 0$. Then for an appropriate choice of the parameters η and γ (which depend on $|\mathcal{A}|, T, \beta, \tau$), for *any* sequence of reward functions $\{r_t\}$ taking values in $[0,1]$, for

$$T > \max\left\{c_1 \frac{\left(|\mathcal{A}|\tau + \frac{\tau^3}{|\mathcal{A}|}\right)\ln|\mathcal{A}|}{\beta^3}, h\left(c_2 \frac{\left(\frac{|\mathcal{A}|}{\tau} + \frac{\tau}{|\mathcal{A}|}\right)\ln|\mathcal{A}|}{\beta}\right)\right\}$$

and $\tau \geq 1$[8] the regret of the algorithm MDP-EXP3 can be bounded as

$$\widehat{L}_T \leq C\sqrt{\frac{\tau^3 T|\mathcal{A}|\ln(|\mathcal{A}|)\ln(T)}{\beta}} + C'\tau^2 \ln T$$

for some universal constants $c_1, c_2, C, C' > 0$.

Note that with the specific choice of parameters the total cost of the algorithm for a time horizon of T is $O\left(T|\mathcal{X}|^2(\tau\ln(T) + |\mathcal{X}|+|\mathcal{A}|)\right)$.

[8]The choice of the lower bound on τ is arbitrary, but the constants in the theorem depend on it. Furthermore, with some extra work, our proof also gives rise to a bound for the case when $\tau \to 0$, but for simplicity we decided to leave out this analysis.

The proof is presented in the next section. For comparison, we give now the analogue result for the algorithm of Even-Dar et al. [7] that was developed for the full-information case when the algorithm is given r_t in each time step. As hinted on before, our algorithm reduces to this algorithm if we set $N = 1$, $\hat{\mathbf{r}}_t = r_t$ and $\gamma = 0$. We call this algorithm MDP-E after Even-Dar et al. [7]. The following regret bound holds for this algorithm:

Theorem 2 (Regret under full-information feedback): Fix $T > 0$. Let the transition probability kernel P satisfy Assumption A1. Then, for an appropriate choice of the parameter η (which depends on $|\mathcal{A}|, T, \tau$), for *any* sequence of reward functions $\{r_t\}$ taking values in $[0,1]$, the regret of the algorithm MDP-E can be bounded as

$$\hat{L}_T \leq 4(\tau+1) + \sqrt{2T(2\tau+3)(2\tau^2+6\tau+5)\ln|\mathcal{A}|}\,. \tag{14}$$

For pedagogical reasons, we shall present the proof in the next section, too. Note that the constants in this bound are different from those presented in Theorem 5.1 of Even-Dar et al. [7]. In particular, the leading term here is $2\tau^{3/2}\sqrt{2T\ln|\mathcal{A}|}$, while their leading term is $4\tau^2\sqrt{T\ln|\mathcal{A}|}$. The above bound both corrects some small mistakes in their calculations and improves the result at the same time.[9]

As Even-Dar et al. [7] note, the regret bound (14) does not depend directly on the number of states, $|\mathcal{X}|$, but the dependence appears implicitly through τ only. Even-Dar et al. [7] also note that a tighter bound, where only the mixing times of the actual policies chosen appear, can be derived. However, it is unclear whether in the worst-case this could be used to improve the bound. Similarly to (14), our bound depends on $|\mathcal{X}|$ through other constants. In the bandit case, these are β and τ. Comparing the theorems it seems that the main price of not seeing the rewards is the appearance of $|\mathcal{A}|$ instead of $\ln|\mathcal{A}|$ (a typical difference between the bandit and full observation cases) and the appearance of a $\sqrt{1/\beta}$ term in the bound.

D. Discussion and future work

In this paper, we have presented an online learning algorithm, MDP-EXP3 for adversarial MDPs, that is, finite stochastic Markovian decision environments where the reward function may change after each transition. This is the first algorithm for this setting that has a rigorously proved $O(\sqrt{T\ln T})$ bound on its regret. We discuss the features of the algorithm, along with future research directions below.

a) Extensions: We considered the expected regret relative to the best fixed policy selected in hindsight. A typical extension is to prove a high probability bound on the regret, which we think can be done in a standard way using concentration inequalities. Note, however, that the extension is more complicated than for the bandit problems because the mixing property has to be used together with the martingale reasoning. Another potential extension is to compete with larger policy classes, such as with sequences of policies with a bounded number of policy-switches. Similarly to Neu et al. [13, 15], the MDP-EXP3 algorithm should then be modified by replacing EXP3 with the EXP3.S algorithm of Auer et al. [2], specifically designed to compete with switching experts in place of EXP3. Note that, again, the analysis will be more complicated than in the bandit case, and requires to bound the maximum regret of EXP3.S relative to any fixed policy over any time window. When compared to a policy with C switches, the resulting regret bound is expected to be C times

[9]One of the mistakes is in the proof of Theorem 4.1 of Even-Dar et al. [7] where they failed to notice that $q_t^{\pi_t}$ can take on negative values. Thus, their Assumption 3.1 is not met by $\{q_t^{\pi_t}\}$ (one needs to extend the upper bound given in their Lemma 2.2 with a lower bound and change Assumption 3.1). As a result, Assumption 3.1 cannot be used to show that the inequality in the proof of Theorem 4.1 holds. This mistake, as well as the others, can easily be corrected, as we show it here.

larger than that of Theorem 1, while the algorithm would not need to know the number of switches C.

b) Tuning and complexity: Setting up and running the algorithm MDP-EXP3 may actually be computationally demanding. Setting the parameters of the algorithm (η and γ) requires a known lower bound β^* on the visitation probabilities such that $\beta = \inf_{\pi,x} \mu^\pi_{\rm st}(x) > \beta^* > 0$ and also the knowledge of an upper bound τ^* on the mixing time τ. While these quantities can be determined in principle from the transition probability kernel P, it is not clear how to compute efficiently the minimum over all policies. Computational issues also arise during running the algorithm: as it is discussed in Section IV-B, each step of the MDP-EXP3 algorithm requires $O\left(|\mathcal{X}|^2(\tau \ln T + |\mathcal{X}|+|\mathcal{A}|)\right)$ computations, which may be too demanding if, e.g., the size of the state space is large. It is an interesting problem to design a more efficient method that achieves similar performance guarantees.

c) Assumptions on the Markovian dynamics: We believe that it should be possible to extend our main result beyond Assumption A1, requiring only the existence of a unique stationary distribution for any policy π (we will refer to this latter assumption as the unichain assumption). Using that the distribution of any unichain Markov chain converges exponentially fast to its stationary distribution, and that it is enough to verify Assumption A1 for deterministic policies only, one can easily show that if P satisfies the unichain assumption, then there exists an integer $K > 0$ such that $(P^\pi)^K$ is a scrambling matrix for any policy π. Then, we conjecture that the MDP-EXP3 algorithm will work as it is, except that the regret will be increased. The key to prove this result is to generalize Lemmas 4 and 5 to this case.

Finally, one may also consider the case when the Markov chains corresponding to P^π are periodic. We speculate that this may be dealt with using *occupancy probabilities* and Cesaro-averages instead of the stationary and state distributions, respectively.

V. PROOFS

In this section we present the proofs of Theorem 1 and Theorem 2. We start with the proof of Theorem 2 as this is a simpler result. The proof of this result is presented partly for the sake of completeness and partly so that we can be more specific about the corrections required to fix the main result (Theorem 5.2) of Even-Dar et al. [7]. Further, the proof will also serve as a starting point for the proof of our main result, Theorem 1. Nevertheless, the impatient reader may skip this next section and jump immediately to the proof of Theorem 1, which apart from referring to some general lemmas developed in the next subsection, is entirely self-contained.

A. Proof of Theorem 2

Throughout this section we consider the MDP-E algorithm (given by Algorithm 1 with $N = 1$, $\hat{\mathbf{r}}_t = r_t$ and $\gamma = 0$), and we suppose that P satisfies Assumption A1. Let π_t denote the policy used in step t of the algorithm. Note that π_t is not random since by assumption the reward function is available at all states (not just the visited ones). Hence, the sequence of policies chosen does not depend on the states visited by the algorithm but is deterministic. Remember that $\rho_t = \rho^{\pi_t}_t$ denotes the average reward of policy π_t measured with respect to the reward function r_t. Following Even-Dar et al. [7], fix some policy π and consider the decomposition of the regret relative to π:

$$R^\pi_T - \widehat{R}_T = \left(R^\pi_T - \sum_{t=1}^T \rho^\pi_t\right) + \left(\sum_{t=1}^T \rho^\pi_t - \sum_{t=1}^T \rho_t\right) + \left(\sum_{t=1}^T \rho_t - \widehat{R}_T\right). \tag{15}$$

The first and the last terms measure the difference between the sum of (asymptotic) average rewards and the actual expected reward. The mixing assumption (Assumption A1) ensures that these differences are not large. In particular, in the case of a fixed policy, this difference is bounded by a constant of order τ:

Lemma 1: For any $T \geq 1$ and any policy π, it holds that

$$R^\pi_T - \sum_{t=1}^T \rho^\pi_t \leq 2\tau + 2\,. \tag{16}$$

This lemma is also stated in [7]. We give the proof for completeness also to correct slight inaccuracies of the proof given in [7].

Proof: Let $\{(\mathbf{x}_t, \mathbf{a}_t)\}$ be the trajectory when π is followed. Note that the difference between R^π_T and $\sum_{t=1}^T \rho^\pi_t$ is caused by the difference between the initial distribution of $\mathbf{x}_1$ and the stationary distribution of π. To quantify the difference, write

$$R^\pi_T - \sum_{t=1}^T \rho^\pi_t = \sum_{t=1}^T \sum_x (\nu^\pi_t(x) - \mu^\pi_{\rm st}(x)) \sum_a \pi(a|x) r_t(x,a),$$

where $\nu^\pi_t(x) = \mathbb{P}[\mathbf{x}_t = x]$ is the state distribution at time step t. Viewing ν^π_t as a row vector, we have $\nu^\pi_t = \nu^\pi_{t-1} P^\pi$. Consider the $t^{\rm th}$ term of the above difference. Then, using $r_t(x,a) \in [0,1]$ and Assumption A1 we get[10]

$$\begin{aligned}
&\sum_x (\nu^\pi_t(x) - \mu^\pi_{\rm st}(x)) \sum_a \pi(a|x) r_t(x,a) \\
&\quad \leq \|\nu^\pi_t - \mu^\pi_{\rm st}\|_1 = \|\nu^\pi_{t-1} P^\pi - \mu^\pi_{\rm st} P^\pi\|_1 \\
&\quad \leq e^{-1/\tau} \|\nu^\pi_{t-1} - \mu^\pi_{\rm st}\|_1 \leq \ldots \leq e^{-(t-1)/\tau} \|\nu^\pi_1 - \mu^\pi_{\rm st}\|_1 \\
&\quad \leq 2e^{-(t-1)/\tau}.
\end{aligned}$$

This, together with the elementary inequality $\sum_{t=1}^T e^{-(t-1)/\tau} \leq 1 + \int_0^\infty e^{-t/\tau}\, dt = 1 + \tau$ gives the desired bound. ∎

Consider now the second term of (15) and in particular its $t^{\rm th}$ term $\rho^\pi_t - \rho_t = \rho^\pi_t - \rho^{\pi_t}_t$. This term is the difference of the average reward obtained by π and π_t. The following lemma shows that this difference can be rewritten in terms of the state-wise action-disadvantages underlying π_t:

Lemma 2 (Performance difference lemma): Consider an MDP specified by the transition probability kernel P and reward function r. Let $\pi, \hat{\pi}$ be two (stochastic stationary) policies in the MDP. Assume that $\mu^\pi_{\rm st}$, $\rho^{\hat{\pi}}$ and $q^{\hat{\pi}}$ are well-defined.[11] Then,

$$\rho^\pi - \rho^{\hat{\pi}} = \sum_{x,a} \mu^\pi_{\rm st}(x) \pi(a|x) \left[q^{\hat{\pi}}(x,a) - v^{\hat{\pi}}(x)\right]\,.$$

This lemma appeared as Lemma 4.1 in [7], but similar statements have been known for a while. For example, the book of Cao [3] also puts performance difference statements in the center of the theory of MDPs. For the sake of completeness, we include the easy proof. Note that the statement of the lemma continues to hold even when $q_{\hat{\pi}}$ and $v^{\hat{\pi}}$ are shifted by the same constant function.

Proof: We have

$$\begin{aligned}
&\sum_{x,a} \mu^\pi_{\rm st}(x) \pi(a|x) q^{\hat{\pi}}(x,a) \\
&\quad = \sum_{x,a} \mu^\pi_{\rm st}(x) \pi(a|x) \left[r(x,a) - \rho^{\hat{\pi}} + \sum_{x'} P(x'|x,a) v^{\hat{\pi}}(x')\right] \\
&\quad = \rho^\pi - \rho^{\hat{\pi}} + \sum_x \mu^\pi_{\rm st}(x) v^{\hat{\pi}}(x),
\end{aligned}$$

where the second equality holds since $\sum_{x,a} \mu^\pi_{\rm st}(x)\pi(a|x)P(x'|x,a) = \mu^\pi_{\rm st}(x')$. Reordering the terms gives the desired result. ∎

[10] Even-Dar et al. [7] mistakenly uses $\|\nu^\pi_t - \mu^\pi_{\rm st}\|_1 \leq e^{-t/\tau} \|\nu^\pi_1 - \mu^\pi_{\rm st}\|_1$ in their paper ($t = 1$ immediately shows that this can be false). See, e.g., the proofs of their Lemmas 2.2 and 5.2.

[11] This lemma does not need Assumption A1 and in fact the assumptions we make could be further relaxed with a slight change to the claim.

Because of this lemma, $\rho_t^\pi - \rho_t = \sum_{x,a} \mu_{\rm st}^\pi(x)\pi(a|x)\left(q_t^{\pi_t}(x,a) - v_t^{\pi_t}(x)\right)$. Thus, by flipping the sum that runs over time with the one that runs over the state-action pairs, we get: $\sum_{t=1}^T \rho_t^\pi - \sum_{t=1}^T \rho_t = \sum_{x,a} \mu_{\rm st}^\pi(x)\pi(a|x)\sum_{t=1}^T \left(q_t^{\pi_t}(x,a) - v_t^{\pi_t}(x)\right)$. Thus, it suffices to bound, for a fixed state-action pair (x,a), the sum

$$\sum_{t=1}^T \left(q_t^{\pi_t}(x,a) - v_t^{\pi_t}(x)\right) = \sum_{t=1}^T \left(q_t^{\pi_t}(x,a) - \sum_{a'} \pi_t(a'|x) q_t^{\pi_t}(x,a')\right).$$

By construction, $\pi_t(a|x) \propto \exp(\eta \sum_{s=1}^{t-1} q_s^{\pi_s}(x,a))$ (recall that $\gamma = 0$ in this version of the algorithm), which means that the sum is the regret of the so-called exponential weights algorithm (EWA) against action a when the algorithm is used on the sequence $\{q_t^{\pi_t}(x,\cdot)\}$. Assume for a moment that $K > 0$ is such that $\|q_t^{\pi_t}\|_\infty \le K$ holds for $1 \le t \le T$. Then, since $q_t^{\pi_t}$ takes its values from an interval of length $2K$, Theorem 2.2 in [5] implies that the regret of EWA can be bounded by

$$\frac{\ln|\mathcal{A}|}{\eta} + \frac{K^2\eta T}{2}\,. \tag{17}$$

Notice that $\{q_t^{\pi_t}\}$ is a sequence that is sequentially generated from $\{r_t\}$. It is Lemma 4.1 of [5] that shows that the bound of Theorem 2.2 of [5] continues to hold for such sequentially generated functions. Putting the inequalities together, we obtain

$$\sum_{t=1}^T \rho_t^\pi - \sum_{t=1}^T \rho_t \le \frac{\ln|\mathcal{A}|}{\eta} + \frac{K^2\eta T}{2}\,. \tag{18}$$

According to the next lemma an appropriate value for K is $2\tau+3$. The lemma is stated in a greater generality than what is needed here because the more general form will be used later.

Lemma 3: Pick any policy π in an MDP (P,r). Assume that the mixing time of π is τ in the sense of (1). If $|\sum_a \pi(a|x) r(x,a)| \le R \le \|r\|_\infty$ holds for any $x \in \mathcal{X}$, then $|v^\pi(x)| \le 2R(\tau+1)$ holds for all $x \in \mathcal{X}$. Furthermore, for any $(x,a) \in \mathcal{X}\times\mathcal{A}$, $|q^\pi(x,a)| \le R(2\tau+3) + |r(x,a)|$ and, if, in addition, $r(x,a) \ge 0$ for any $(x,a) \in \mathcal{X}\times\mathcal{A}$, then $|q^\pi(x,a)| \le (2\tau+3)\|r\|_\infty$.

Proof: As it is well known and is easy to see from the definitions, the (differential) value of policy π at state x can be written as

$$v^\pi(x) = \sum_{s=1}^\infty \sum_{x'} (\nu_{s,x}^\pi(x') - \mu_{\rm st}^\pi(x')) \sum_a \pi(a|x') r(x',a),$$

where $\nu_{s,x}^\pi = e_x (P^\pi)^{s-1}$ is the state distribution when following π for $s-1$ steps starting from state x. The triangle inequality and then the bound on $\sum_a \pi(a|x')r(x',a)$ gives

$$|v^\pi(x)| \le R \sum_{s=1}^\infty \sum_{x'} |\nu_{s,x}^\pi(x') - \mu_{\rm st}^\pi(x')| \le 2R(\tau+1)\,,$$

where in the second inequality we used $\left\|\nu_{s,x}^\pi - \mu_{\rm st}^\pi\right\|_1 \le 2e^{-(s-1)/\tau}$ and that $\sum_{s=1}^\infty e^{-(s-1)/\tau} \le \tau+1$ (cf. the proof of Lemma 1). This proves the first inequality. The inequalities on $|q^\pi(x,a)|$ follow from the first part and the Bellman equation:

$$\begin{aligned}
|q^\pi(x,a)| &\le |r(x,a)| + |\rho^\pi| + \sum_{x'} P(x'|x,a)|v^\pi(x')| \\
&\le R(2\tau+3) + |r(x,a)|, \\
|q^\pi(x,a)| &\le |r(x,a) - \rho^\pi| + \sum_{x'} P(x'|x,a)|v^\pi(x')| \\
&\le (2\tau+3)\|r\|_\infty.
\end{aligned}$$

Here, in the first inequality we used that $|\rho^\pi| \le \sum_x \mu_{\rm st}^\pi(x) |\sum_a \pi(a|x) r(x,a)| \le R$, while the second inequality holds since $|r(x,a) - \rho^\pi|$, $R \in [0, \|r\|_\infty]$. ∎

Let us now consider the third term of (15), $\sum_{t=1}^T \rho_t - \widehat{R}_T$. The $t^{\rm th}$ term of this difference is the difference between the average reward of π_t and the expected reward obtained in step π_t. If $\nu_t(x)$ is the distribution of states in time step t, $\sum_{t=1}^T \rho_t - \widehat{R}_T = \sum_{t=1}^T \sum_x (\mu_{\rm st}^\pi(x) - \nu_t(x)) \sum_a \pi(a|x) r_t(x,a)$. Thus,

$$\sum_{t=1}^T \rho_t - \widehat{R}_T \le \sum_{t=1}^T \|\mu_{\rm st}^{\pi_t} - \nu_t\|_1 \tag{19}$$

and so remains to bound the ℓ^1 distances between the distributions $\mu_{\rm st}^{\pi_t}$ and ν_t. For this, we will use two general lemmas that will again come useful later. For $f : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$, introduce the mixed norm $\|f\|_{1,\infty} = \max_x \sum_a |f(a|x)|$, where $f(a|x)$ is identified with $f(x,a)$. Clearly, $\left\|\nu P^\pi - \nu P^{\hat\pi}\right\|_1 \le \|\pi - \hat\pi\|_{1,\infty}$ holds for any two policies $\pi, \hat\pi$ and any distribution ν (cf. Lemma 5.1 in [7]). The first lemma shows that the map $\pi \mapsto \mu_{\rm st}^\pi$ as a map from the space of stationary policies equipped with the mixed norm $\|\cdot\|_{1,\infty}$ to the space of distributions equipped with the ℓ^1-norm is $(\tau+1)$-Lipschitz:

Lemma 4: Let P be a transition probability kernel over $\mathcal{X}\times\mathcal{A}$ such that the mixing time of P is $\tau < \infty$. For any two policies, $\pi, \hat\pi$, it holds that

$$\left\|\mu_{\rm st}^\pi - \mu_{\rm st}^{\hat\pi}\right\|_1 \le (\tau+1)\|\pi - \hat\pi\|_{1,\infty}\,.$$

Proof: The statement follows from solving

$$\begin{aligned}
\left\|\mu_{\rm st}^\pi - \mu_{\rm st}^{\hat\pi}\right\|_1 &\le \left\|\mu_{\rm st}^\pi P^\pi - \mu_{\rm st}^{\hat\pi} P^\pi\right\|_1 + \left\|\mu_{\rm st}^{\hat\pi} P^\pi - \mu_{\rm st}^{\hat\pi} P^{\hat\pi}\right\|_1 \\
&\le e^{-1/\tau}\left\|\mu_{\rm st}^\pi - \mu_{\rm st}^{\hat\pi}\right\|_1 + \|\pi - \hat\pi\|_{1,\infty}
\end{aligned}$$

for $\left\|\mu_{\rm st}^\pi - \mu_{\rm st}^{\hat\pi}\right\|_1$ and using

$$1/(1 - e^{-1/\tau}) \le \tau + 1. \tag{20}$$

∎

The next lemma allows us to compare an n-step distribution under a policy sequence with the stationary distribution of the sequence's last policy:

Lemma 5: Let P be a transition probability kernel over $\mathcal{X}\times\mathcal{A}$ such that the mixing time of P is $\tau < \infty$. Take any probability distribution ν_1 over $\mathcal{X}$, integer $n \ge 1$ and policies $\pi_1, \ldots, \pi_n$. Consider the distribution $\nu_n = \nu_1 P^{\pi_1} \cdots P^{\pi_{n-1}}$. Then, it holds that

$$\|\nu_n - \mu_{\rm st}^{\pi_n}\|_1 \le 2e^{-(n-1)/\tau} + (\tau+1)^2 \max_{1\le t\le n} \|\pi_t - \pi_{t-1}\|_{1,\infty}\,,$$

where, for convenience, we have introduced $\pi_0 = \pi_1$.

Proof: If $n = 1$ the result is obtained from $\|\nu_1 - \mu_{\rm st}^{\pi_1}\|_1 \le 2$. Thus, in what follows we assume $n \ge 2$. Let $c = \max_{1\le t\le n} \|\pi_t - \pi_{t-1}\|_{1,\infty}$. By the triangle inequality,

$$\begin{aligned}
\|\nu_n - \mu_{\rm st}^{\pi_n}\|_1 &\le \left\|\nu_n - \mu_{\rm st}^{\pi_{n-1}}\right\|_1 + \left\|\mu_{\rm st}^{\pi_{n-1}} - \mu_{\rm st}^{\pi_n}\right\|_1 \\
&\le e^{-1/\tau}\left\|\nu_{n-1} - \mu_{\rm st}^{\pi_{n-1}}\right\|_1 + (\tau+1)c\,,
\end{aligned}$$

where we used that by the previous lemma $\left\|\mu_{\rm st}^{\pi_{n-1}} - \mu_{\rm st}^{\pi_n}\right\|_1 \le (\tau+1)\|\pi_{n-1} - \pi_n\|_{1,\infty} \le (\tau+1)c$. Continuing recursively, we get

$$\begin{aligned}
&\|\nu_n - \mu_{\rm st}^{\pi_n}\|_1 \\
&\le e^{-1/\tau}\left(e^{-1/\tau}\left\|\nu_{n-2} - \mu_{\rm st}^{\pi_{n-2}}\right\|_1 + (\tau+1)c\right) + (\tau+1)c \\
&\quad\vdots \\
&\le e^{-\frac{n-1}{\tau}}\|\nu_1 - \mu_{\rm st}^{\pi_1}\|_1 + (\tau+1)c\left(1 + e^{-\frac{1}{\tau}} + \ldots + e^{-\frac{n-2}{\tau}}\right) \\
&\le 2e^{-(n-1)/\tau} + (\tau+1)^2 c\,,
\end{aligned}$$

where we bounded the geometrical series by $1/(1-e^{-1/\tau})$ and used (20). ∎

Applying this lemma to $\|\nu_t - \mu_{\mathrm{st}}^{\pi_t}\|_1$ we get

$$\|\nu_t - \mu_{\mathrm{st}}^{\pi_t}\|_1 \le 2e^{-(t-1)/\tau} + (\tau+1)^2 K',$$

where K' is a bound on $\max_{2\le t\le n} \|\pi_t - \pi_{t-1}\|_{1,\infty}$.[12] Therefore, by (19), we have

$$\sum_{t=1}^{T} \rho_t - \widehat{R}_T \le 2\sum_{t=1}^{T} e^{-t/\tau} + (\tau+1)^2 K' T \le 2\tau + (\tau+1)^2 K' T.$$

Thus, it remains to find an appropriate value for K'. It is a well known property of EWA that $\|\pi_t(\cdot|x) - \pi_{t-1}(\cdot|x)\|_1 \le \eta \left\|q_{t-1}^{\pi_{t-1}}(x,\cdot)\right\|_\infty$. Indeed, applying Pinsker's inequality and Hoeffding's lemma (see Section A.2 and Lemma A.6 in Cesa-Bianchi and Lugosi 5), we get for any $x \in \mathcal{X}$

$$\begin{aligned}
&\|\pi_t(\cdot|x) - \pi_{t-1}(\cdot|x)\|_1 \le \sqrt{2D(\pi_{t-1}(\cdot|x)\|\pi_t(\cdot|x))} \\
&= \sqrt{2\left[\ln\left(\sum_b \pi_{t-1}(b|x) e^{\eta q_{t-1}^{\pi_{t-1}}(b,x)}\right) - \sum_a \eta \pi_{t-1}(a|x) q_{t-1}^{\pi_{t-1}}(b,x)\right]} \\
&\le \eta \|q_{t-1}^{\pi_{t-1}}(x,\cdot)\|_\infty
\end{aligned}$$

where, for two distributions $D(v\|v') = \sum_i v_i \ln(v_i/v_i')$ denotes the Kullback-Leibler divergence of the distributions v and v'. Thus, $\|\pi_t - \pi_{t-1}\|_{1,\infty} \le \eta \left\|q_{t-1}^{\pi_{t-1}}\right\|_\infty$. Now, by Lemma 3, $\|q_t^{\pi_t}\|_\infty \le 2\tau + 3$, showing that $K' = \eta(2\tau+3)$ is suitable. Putting together the inequalities obtained, we get

$$\sum_{t=1}^{T} \rho_t - \widehat{R}_T \le 2\tau + (2\tau+3)(\tau+1)^2 \eta T.$$

Combining (16), (18) and this last bound, we obtain

$$R_T^\pi - \widehat{R}_T = 4\tau + 2 + \frac{\ln|\mathcal{A}|}{\eta} + \frac{\eta T(2\tau+3)(2\tau^2+6\tau+5)}{2}.$$

Setting

$$\eta = \sqrt{\frac{2\ln|\mathcal{A}|}{T(2\tau+3)(2\tau^2+6\tau+5)}},$$

we get the bound stated in Theorem 2.

B. Proof of Theorem 1

Throughout this section we consider the MDP-EXP3 algorithm and suppose that both Assumptions A1 and A2 hold for P. We start from the decomposition (15), which is repeated to emphasize the difference that some of the terms are random now:

$$R_T^\pi - \widehat{R}_T = \left(R_T^\pi - \sum_{t=1}^{T} \rho_t^\pi\right) + \left(\sum_{t=1}^{T} \rho_t^\pi - \sum_{t=1}^{T} \boldsymbol{\rho}_t\right) + \left(\sum_{t=1}^{T} \boldsymbol{\rho}_t - \widehat{R}_T\right) \tag{21}$$

As before, Lemma 1 shows that the first term is bounded by $2(\tau+1)$. Thus, it remains to bound the expectation of the other two terms. This is done in the following two propositions whose proofs are deferred to the next subsections:

[12] Lemma 5.2 of Even-Dar et al. [7] gives a bound on $\|\nu_t - \mu_{\mathrm{st}}^{\pi_t}\|_1$ with a slightly different technique. However, there are multiple mistakes in the proof. Once the mistakes are removed, their bounding technique gives the same result as ours. One of the mistakes is that Assumption 3.1 states that $K' = \sqrt{\ln|\mathcal{A}|/T}$, whereas since the range of the action-value functions scales with τ, K' should also scale with τ. Unfortunately, in [16] we committed the same mistake, which we correct here. We choose to present an alternate proof, as we find it somewhat cleaner and it also gave us the opportunity to present Lemma 4.

Proposition 1: Let $L = \frac{2}{\beta}(2\tau+3)$, $V_{\hat{\mathbf{q}}} = \frac{2}{\beta}\left(\frac{|\mathcal{A}|}{\gamma} + 2\tau + 2\right)$, $U_{\hat{\mathbf{v}}} = \frac{4}{\beta}(\tau+1)$, $U_{\boldsymbol{\pi}\hat{\mathbf{q}}} = \frac{4}{\beta}(\tau+2)$, $U_q = 2\tau+3$, $U_{\bar{\mathbf{q}}} = 2\tau+4$, $e' = e-1$, $e'' = e-2$,

$$\begin{aligned}
c &= \eta \frac{e(U_{\hat{\mathbf{v}}} + L + \gamma V_{\hat{\mathbf{q}}})}{1-\gamma-\eta e N V_{\hat{\mathbf{q}}}}, \\
c' &= \eta \frac{e'(L+\gamma V_{\hat{\mathbf{q}}}) + (e' U_{\boldsymbol{\pi}\hat{\mathbf{q}}} + U_{\hat{\mathbf{v}}}) U_{\bar{\mathbf{q}}} |\mathcal{A}|}{1-\gamma-\eta e(N+1) V_{\hat{\mathbf{q}}}},
\end{aligned}$$

and assume that $\gamma \in (0,1)$, $c(\tau+1)^2 < \beta/2$, $N \ge 1 + \left\lceil \tau \ln\left(\frac{4}{\beta - 2c(\tau+1)^2}\right)\right\rceil$, $0 < \eta < \frac{\beta(1-\gamma)}{2e(N+1)(|\mathcal{A}|/\gamma + 2\tau + 2)}$. Then, for any policy π, we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\rho_t^\pi - \boldsymbol{\rho}_t\right] \le \frac{\ln|\mathcal{A}|}{\eta} &+ (N-1)(U_{\bar{\mathbf{q}}} + U_q + 1) \\
&+ (T - 2N + 2)\Big(c'(N-1)(1+\eta e'' V_{\hat{\mathbf{q}}}) \\
&+ \gamma U_q + \eta e'' |\mathcal{A}| U_{\boldsymbol{\pi}\hat{\mathbf{q}}} U_{\bar{\mathbf{q}}}\Big).
\end{aligned}$$

Proposition 2: Assume that the conditions of Proposition 1 hold. Then,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\boldsymbol{\rho}_t\right] - \widehat{R}_T \le N - 1 &+ (T - N + 1)\, c\, (\tau+1)^2 \\
&+ 2(T-N+1)e^{-(N-1)/\tau}.
\end{aligned} \tag{22}$$

Note that setting $N \ge 1 + \lceil \tau \ln T \rceil$, as suggested in Theorem 1, the last term in the right-hand side of (22) becomes $O(1)$, while for T sufficiently large all the conditions of the last two propositions will be satisfied. This leads to the proof of Theorem 1:

Proof of Theorem 1: If $|\mathcal{A}| = 1$ then, due to $\hat{L}_T = 0$, the statement is trivial, so we assume $|\mathcal{A}| \ge 2$ from now on. Define $\alpha = \frac{\beta}{2} e(U_{\hat{\mathbf{v}}} + L + \gamma V_{\hat{\mathbf{q}}})$, $\alpha' = \frac{\beta}{2}\{e'(L+\gamma V_{\hat{\mathbf{q}}}) + (e' U_{\boldsymbol{\pi}\hat{\mathbf{q}}} + U_{\hat{\mathbf{v}}}) U_{\bar{\mathbf{q}}} |\mathcal{A}|\}$ so that $c = 2\eta \frac{\alpha}{\beta(1-\gamma) - 2\eta e N V'_{\hat{\mathbf{q}}}}$ and $c' = 2\eta \frac{\alpha'}{\beta(1-\gamma) - 2\eta e(N+1) V'_{\hat{\mathbf{q}}}}$, where $V'_{\hat{\mathbf{q}}} = \frac{\beta}{2} V_{\hat{\mathbf{q}}} = |\mathcal{A}|/\gamma + 2\tau + 2$. In the following we will use the notation $f \sim g$ for two positive-valued functions $f, g : D \to \mathbb{R}^+$ defined on the same domain D to denote that they are equivalent up to a constant factor, that is, $\sup_{x \in D} \max\{f(x)/g(x), g(x)/f(x)\} < \infty$. With this notation, on $|\mathcal{A}| \ge 2, \tau \ge 1$ and as long as $\gamma \le 1$, we have

$$\alpha \sim |\mathcal{A}| + \tau \quad \text{and} \quad \alpha' \sim |\mathcal{A}| \tau^2 \tag{23}$$

independently of the value of β and of the choice of η, γ, N. In what follows all the equivalences will be stated for the domain $|\mathcal{A}| \ge 2, \tau \ge 1$.

We now show how to choose η, γ and N so as to achieve a small regret bound. In order to do so we will choose these constants so that the conditions of Propositions 1 and 2 are satisfied. For simplicity, we add the constraint $\gamma \le 1/2$ that we will also show to hold. Under this additional constraint, the inequality

$$\eta < \frac{\beta(1-\gamma)}{2e(N+1)(|\mathcal{A}|/\gamma + 2\tau + 2)} \tag{24}$$

will be satisfied if we choose $\gamma = 8e\eta(N+1)(|\mathcal{A}| + \tau + 1)/\beta$. Indeed, the said inequality holds since it is equivalent to $D = \beta(1-\gamma) - 2\eta e(N+1)(|\mathcal{A}|/\gamma + 2\tau + 2) > 0$ and

$$\begin{aligned}
D &= \beta(1-\gamma) - \frac{2\eta e(N+1)(|\mathcal{A}| + \gamma(2\tau+2))}{\gamma} \\
&\ge \frac{\beta}{2} - \frac{2\eta e(N+1)(|\mathcal{A}| + \tau + 1)}{\gamma} = \frac{\beta}{4} > 0,
\end{aligned}$$

where the first inequality holds because $\gamma \le 1/2$ and the second equality holds by the definition of γ. Since $c \le 2\eta\alpha/D$ and $c' =$

$2\eta\alpha'/D$, this also implies

$$c \le 8\eta\alpha/\beta \quad \text{and} \quad c' \le 8\eta\alpha'/\beta. \tag{25}$$

Due to this upper bound on c, $c(\tau+1)^2 < \beta/2$ will be satisfied if

$$\eta < \frac{\beta^2}{16\alpha(\tau+1)^2}. \tag{26}$$

To satisfy $\gamma \le 1/2$, the inequality

$$\eta \le \frac{\beta}{16e(N+1)(|\mathcal{A}|+\tau+1)} \tag{27}$$

has to be satisfied, too. Before proving (26) and (27), we derive the regret bound they imply.

Taking expectation in (21) and using the bounds of Lemma 1 and Propositions 1 and 2, we get

$$\begin{aligned}\widehat{L}_T &\le 2\tau + 2 + \frac{\ln|\mathcal{A}|}{\eta} \\ &+T\Big[c'(N-1)(1+\eta\, e''V_{\hat{\mathbf{q}}}) + \gamma\, U_q + \eta\, e''|\mathcal{A}|\, U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\, U_{\bar{\mathbf{q}}} + c(\tau+1)^2\Big] \\ &+(N-1)(U_{\bar{\mathbf{q}}} + U_q + 2) + 2Te^{-(N-1)/\tau}.\end{aligned}$$

Choosing $N = 1 + \lceil \tau \ln T \rceil$, we have $2Te^{-(N-1)/\tau} \le 2$. Furthermore, (24) implies $\eta\, e''V_{\hat{\mathbf{q}}} \le \frac{e''}{e(N+1)}$. This, together with the definition of the different constants above and the bound (25) on c and c' gives

$$\begin{aligned}\widehat{L}_T &\le 2\tau + 2 + \frac{\ln|\mathcal{A}|}{\eta} \\ &+ \frac{\eta T}{\beta}\Bigg[8\alpha'(N-1)\left(1+\frac{e''}{e(N+1)}\right) \\ &\qquad + 8e\beta(N+1)(|\mathcal{A}|+\tau+1)(2\tau+3) \\ &\qquad + 8e''|\mathcal{A}|\,(\tau+2)^2 + 8\alpha(\tau+1)^2\Bigg] \\ &+ (N-1)(U_{\bar{\mathbf{q}}} + U_q + 2) + 2 \\ &\le \frac{\ln|\mathcal{A}|}{\eta} + \frac{\eta T}{\beta}B + C_1\tau N,\end{aligned}$$

where we introduced B to denote the expression in the squared brackets and used the fact that $2\tau+4+(N-1)(U_{\bar{\mathbf{q}}}+U_q+2) \le C_1\tau N$ for some constant $C_1 > 0$. Note that (23) implies that

$$B \sim N|\mathcal{A}|\tau^2 + \tau^3 \sim |\mathcal{A}|\tau^3 \ln T \tag{28}$$

since $N \sim \tau \ln T$ for $\tau \ge 1, T \ge 2$. Now, choose $\eta = \sqrt{\frac{\beta \ln|\mathcal{A}|}{TB}}$. Then,

$$\begin{aligned}\widehat{L}_T &= 2\sqrt{\frac{TB\ln|\mathcal{A}|}{\beta}} + C_1\tau N \\ &\le C_2\sqrt{\frac{\tau^3 T|\mathcal{A}|\ln(|\mathcal{A}|)\ln(T)}{\beta}} + C_3\tau^2\ln T\end{aligned}$$

for some appropriate constants $C_2, C_3 > 0$.

It remains to show that for T large enough, inequalities (26) and (27) will hold, and also the lower bound on N in the propositions will be satisfied. Instead of (26) we will choose a lower bound on T to guarantee the stronger condition

$$\eta \le \frac{\beta^2}{32\alpha(\tau+1)^2}, \tag{29}$$

which, together with (25), also ensures $c(\tau+1)^2 \le \beta/4$. The latter inequality implies that the lower bound on N in the propositions is satisfied for $T \ge 8/\beta$. Using the choice of η and the respective equivalent forms (23) and (28) for α and B, one can see that condition (29) is satisfied if $T\ln T \ge C_4 \frac{(|\mathcal{A}|\tau+\tau^3/|\mathcal{A}|)\ln|\mathcal{A}|}{\beta^3}$ for some appropriate constant $C_4 > 0$. To keep things simple, notice that selecting $T \ge C_4 \frac{(|\mathcal{A}|\tau+\tau^3/|\mathcal{A}|)\ln|\mathcal{A}|}{\beta^3}$ implies (29), and also $T \ge 8/\beta$ if $C_4 \ge 8$.[13] Furthermore, one can similarly show that (27) is satisfied if $\frac{T}{\ln T} \ge C_5 \frac{(|\mathcal{A}|/\tau+\tau/|\mathcal{A}|)\ln|\mathcal{A}|}{\beta}$ for some appropriate constant $C_5 > 0$. By Proposition 3 of Antos et al. [1], for any $u > 0$, $t/\ln t > u$ if $t \ge h(u) \stackrel{\text{def}}{=} 2u\ln u$. Thus, the last condition on T is satisfied if $T > h\left(C_5 \frac{(|\mathcal{A}|/\tau+\tau/|\mathcal{A}|)\ln|\mathcal{A}|}{\beta}\right)$. This finishes the proof of the theorem. ∎

C. General tools for the proofs of Propositions 1 and 2

Just like in the previous section, throughout this section we suppose that both Assumptions A1 and A2 hold for P and the rewards are in the $[0,1]$ interval. We proceed with a series of lemmas to bound the rate of change of the policies generated by MDP-EXP3.

Lemma 6: Let $1 \le t \le T$ and assume that $\boldsymbol{\mu}_t^N(x) \ge \beta/2$ holds for all states x. Then, for any $(x,a) \in \mathcal{X}\times\mathcal{A}$, we have $|\hat{\mathbf{v}}_t(x)| \le \frac{4\tau+4}{\beta}$ and $-\frac{2}{\beta}(2\tau+3) \le \hat{\mathbf{q}}_t(x,a) \le \frac{2}{\beta}\left(\frac{|\mathcal{A}|}{\gamma} + 2\tau + 2\right)$.

Proof: Since $|\sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{r}}_t(x,a)| = \boldsymbol{\pi}_t(\mathbf{a}_t|x)\hat{\mathbf{r}}_t(x,\mathbf{a}_t)\mathbb{I}_{\{x=\mathbf{x}_t\}} \le 1/\boldsymbol{\mu}_t^N(x) \le 2/\beta$ by assumption and $\hat{\mathbf{v}}_t = v^{\boldsymbol{\pi}_t}$, the first statement of the lemma follows from Lemma 3.

To prove the bounds on $\hat{\mathbf{q}}_t$, notice that

$$\begin{aligned}0 \le \hat{\boldsymbol{\rho}}_t &= \sum_{x,a} \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x)\, \boldsymbol{\pi}_t(a|x)\hat{\mathbf{r}}_t(x,a) \\ &= \sum_x \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x) \sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{r}}_t(x,a) \le \frac{2}{\beta}.\end{aligned}$$

Applying the above inequalities to the Bellman equations (5), we obtain $\hat{\mathbf{q}}_t(x,a) = \hat{\mathbf{r}}_t(x,a) - \hat{\boldsymbol{\rho}}_t + \sum_{x'} P(x'|x,a)\hat{\mathbf{v}}_t(x') \ge -\frac{2}{\beta} - \frac{4\tau+4}{\beta} = -\frac{2}{\beta}(2\tau+3)$. Since $\boldsymbol{\pi}_t(a|x) \ge \gamma/|\mathcal{A}|$, the assumption on $\boldsymbol{\mu}_t^N$ and the definition of $\hat{\mathbf{r}}_t$ imply $\hat{\mathbf{r}}_t(x,a) \le \frac{2|\mathcal{A}|}{\gamma\beta}$. Thus, we get the upper bound $\hat{\mathbf{q}}_t(x,a) \le \frac{2|\mathcal{A}|}{\gamma\beta} + \frac{4\tau+4}{\beta} = \frac{2}{\beta}\left(\frac{|\mathcal{A}|}{\gamma} + 2\tau + 2\right)$. ∎

The previous result can be strengthened if one is interested in a bound on $\mathbb{E}\left[|\hat{\mathbf{v}}_t(x)|\,|\,\mathbf{u}_{t-N}\right]$:

Lemma 7: Let $1 \le t \le T$ and assume that $\boldsymbol{\mu}_t^N(x) > 0$ holds for all states x. Then, for any $x \in \mathcal{X}$, we have $\mathbb{E}\left[|\hat{\mathbf{v}}_t(x)|\,|\,\mathbf{u}_{t-N}\right] \le 2(\tau+1)$.

Proof: Proceeding as in the proof of Lemma 3 and then taking expectations, we get

$$\begin{aligned}&\mathbb{E}\left[|\hat{\mathbf{v}}_t(x)|\,|\,\mathbf{u}_{t-N}\right] \\ &\quad\le \sum_{s=1}^{\infty}\sum_{x'} |\nu_{s,x}^{\pi_t}(x') - \mu_{\mathrm{st}}^{\pi_t}(x')|\; \mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x')\hat{\mathbf{r}}_t(x',a)\,\middle|\,\mathbf{u}_{t-N}\right],\end{aligned}$$

where we have exploited that $\hat{\mathbf{r}}_t$ is well-defined by our assumption on $\boldsymbol{\mu}_t^N$ and it takes only nonnegative values. Now, by (9) and (11),

$$\begin{aligned}\mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x')\hat{\mathbf{r}}_t(x',a)\,\middle|\,\mathbf{u}_{t-N}\right] &= \sum_a \boldsymbol{\pi}_t(a|x')\mathbb{E}\left[\hat{\mathbf{r}}_t(x',a)\big|\,\mathbf{u}_{t-N}\right] \\ &= \sum_a \boldsymbol{\pi}_t(a|x')r_t(x',a),\end{aligned}$$

which is bounded between 0 and 1. Hence, $\mathbb{E}\left[|\hat{\mathbf{v}}_t(x)|\,|\,\mathbf{u}_{t-N}\right] \le \sum_{s=1}^{\infty}\sum_{x'}|\nu_{s,x}^{\pi}(x') - \mu_{\mathrm{st}}^{\pi}(x')|$. Finishing as in the proof of Lemma 1 or 3, we get the statement. ∎

Similarly, we will also need a bound on the expected value of $\mathbb{E}\left[|\hat{\mathbf{q}}_t(x,a)|\,|\,\mathbf{u}_{t-N}\right]$:

[13] With some extra work one can show that it is sufficient to choose $T \ge h_1\left(C_4\frac{\tau^3|\mathcal{A}|\ln|\mathcal{A}|}{\beta^3}\right)$ with $C_4 \ge 64$ and $h_1(y) = y/(\ln y - \ln\ln y)$.

Lemma 8: Let $1 \le t \le T$ and assume that $\boldsymbol{\mu}_t^N(x) > 0$ holds for all states x. Then, for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have that $\mathbb{E}\left[\,|\hat{\mathbf{q}}_t(x,a)|\,|\,\mathbf{u}_{t-N}\right] \le 2\tau+4$ and also $\mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\right] \le 2\tau+4$.

Proof: By the Bellman equations (5),

$$\mathbb{E}\left[\,|\hat{\mathbf{q}}_t(x,a)|\,|\,\mathbf{u}_{t-N}\right] \le \mathbb{E}\left[\,|\hat{\mathbf{r}}_t(x,a)|\,|\,\mathbf{u}_{t-N}\right] + \mathbb{E}\left[\,|\hat{\boldsymbol{\rho}}_t|\,|\,\mathbf{u}_{t-N}\right] + \sum_{x'} P(x'|x,a)\mathbb{E}\left[\,|\hat{\mathbf{v}}_t(x')|\,\middle|\,\mathbf{u}_{t-N}\right].$$

As before, $\mathbb{E}\left[\,|\hat{\mathbf{r}}_t(x,a)|\,|\,\mathbf{u}_{t-N}\right] \le 1$, and also $\mathbb{E}\left[\,|\hat{\boldsymbol{\rho}}_t|\,|\,\mathbf{u}_{t-N}\right] \le 1$. Combining these with the result of the previous lemma, we get the first part of the statement. To get the second part note that

$$\begin{aligned}\mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\,\middle|\,\mathbf{u}_{t-N}\right]\right]\\ &= \mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x)\,\mathbb{E}\left[\,|\hat{\mathbf{q}}_t(x,a)|\,|\,\mathbf{u}_{t-N}\right]\right]\\ &\le \mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x)\,(2\tau+4)\right] \le 2\tau+4\,.\end{aligned}$$

■

The quantity $\boldsymbol{\pi}_t(x,a)|\hat{\mathbf{q}}_t(x,a)|$ also enjoys a bound which is independent of the exploration rate γ:

Lemma 9: Let $1 \le t \le T$ and assume that $\boldsymbol{\mu}_t^N(x) \ge \beta/2$ holds for all states x. Then, for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, it holds that $\boldsymbol{\pi}_t(x,a)\,|\hat{\mathbf{q}}_t(x,a)| \le \frac{4}{\beta}(\tau+2)$.

Proof: By our assumption on $\boldsymbol{\mu}_t^N$ and the construction of $\hat{\mathbf{r}}_t(x,a)$,

$$\boldsymbol{\pi}_t(x,a)|\hat{\mathbf{r}}_t(x,a)| \le \frac{2}{\beta}. \tag{30}$$

Since $\hat{\mathbf{r}}_t(x,a) = 0$ unless $a = \mathbf{a}_t$, Lemma 3 can be applied with $R = 2/\beta$ to obtain $|\hat{\mathbf{q}}_t(x,a)| \le \frac{2}{\beta}(2\tau+3) + |\hat{\mathbf{r}}_t(x,a)|$. Multiplying both sides by $\boldsymbol{\pi}_t(x,a)$ and using (30) again finishes the proof. ■

Now we show that if the policies that we follow up to time step t change slowly, $\boldsymbol{\mu}_t^N$ is "close" to $\mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}$:

Lemma 10: Let $1 \le N \le t \le T$ and $c > 0$ be such that $\|\boldsymbol{\pi}_{s+1} - \boldsymbol{\pi}_s\|_{1,\infty} \le c$ holds for $1 \le s \le t-1$. Then we have $\left\|\boldsymbol{\mu}_t^N - \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}\right\|_1 \le c\,(\tau+1)^2 + 2e^{-(N-1)/\tau}$.

Proof: This follows directly from Lemma 5 since, thanks to the recursive form of $\boldsymbol{\mu}_t^N$, $\boldsymbol{\mu}_t^N = \boldsymbol{\mu}_1 P^{\boldsymbol{\pi}_{t-N+1}} \cdots P^{\boldsymbol{\pi}_{t-1}}$, where $\boldsymbol{\mu}_1 = e_{\mathbf{x}_{t-N}} P^{\mathbf{a}_{t-N}}$ for $t \ge N+1$ and $\boldsymbol{\mu}_1 = P_1$ if $t = N$. ■

In the lemma that follows we compute the rate of change of the policies produced by MDP-Exp3. We will use this lemma for multiple purposes, including showing that for a large enough value of N, $\boldsymbol{\mu}_t^N$ can be uniformly bounded from below by $\beta/2$.

To simplify the presentation, we recall some short-hand notation from Proposition 1. In particular, we denote the lower and upper bounds for $\hat{\mathbf{q}}_t$ by $-L$ and $V_{\hat{\mathbf{q}}}$ of Lemma 6, respectively, and the upper bound on $|\hat{\mathbf{v}}_t|$ from the same lemma by $U_{\hat{\mathbf{v}}}$. Thus setting

$$L = \frac{2}{\beta}(2\tau+3), \quad V_{\hat{\mathbf{q}}} = \frac{2}{\beta}\left(\frac{|\mathcal{A}|}{\gamma} + 2\tau + 2\right), \quad U_{\hat{\mathbf{v}}} = \frac{4}{\beta}(\tau+1)$$

we have $-L \le \hat{\mathbf{q}}_t(x,a) \le V_{\hat{\mathbf{q}}}$, and $|\hat{\mathbf{v}}_t(x)| \le U_{\hat{\mathbf{v}}}$ for all state-action pairs (x,a).

Lemma 11: Assume that $0 < \eta \le 1/(V_{\hat{\mathbf{q}}}+L)$. For $(x,a) \in \mathcal{X}\times\mathcal{A}$, $1 \le t \le T$, $i \in \{0,1\}$ let

$$\boldsymbol{d}_{t,i}(x,a) = \hat{\mathbf{q}}_t(x,a) - \sum_{b\in\mathcal{A}} \frac{\boldsymbol{\pi}_{t+N-1+i}(b|x) - \frac{\gamma}{|\mathcal{A}|}}{1-\gamma}\,\hat{\mathbf{q}}_t(x,b). \tag{31}$$

Then, for all $N \le t \le T$,

$$\begin{aligned}&|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)|\\ &\quad\le \eta\boldsymbol{\pi}_{t+N-1}(a|x)\max\left\{(e-1)\boldsymbol{d}_{t,0}(x,a), -\boldsymbol{d}_{t,1}(x,a)\right\}.\end{aligned}$$

Proof: Fix some state-action pair $(x,a) \in \mathcal{X}\times\mathcal{A}$ and let $\mathbf{W}_t(x) = \sum_a \mathbf{w}_t(x,a)$ where $t = 1,2,\ldots,T$. Since $\mathbf{w}_{t+N}$ is computed using the exponential weight update for $t \ge N$, we have

$$\begin{aligned}&|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)|\\ &\quad= (1-\gamma)\left|\frac{\mathbf{w}_{t+N-1}(x,a)}{\mathbf{W}_{t+N-1}(x)} - \frac{\mathbf{w}_{t+N}(x,a)}{\mathbf{W}_{t+N}(x)}\right|\\ &\quad= (1-\gamma)\frac{\mathbf{w}_{t+N-1}(x,a)}{\mathbf{W}_{t+N-1}(x)}\left|1 - \frac{\mathbf{w}_{t+N}(x,a)}{\mathbf{w}_{t+N-1}(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\right|\\ &\quad= (1-\gamma)\frac{\mathbf{w}_{t+N-1}(x,a)}{\mathbf{W}_{t+N-1}(x)}\left|1 - e^{\eta\hat{\mathbf{q}}_t(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\right|\\ &\quad\le \boldsymbol{\pi}_{t+N-1}(a|x)\left|1 - e^{\eta\hat{\mathbf{q}}_t(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\right|.\end{aligned} \tag{32}$$

We examine two separate cases depending on the sign of the expression in the absolute value on the right-hand side.

Case a) $1 - e^{\eta\hat{\mathbf{q}}_t(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)} \le 0$: First notice that the logarithm of the second term is positive by the condition, that is, $\eta\hat{\mathbf{q}}_t(x,a) + \ln\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)} \ge 0$. Furthermore, it is bounded from above by 1. Indeed, by Jensen's inequality,

$$\begin{aligned}\ln\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)} &= -\ln\sum_b \frac{\mathbf{w}_{t+N-1}(x,b)}{\mathbf{W}_{t+N-1}(x)}e^{\eta\hat{\mathbf{q}}_t(x,b)}\\ &\le -\eta\sum_b \frac{\mathbf{w}_{t+N-1}(x,b)}{\mathbf{W}_{t+N-1}(x)}\,\hat{\mathbf{q}}_t(x,b)\end{aligned} \tag{33}$$

and thus

$$\begin{aligned}&\eta\hat{\mathbf{q}}_t(x,a) + \ln\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\\ &\quad\le \eta\hat{\mathbf{q}}_t(x,a) - \eta\sum_b \frac{\mathbf{w}_{t+N-1}(x,b)}{\mathbf{W}_{t+N-1}(x)}\hat{\mathbf{q}}_t(x,b)\\ &\quad\le \eta(V_{\hat{\mathbf{q}}} + L) \le 1\,,\end{aligned}$$

where the second inequality holds by our choice of $V_{\hat{\mathbf{q}}}$ and L, while the third one holds by our assumption on η. Thus, using $e^z - 1 \le (e-1)z$, which holds for any $0 \le z \le 1$, we get

$$\begin{aligned}&\left|1 - e^{\eta\hat{\mathbf{q}}_t(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\right|\\ &\quad\le (e-1)\left(\eta\hat{\mathbf{q}}_t(x,a) + \ln\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\right).\end{aligned}$$

From this inequality, (32) and (33) and using the definition of $\boldsymbol{\pi}_{t+N-1}$, we get

$$\begin{aligned}&|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)|\\ &\quad\le \eta(e-1)\,\boldsymbol{\pi}_{t+N-1}(a|x)\left(\hat{\mathbf{q}}_t(x,a) - \sum_b \frac{\boldsymbol{\pi}_{t+N-1}(b|x) - \frac{\gamma}{|\mathcal{A}|}}{1-\gamma}\hat{\mathbf{q}}_t(x,b)\right).\end{aligned}$$

Case b) $1 - e^{\eta\hat{\mathbf{q}}_t(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)} \ge 0$: Using $1 - e^z \le -z$ (which holds for all $z \in \mathbb{R}$), we get

$$\left|1 - e^{\eta\hat{\mathbf{q}}_t(x,a)}\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}\right| \le -\eta\hat{\mathbf{q}}_t(x,a) - \ln\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)}.$$

Applying Jensen's inequality, the second term can be bounded as

$$\begin{aligned}\ln\frac{\mathbf{W}_{t+N-1}(x)}{\mathbf{W}_{t+N}(x)} &= \ln\sum_b \frac{\mathbf{w}_{t+N}(x,b)}{\mathbf{W}_{t+N}(x)}e^{-\eta\hat{\mathbf{q}}_t(x,b)}\\ &\ge -\eta\sum_b \frac{\boldsymbol{\pi}_{t+N}(b|x) - \frac{\gamma}{|\mathcal{A}|}}{1-\gamma}\,\hat{\mathbf{q}}_t(x,b)\,.\end{aligned}$$

Combining these inequalities with (32), we get

$$|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)| \le \eta\, \boldsymbol{\pi}_{t+N-1}(a|x) \left(-\hat{\mathbf{q}}_t(x,a) + \sum_b \frac{\boldsymbol{\pi}_{t+N}(b|x) - \frac{\gamma}{|\mathcal{A}|}}{1-\gamma} \hat{\mathbf{q}}_t(x,b) \right).$$

The two cases together prove the lemma. ∎

By the lemma just proved, the rate of change $\{\boldsymbol{\pi}_t\}$ is partially governed by $\{\boldsymbol{d}_{t,i}\}$. To further bound the rate of change, we develop lower and upper bounds on $\boldsymbol{d}_{t,i}$. To facilitate this, we rewrite $\boldsymbol{d}_{t,i}$ by grouping the terms with identical signs:

$$\begin{aligned} \boldsymbol{d}_{t,i}(x,a) = \hat{\mathbf{q}}_t(x,a) + \frac{\gamma}{1-\gamma} \frac{1}{|\mathcal{A}|} \sum_{b\in\mathcal{A}} \hat{\mathbf{q}}_t(x,b) \\ - \frac{1}{1-\gamma} \sum_{b\in\mathcal{A}} \boldsymbol{\pi}_{t+N-1+i}(b|x) \hat{\mathbf{q}}_t(x,b). \end{aligned} \tag{34}$$

Then, from $-L \le \hat{\mathbf{q}}_t(x,a) \le V_{\hat{\mathbf{q}}}$, as long as $0 \le \gamma < 1$, we have

$$\begin{aligned} -\frac{L + \sum_b \boldsymbol{\pi}_{t+N-1+i}(b|x) \hat{\mathbf{q}}_t(x,b)}{1-\gamma} \le \boldsymbol{d}_{t,i}(x,a) \\ \le \hat{\mathbf{q}}_t(x,a) + \frac{L + \gamma V_{\hat{\mathbf{q}}}}{1-\gamma}. \end{aligned} \tag{35}$$

Notice that since $V_{\hat{\mathbf{q}}}$ scales with $1/\gamma$, we avoided upper bounding $\hat{\mathbf{q}}_t$ by $V_{\hat{\mathbf{q}}}$ except when $\hat{\mathbf{q}}_t$ is multiplied by γ, and so the bounds will not "blow up" as $\gamma \to 0$. In fact, this is one of the main reasons that in this paper we succeed in proving an $\tilde{O}(T^{1/2})$ regret bound as compared to the $O(T^{2/3})$ regret bound of [16].

Let us now show that, provided $\boldsymbol{\mu}_t^N$ is uniformly bounded away from zero, the sequence $\{\boldsymbol{\pi}_t\}$ changes slowly.

Lemma 12: Assume that $0 \le \gamma < 1$, $0 < \eta < \min\left(\frac{1}{V_{\hat{\mathbf{q}}}+L}, \frac{1-\gamma}{eNV_{\hat{\mathbf{q}}}}\right) = \frac{\beta(1-\gamma)}{2eN(|\mathcal{A}|/\gamma+2\tau+2)}$ and that for all $N \le t \le T$ and states x, $\boldsymbol{\mu}_t^N(x) \ge \beta/2$ holds true. Set $c = \eta\, e \frac{U_{\hat{\mathbf{v}}}+L+\gamma V_{\hat{\mathbf{q}}}}{1-\gamma-\eta e N V_{\hat{\mathbf{q}}}}$. Then, for any $1 \le t \le T$,

$$\|\boldsymbol{\pi}_{t+N-1} - \boldsymbol{\pi}_{t+N}\|_{1,\infty} \le c. \tag{36}$$

Proof: We prove the statement by induction on t. To show the bound for time step t assume that $\|\boldsymbol{\pi}_{s+N-1} - \boldsymbol{\pi}_{s+N}\|_{1,\infty} \le c$ holds for all $s = 1, 2, \ldots, t-1$. As $\boldsymbol{\pi}_{s+N-1} = \boldsymbol{\pi}_{s+N}$ for all $s = 1, 2, \ldots, N-1$, the assumption holds for $t = 1, \ldots, N-1$ and we are left with proving the induction step for $t \ge N$.

Fix $x \in \mathcal{X}$. For any $a \in \mathcal{A}$, by Lemma 11,

$$|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)| \le \eta \boldsymbol{\pi}_{t+N-1}(a|x) \max\left\{(e-1)\boldsymbol{d}_{t,0}(x,a), -\boldsymbol{d}_{t,1}(x,a)\right\}.$$

Our goal is to upper-bound $\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,0}(x,a)$ and lower-bound $\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,1}(x,a)$. As before, we make an effort to avoid terms that scale with $1/\gamma$, but we allow terms that scale with c/γ as c will be seen to scale with γ (and η).

Consider first an upper bound on $\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,0}(x,a)$. From (35), it remains to bound $\boldsymbol{\pi}_{t+N-1}(a|x)\hat{\mathbf{q}}_t(x,a)$. By a simple telescoping argument, we bound this by

$$\begin{aligned} &\boldsymbol{\pi}_{t+N-1}(a|x)\hat{\mathbf{q}}_t(x,a) \\ &= \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_t(x,a) + \left[\sum_{s=t}^{t+N-2} (\boldsymbol{\pi}_{s+1}(a|x) - \boldsymbol{\pi}_s(a|x))\right] \hat{\mathbf{q}}_t(x,a) \\ &\le \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_t(x,a) + \boldsymbol{F}_t(x,a) V_{\hat{\mathbf{q}}}, \end{aligned}$$

where we have introduced

$$\boldsymbol{F}_t(x,a) = \sum_{s=t}^{t+N-2} |\boldsymbol{\pi}_{s+1}(a|x) - \boldsymbol{\pi}_s(a|x)|.$$

Now, using (34),

$$\begin{aligned} &\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,0}(x,a) \\ &\le \frac{1}{1-\gamma} \left(\boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_t(x,a) + \boldsymbol{F}_t(x,a) V_{\hat{\mathbf{q}}} + \boldsymbol{\pi}_{t+N-1}(a|x)(L + \gamma V_{\hat{\mathbf{q}}}) \right). \end{aligned}$$

Now, let us consider upper bounding $-\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,1}(x,a)$. In this case, we use telescoping for the second term on the left-hand side of (35):

$$\begin{aligned} &\sum_b \boldsymbol{\pi}_{t+N}(b|x)\hat{\mathbf{q}}_t(x,b) \\ &= \sum_b \boldsymbol{\pi}_t(b|x)\hat{\mathbf{q}}_t(x,b) + \sum_{s=t}^{t+N-1} \sum_b \Big(\boldsymbol{\pi}_{s+1}(b|x) - \boldsymbol{\pi}_s(b|x)\Big)\hat{\mathbf{q}}_t(x,b) \\ &\le \hat{\mathbf{v}}_t(x) + NcV_{\hat{\mathbf{q}}} \le U_{\hat{\mathbf{v}}} + NcV_{\hat{\mathbf{q}}}, \end{aligned}$$

where we used $\hat{\mathbf{q}}_t(x,a) \le V_{\hat{\mathbf{q}}}$, $\hat{\mathbf{v}}_t(x) \le U_{\hat{\mathbf{v}}}$ and the induction hypothesis. Plugging this into the lower bound in (35), we get

$$-\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,1}(x,a) \le \frac{\boldsymbol{\pi}_{t+N-1}(a|x)}{1-\gamma} \left(L + U_{\hat{\mathbf{v}}} + NcV_{\hat{\mathbf{q}}}\right).$$

Combining the two cases, we get

$$\begin{aligned} &|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)| \\ &\le \max\Bigg(\frac{\eta(e-1)}{1-\gamma} \Big(\boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_t(x,a) + \boldsymbol{F}_t(x,a)V_{\hat{\mathbf{q}}} \\ &\qquad + \boldsymbol{\pi}_{t+N-1}(a|x)(L + \gamma V_{\hat{\mathbf{q}}}) \Big), \\ &\qquad \frac{\eta \boldsymbol{\pi}_{t+N-1}(a|x)}{1-\gamma} \left(U_{\hat{\mathbf{v}}} + NcV_{\hat{\mathbf{q}}} + L\right) \Bigg) \\ &\le \frac{\eta}{1-\gamma} \Big((e-1)(\boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_t(x,a) + \boldsymbol{F}_t(x,a)V_{\hat{\mathbf{q}}}) \\ &\qquad + \boldsymbol{\pi}_{t+N-1}(a|x)(U_{\hat{\mathbf{v}}} + NcV_{\hat{\mathbf{q}}}) \\ &\qquad + (e-1)\,\boldsymbol{\pi}_{t+N-1}(a|x)(L + \gamma V_{\hat{\mathbf{q}}}) \Big). \end{aligned}$$

Summing these inequalities for all a and taking the maximum over x gives

$$\|\,\boldsymbol{\pi}_{t+N-1} - \boldsymbol{\pi}_{t+N}\,\|_{1,\infty} \le \frac{\eta\, e}{1-\gamma} \Big(U_{\hat{\mathbf{v}}} + NcV_{\hat{\mathbf{q}}} + L + \gamma V_{\hat{\mathbf{q}}} \Big),$$

where we upper bounded $(e-1)(L+\gamma V_{\hat{\mathbf{q}}})$ by $e(L+\gamma V_{\hat{\mathbf{q}}})$ and used that the inequality $\sum_a \boldsymbol{F}_t(x,a) \le (N-1)c \le Nc$ holds by the induction hypothesis. Now, the result follows because, thanks to the definition of c, the right-hand side equals c (in fact, this is how the definition of c is obtained). ∎

Lemma 13: Let c, η, γ be as in Lemma 12. Assume further that $c(\tau+1)^2 < \beta/2$, and let

$$N \ge 1 + \left\lceil \tau \ln\left(\frac{4}{\beta - 2c(\tau+1)^2}\right) \right\rceil. \tag{37}$$

Then, for all $N \le t \le T$, $x \in \mathcal{X}$, we have $\boldsymbol{\mu}_t^N(x) \ge \beta/2$ and $\|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\|_{1,\infty} \le c$.

Proof: We prove the lemma by induction on t. The induction hypothesis is that for $N \le t \le T$, $\min_x \boldsymbol{\mu}_s^N(x) \ge \beta/2$ and $\max_x \sum_a |\boldsymbol{\pi}_{s+1}(a|x) - \boldsymbol{\pi}_s(a|x)| \le c$ hold for all $N \le s \le t$.

Let us first show that this hypothesis holds when $N \le t \le 2N-2$. By the construction of the policies, we have $\max_x \sum_a |\boldsymbol{\pi}_{t+1}(a|x) - \boldsymbol{\pi}_t(a|x)| = 0 \le c$ for all $1 \le t \le 2N-2$. Thus, by Lemma 10, we get that $\left\|\boldsymbol{\mu}_t^N - \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}\right\|_1 \le c(\tau+1)^2 + 2e^{-(N-1)/\tau}$ holds for all $N \le t \le 2N-2$. By our assumption about N, we have

$$c(\tau+1)^2 + 2e^{-(N-1)/\tau} \le \beta/2, \tag{38}$$

thus for any $N \le t \le 2N-2$,

$$\left\|\boldsymbol{\mu}_t^N - \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}\right\|_\infty \le \left\|\boldsymbol{\mu}_t^N - \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}\right\|_1 \le \beta/2. \tag{39}$$

Since, by assumption, $\mu_{\mathrm{st}}^{\pi}(x) \ge \beta$ holds for any stationary policy π, we also have $\mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x) \ge \beta$ ($x \in \mathcal{X}$). This, together with (39) gives that $\boldsymbol{\mu}_t^N(x) \ge \beta/2$ holds for any $x \in \mathcal{X}$.

Now, fix a time index $2N-1 \le t \le T$ and assume that the induction hypothesis holds for time $t-1$. Then, thanks to $\min_x \boldsymbol{\mu}_{t-N+1}^N(x) \ge \beta/2$, Lemma 12 implies $\|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\|_{1,\infty} \le c$. Now, by Lemma 10, we have $\left\|\boldsymbol{\mu}_t^N - \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}\right\|_1 \le c(\tau+1)^2 + 2e^{-(N-1)/\tau}$ Using the same reasoning as above, we finish the inductive step and thus the proof. ∎

In our final result we study the weighted sums

$$\boldsymbol{\Delta}_{t,s}(x) \stackrel{\mathrm{def}}{=} \sum_{a\in\mathcal{A}} |\boldsymbol{\pi}_{t+1}(a|x) - \boldsymbol{\pi}_t(a|x)|\,|\hat{\mathbf{q}}_s(x,a)|$$

for $x \in \mathcal{X}$, $t, s \ge 1$. To state this result recall the definitions

$$U_{\boldsymbol{\pi}\hat{\mathbf{q}}} = \frac{4}{\beta}(\tau+2), \qquad U_{\bar{\mathbf{q}}} = 2\tau + 4,$$

from Proposition 1. Note that by Lemma 9, $\boldsymbol{\pi}_t(x,a)|\hat{\mathbf{q}}_t(x,a)| \le U_{\boldsymbol{\pi}\hat{\mathbf{q}}}$ and by Lemma 8, $\mathbb{E}\left[|\hat{\mathbf{q}}_s(x,a)|\right] \le U_{\bar{\mathbf{q}}}$.

Lemma 14: Let $e' = e-1$, $C_1 = e'(L+\gamma V_{\hat{\mathbf{q}}}) + (e'U_{\boldsymbol{\pi}\hat{\mathbf{q}}} + U_{\hat{\mathbf{v}}})U_{\bar{\mathbf{q}}}|\mathcal{A}|$, and $c' = \frac{\eta C_1}{1-\gamma-\eta e(N+1)V_{\hat{\mathbf{q}}}}$. Assume that $0 < \gamma \le 1$, $0 \le \eta < \min\left(\frac{1}{V_{\hat{\mathbf{q}}}+L}, \frac{1-\gamma}{e(N+1)V_{\hat{\mathbf{q}}}}\right) = \frac{\beta(1-\gamma)}{2e(N+1)(|\mathcal{A}|/\gamma+2\tau+2)}$ and that for all $N \le t \le T$ and states x, $\boldsymbol{\mu}_t^N(x) \ge \beta/2$ holds true. Then, for any $1 \le t \le T$, $1 \le s \le T$ and $x \in \mathcal{X}$, it holds that $\mathbb{E}\left[\boldsymbol{\Delta}_{t,s}(x)\right] \le c'$. An observation that will be needed later is that the conditions of this lemma on η, γ and N imply those of Lemma 12.

Proof: Fix $x \in \mathcal{X}$. We will prove the result by induction. Since in the algorithm the weights are kept fixed for $1 \le t \le 2N-1$, $\boldsymbol{\Delta}_{t,s} = 0$ when $1 \le t \le 2N-2$. This establishes the base case of the induction. Thus, fix $t \ge N$ and assume that $\mathbb{E}\left[\boldsymbol{\Delta}_{t',s}(x)\right] \le c'$ holds for all pairs (t', s) such that $1 \le t' \le t+N-2$, $1 \le s \le T$.

By Lemma 11,

$$\begin{aligned}&\boldsymbol{\Delta}_{t+N-1,s}(x)\\ &\le \eta \sum_a \boldsymbol{\pi}_{t+N-1}(a|x)\max\left\{e'\boldsymbol{d}_{t,0}(x,a), -\boldsymbol{d}_{t,1}(x,a)\right\}|\hat{\mathbf{q}}_s(x,a)|.\end{aligned} \tag{40}$$

As in the proof of Lemma 12, we upper bound the two terms resulting from the maximum on the right-hand side of the above expression separately. Considering the first of these, using the upper bound from (35), we get

$$\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,0}(x,a) \le \boldsymbol{\pi}_{t+N-1}(a|x)\frac{\hat{\mathbf{q}}_t(x,a) + L + \gamma V_{\hat{\mathbf{q}}}}{1-\gamma}.$$

We use telescoping to bound the first term in the numerator on the right-hand side:

$$\begin{aligned}\boldsymbol{\pi}_{t+N-1}(a|x)\hat{\mathbf{q}}_t(x,a) \le\ & \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\\ &+ \sum_{t'=t}^{t+N-2} |\boldsymbol{\pi}_{t'+1}(a|x) - \boldsymbol{\pi}_{t'}(a|x)|\,|\hat{\mathbf{q}}_t(x,a)|.\end{aligned}$$

Hence,

$$\begin{aligned}&\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,0}(x,a) \le\\ &\quad\frac{1}{1-\gamma}\Bigg(\boldsymbol{\pi}_{t+N-1}(a|x)(L+\gamma V_{\hat{\mathbf{q}}}) + \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\\ &\qquad + \sum_{t'=t}^{t+N-1} |\boldsymbol{\pi}_{t'+1}(a|x) - \boldsymbol{\pi}_{t'}(a|x)|\,|\hat{\mathbf{q}}_t(x,a)|\Bigg).\end{aligned}$$

Now, considering the second branch of the maximum, using this time the lower bound from (35),

$$\begin{aligned}&-\boldsymbol{\pi}_{t+N-1}(a|x)\boldsymbol{d}_{t,1}(x,a)\\ &\le \boldsymbol{\pi}_{t+N-1}(a|x)\frac{L + \sum_b \boldsymbol{\pi}_{t+N}(b|x)\hat{\mathbf{q}}_t(x,b)}{1-\gamma}\\ &\le \frac{\boldsymbol{\pi}_{t+N-1}(a|x)}{1-\gamma}\Bigg(L + \left|\sum_b \boldsymbol{\pi}_t(b|x)\hat{\mathbf{q}}_t(x,b)\right|\\ &\qquad + \sum_b \sum_{t'=t}^{t+N-1} |\boldsymbol{\pi}_{t'+1}(b|x) - \boldsymbol{\pi}_{t'}(b|x)|\,|\hat{\mathbf{q}}_t(x,b)|\Bigg).\end{aligned}$$

Combining these two inequalities, introducing $\hat{C} = e'\frac{L+\gamma V_{\hat{\mathbf{q}}}}{1-\gamma}$, we get

$$\begin{aligned}&\boldsymbol{\pi}_{t+N-1}(a|x)\max\left\{e'\boldsymbol{d}_{t,0}(x,a), -\boldsymbol{d}_{t,1}(x,a)\right\}\\ &\le \boldsymbol{\pi}_{t+N-1}(a|x)\hat{C} + \frac{e'}{1-\gamma}\boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\\ &\quad + \frac{e'}{1-\gamma}\sum_{t'=t}^{t+N-2} |\boldsymbol{\pi}_{t'+1}(a|x) - \boldsymbol{\pi}_{t'}(a|x)|\,|\hat{\mathbf{q}}_t(x,a)|\\ &\quad + \frac{\boldsymbol{\pi}_{t+N-1}(a|x)}{1-\gamma}\Bigg[\left|\sum_b \boldsymbol{\pi}_t(b|x)\hat{\mathbf{q}}_t(x,b)\right|\\ &\qquad + \sum_b \sum_{t'=t}^{t+N-1} |\boldsymbol{\pi}_{t'+1}(b|x) - \boldsymbol{\pi}_{t'}(b|x)|\,|\hat{\mathbf{q}}_t(x,a)|\Bigg].\end{aligned}$$

Plugging this into (40), we get

$$\begin{aligned}&\boldsymbol{\Delta}_{t+N-1,s}(x) \le \eta\hat{C}\\ &\quad + \frac{\eta e'}{1-\gamma}\sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\,|\hat{\mathbf{q}}_s(x,a)|\\ &\quad + \frac{\eta e'}{1-\gamma}\sum_{t'=t}^{t+N-2}\sum_a |\boldsymbol{\pi}_{t'+1}(a|x) - \boldsymbol{\pi}_{t'}(a|x)|\,|\hat{\mathbf{q}}_t(x,a)|\,|\hat{\mathbf{q}}_s(x,a)|\\ &\quad + \eta\sum_a |\hat{\mathbf{q}}_s(x,a)|\frac{\boldsymbol{\pi}_{t+N-1}(a|x)}{1-\gamma}\left[\left|\sum_b \boldsymbol{\pi}_t(b|x)\hat{\mathbf{q}}_t(x,b)\right| + \sum_{t'=t}^{t+N-1}\boldsymbol{\Delta}_{t',t}(x)\right].\end{aligned}$$

Using $\boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)| \le U_{\boldsymbol{\pi}\hat{\mathbf{q}}}$, $|\hat{\mathbf{v}}_t(x)| \le U_{\hat{\mathbf{v}}}$ and $|\hat{\mathbf{q}}_s(x,a)| \le V_{\hat{\mathbf{q}}}$ (where the last inequality holds thanks to $\gamma \le 1$), we obtain

$$\begin{aligned}&\boldsymbol{\Delta}_{t+N-1,s}(x)\\ &\quad\le \eta\hat{C} + \frac{\eta e' U_{\boldsymbol{\pi}\hat{\mathbf{q}}}}{1-\gamma}\sum_a |\hat{\mathbf{q}}_s(x,a)| + \frac{\eta e' V_{\hat{\mathbf{q}}}}{1-\gamma}\sum_{t'=t}^{t+N-2}\boldsymbol{\Delta}_{t',t}(x)\\ &\qquad + \eta\sum_a |\hat{\mathbf{q}}_s(x,a)|\frac{\boldsymbol{\pi}_{t+N-1}(a|x)}{1-\gamma}\left[U_{\hat{\mathbf{v}}} + \sum_{t'=t}^{t+N-1}\boldsymbol{\Delta}_{t',t}(x)\right]\\ &\quad\le \eta\hat{C} + \frac{\eta e' U_{\boldsymbol{\pi}\hat{\mathbf{q}}}}{1-\gamma}\sum_a |\hat{\mathbf{q}}_s(x,a)| + \frac{\eta e' V_{\hat{\mathbf{q}}}}{1-\gamma}\sum_{t'=t}^{t+N-2}\boldsymbol{\Delta}_{t',t}(x)\\ &\qquad + \frac{\eta U_{\hat{\mathbf{v}}}}{1-\gamma}\sum_a |\hat{\mathbf{q}}_s(x,a)| + \frac{\eta V_{\hat{\mathbf{q}}}}{1-\gamma}\sum_{t'=t}^{t+N-1}\boldsymbol{\Delta}_{t',t}(x)\\ &\quad= \eta\hat{C} + \frac{\eta(e' U_{\boldsymbol{\pi}\hat{\mathbf{q}}} + U_{\hat{\mathbf{v}}})}{1-\gamma}\sum_a |\hat{\mathbf{q}}_s(x,a)| + \frac{\eta\, e\, V_{\hat{\mathbf{q}}}}{1-\gamma}\sum_{t'=t}^{t+N-1}\boldsymbol{\Delta}_{t',t}(x).\end{aligned}$$

Now, take the expectation of both sides and use that $\mathbb{E}\left[|\hat{\mathbf{q}}_s(x,a)|\right] \le U_{\bar{\mathbf{q}}}$. Introducing the constant $C_1 = \hat{C}(1-\gamma) + (e'U_{\boldsymbol{\pi}\hat{\mathbf{q}}} + U_{\hat{\mathbf{v}}})U_{\bar{\mathbf{q}}}|\mathcal{A}| = e'(L+\gamma V_{\hat{\mathbf{q}}}) + (e'U_{\boldsymbol{\pi}\hat{\mathbf{q}}} + U_{\hat{\mathbf{v}}})U_{\bar{\mathbf{q}}}|\mathcal{A}|$, we get

$$\mathbb{E}\left[\boldsymbol{\Delta}_{t+N-1,s}(x)\right] \le \frac{\eta\, C_1}{1-\gamma} + \frac{\eta\, e\, V_{\hat{\mathbf{q}}}}{1-\gamma}\sum_{t'=t}^{t+N-1}\mathbb{E}\left[\boldsymbol{\Delta}_{t',t}(x)\right]. \tag{41}$$

Taking $s = t$ in (41), using $1 - \eta\frac{e\,V_{\hat{\mathbf{q}}}}{1-\gamma} > 0$ which holds by our assumption on η and γ, reordering gives

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{\Delta}_{t+N-1,t}(x)\right] \\
&\quad\le \frac{\eta(1-\gamma)}{1-\gamma-\eta\,e\,V_{\hat{\mathbf{q}}}}\left\{\frac{C_1}{1-\gamma}+\frac{e\,V_{\hat{\mathbf{q}}}}{1-\gamma}\sum_{t'=t}^{t+N-2}\mathbb{E}\left[\boldsymbol{\Delta}_{t',t}(x)\right]\right\} \\
&\quad\le \frac{\eta}{1-\gamma-\eta\,e\,V_{\hat{\mathbf{q}}}}\left\{C_1+e\,V_{\hat{\mathbf{q}}}\,(N-1)c'\right\} \\
&\quad\le \frac{\eta}{1-\gamma-\eta\,e\,V_{\hat{\mathbf{q}}}}\left\{C_1+e\,V_{\hat{\mathbf{q}}}\,N\,c'\right\} = c', \qquad (42)
\end{aligned}$$

where the second inequality follows since, by our induction hypothesis, $\mathbb{E}\left[\boldsymbol{\Delta}_{t',t}(x)\right] \le c'$ holds for any t' such that $1 \le t' \le t+N-2$, the third follows by our assumptions on N, η and γ, while the last equality holds by the definition of c'. This shows that the induction hypothesis holds for the pair $(t+N-1, t)$.

Let us now consider the pairs $(t+N-1, s)$, where $s \ne t$. We start from (41) again. Note that by our induction hypothesis, $\mathbb{E}\left[\boldsymbol{\Delta}_{t',t}(x)\right] \le c'$ for $t \le t' \le t+N-2$. Furthermore, by (42), we also have $\mathbb{E}\left[\boldsymbol{\Delta}_{t+N-1,t}(x)\right] \le c'$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Delta}_{t+N-1,s}(x)\right] &\le \frac{\eta\,C_1}{1-\gamma}+\frac{\eta\,e\,V_{\hat{\mathbf{q}}}}{1-\gamma}\,N\,c' \\
&\le \frac{\eta\,C_1}{1-\gamma-\eta e V_{\hat{\mathbf{q}}}}+\frac{\eta\,e\,V_{\hat{\mathbf{q}}}}{1-\gamma-\eta e V_{\hat{\mathbf{q}}}}\,N\,c' = c'.
\end{aligned}$$

■

D. Proof of Proposition 1

For every x, a define $\mathbf{Q}_T(x,a) = \sum_{t=N}^{T}\mathbf{q}_t(x,a)$ and $\mathbf{V}_T(x) = \sum_{t=N}^{T}\mathbf{v}_t(x)$. Lemma 2 shows that in order to prove Proposition 1, it suffices to prove an upper bound on $\mathbb{E}\left[\mathbf{Q}_T(x,a) - \mathbf{V}_T(x)\right]$.

Lemma 15: Let c be as in Lemma 12 and c' be as in Lemma 14. Assume that $\gamma \in (0,1)$, $c(\tau+1)^2 < \beta/2$, $N \ge 1 + \left\lceil \tau \ln\left(\frac{4}{\beta - 2c(\tau+1)^2}\right)\right\rceil$, $0 < \eta < \frac{\beta(1-\gamma)}{2e(N+1)(|\mathcal{A}|/\gamma+2\tau+2)}$, and $T \ge N$ hold. Then, for all $(x,a) \in \mathcal{X}\times\mathcal{A}$,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Q}_T(x,b) - \mathbf{V}_T(x)\right] \le \frac{\ln|\mathcal{A}|}{\eta} &+ (N-1)(U_{\bar{\mathbf{q}}}+U_q) \\
&+ (T-2N+2)\Big(c'(N-1)(1+\eta\,e''V_{\hat{\mathbf{q}}}) \\
&+ \gamma\,U_q + \eta\,e''|\mathcal{A}|\,U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\,U_{\bar{\mathbf{q}}}\Big).
\end{aligned}$$

Proof: Note that if $T < 2N$, the conclusion of the lemma trivially holds. Therefore, in what follows we assume that $T \ge 2N$. First, note that the conditions of both Lemmas 13 and 9 are satisfied. Thus, the conclusions of Lemma 9 and therefore also those of Lemmas 6–9 hold. In particular, by Lemma 9, $\boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)| \le U_{\boldsymbol{\pi}\hat{\mathbf{q}}} = \frac{4}{\beta}(\tau+2)$ holds for any $t \ge N+1$. Now, using Lemma 6, we have $\hat{\mathbf{q}}_t(x,a) \le V_{\hat{\mathbf{q}}} = \frac{2}{\beta}(|\mathcal{A}|/\gamma + 2\tau + 2)$, thus by the constraint on η, $\eta\hat{\mathbf{q}}_t(x,a) \le 1$.

We will follow the steps of the proof in Auer et al. [2]. For $1 \le t \le T$, define $\mathbf{W}_t(x) = \sum_a \mathbf{w}_t(x,a)$. Fix a time step t such that $2N-1 \le t \le T$ and a state-action pair $(x,a) \in \mathcal{X}\times\mathcal{A}$. Recalling $e'' = e-2$, we have the following inequalities:

$$\begin{aligned}
\frac{\mathbf{W}_{t+1}(x)}{\mathbf{W}_t(x)} &= \sum_a \frac{\mathbf{w}_{t+1}(x,a)}{\mathbf{W}_t(x)} = \sum_a \frac{\mathbf{w}_t(x,a)}{\mathbf{W}_t(x)} e^{\eta\hat{\mathbf{q}}_{t-N+1}(x,a)} \\
&= \sum_a \frac{\boldsymbol{\pi}_t(a|x) - \gamma/|\mathcal{A}|}{1-\gamma} e^{\eta\hat{\mathbf{q}}_{t-N+1}(x,a)} \\
&\le \sum_a \frac{\boldsymbol{\pi}_t(a|x) - \gamma/|\mathcal{A}|}{1-\gamma}\left(1+\eta\hat{\mathbf{q}}_{t-N+1}(x,a) + e''\left(\eta\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2\right) \\
&\qquad \text{(as } \eta\hat{\mathbf{q}}_{t-N+1}(x,a) \le 1 \text{ and for } x \le 1, e^x \le 1 + x + e''x^2\text{)} \\
&\le 1 + \frac{\eta}{1-\gamma}\sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_{t-N+1}(x,a) \\
&\quad + \frac{\eta^2 e''}{1-\gamma}\sum_a \boldsymbol{\pi}_t(a|x)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2.
\end{aligned}$$

Introduce

$$\mathbf{q}_{2,t-N+1}(x) = \sum_a \boldsymbol{\pi}_t(a|x)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2.$$

We now show a bound on the expectation of this quantity that will be useful later. For this, write

$$\begin{aligned}
\mathbf{q}_{2,t-N+1}(x) &= \sum_a \boldsymbol{\pi}_t(a|x)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2 \\
&= \sum_a \boldsymbol{\pi}_{t-N+1}(a|x)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2 \\
&\quad + \sum_a \left(\boldsymbol{\pi}_t(a|x) - \boldsymbol{\pi}_{t-N+1}(a|x)\right)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2
\end{aligned}$$

By Lemma 9, the first term on the right-hand side can be bounded as follows:

$$\sum_a \boldsymbol{\pi}_{t-N+1}(a|x)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2 \le U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\sum_a\left|\hat{\mathbf{q}}_{t-N+1}(x,a)\right|,$$

while, thanks to Lemma 14, the second one is bounded, in expectation, by

$$\begin{aligned}
&\mathbb{E}\left[\sum_a\left(\boldsymbol{\pi}_t(a|x) - \boldsymbol{\pi}_{t-N+1}(a|x)\right)\left(\hat{\mathbf{q}}_{t-N+1}(x,a)\right)^2\right] \\
&\qquad \le \mathbb{E}\left[V_{\hat{\mathbf{q}}}\sum_{t'=t-N+1}^{t-1}\boldsymbol{\Delta}_{t',t-N+1}(x)\right] \le (N-1)\,c'\,V_{\hat{\mathbf{q}}},
\end{aligned}$$

where we have used that $|\hat{\mathbf{q}}_t(x,a)| \le V_{\hat{\mathbf{q}}}$ and also that $\mathbb{E}\left[\boldsymbol{\Delta}_{t',s}(x)\right] \le c'$. Combining these inequalities, we get that

$$\mathbb{E}\left[\mathbf{q}_{2,t-N+1}(x)\right] \le |\mathcal{A}|\,U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\,U_{\bar{\mathbf{q}}} + (N-1)\,c'\,V_{\hat{\mathbf{q}}}. \qquad (43)$$

Let us now return to developing an upper-bound on $\frac{\mathbf{W}_{t+1}(x)}{\mathbf{W}_t(x)}$. Defining $\hat{\mathbf{v}}_t^N(x) = \sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_{t-N+1}(x,a)$, we obtain

$$\frac{\mathbf{W}_{t+1}(x)}{\mathbf{W}_t(x)} \le 1 + \frac{\eta}{1-\gamma}\hat{\mathbf{v}}_t^N(x) + \frac{\eta^2 e''}{1-\gamma}\mathbf{q}_{2,t-N+1}(x).$$

Using $1 + x \le e^x$ and then taking logarithms gives

$$\ln\frac{\mathbf{W}_{t+1}(x)}{\mathbf{W}_t(x)} \le \frac{\eta}{1-\gamma}\hat{\mathbf{v}}_t^N(x) + \frac{\eta^2 e''}{1-\gamma}\mathbf{q}_{2,t-N+1}(x).$$

Summing over $t = 2N-1, 2N, \ldots, T$, we get

$$\ln\frac{\mathbf{W}_{T+1}(x)}{\mathbf{W}_{2N-1}(x)} \le \frac{\eta}{1-\gamma}\widehat{\mathbf{V}}_T^N(x) + \frac{\eta^2 e''}{(1-\gamma)}\mathbf{Q}_{2,T}^N(x), \qquad (44)$$

where $\widehat{\mathbf{V}}_T^N(x) = \sum_{t=2N-1}^{T}\hat{\mathbf{v}}_t^N(x)$ and $\mathbf{Q}_{2,T}^N(x) = \sum_{t=N}^{T-N+1}\mathbf{q}_{2,t}(x)$.

Now, considering a lower bound on the left-hand side, we have for any action b,

$$\ln\frac{\mathbf{W}_{T+1}(x)}{\mathbf{W}_{2N-1}(x)} \ge \ln\frac{\mathbf{w}_{T+1}(x,b)}{\mathbf{W}_{2N-1}(x)} = \eta\sum_{t=N}^{T-N+1}\hat{\mathbf{q}}_t(x,b) - \ln|\mathcal{A}|,$$

where we used that $\mathbf{w}_{2N-1}(x,a) = 1$ holds for all $a \in \mathcal{A}$. Combining with (44), we get

$$\widehat{\mathbf{V}}_T^N(x) \ge (1-\gamma)\widehat{\mathbf{Q}}_T^N(x,b) - \frac{\ln|\mathcal{A}|}{\eta} - \eta e''\,\mathbf{Q}_{2,T}^N(x), \qquad (45)$$

where $\widehat{\mathbf{Q}}_T^N(x,b) = \sum_{t=N}^{T-N+1}\hat{\mathbf{q}}_t(x,b)$.

Let us now bound the difference of $\widehat{\mathbf{V}}_T^N(x)$ and

$$\widehat{\mathbf{V}}_T(x) = \sum_{t=N}^{T} \hat{\mathbf{v}}_t(x) = \sum_{t=N}^{T} \sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_t(x,a).$$

Note that

$$\widehat{\mathbf{V}}_T^N(x) = \sum_{t=N}^{T-N+1} \sum_a \boldsymbol{\pi}_{t+N-1}(a|x)\hat{\mathbf{q}}_t(x,a).$$

Therefore,

$$\begin{aligned}
&\widehat{\mathbf{V}}_T^N(x) - \widehat{\mathbf{V}}_T(x) \\
&\le \sum_{t=N}^{T-N+1} \sum_a |\hat{\mathbf{q}}_t(x,a)| \Big| \boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_t(a|x) \Big| \\
&\quad + \sum_{t=T-N+2}^{T} \sum_a \boldsymbol{\pi}_t(a|x) |\hat{\mathbf{q}}_t(x,a)| \\
&= \sum_{t=N}^{T-N+1} \sum_{t'=t}^{t+N-2} \boldsymbol{\Delta}_{t',t}(x) + \sum_{t=T-N+2}^{T} \sum_a \boldsymbol{\pi}_t(a|x) |\hat{\mathbf{q}}_t(x,a)| \,,
\end{aligned}$$

where we used the definition of $\boldsymbol{\Delta}_{t',t}(x)$. Taking the expectation of both sides, using Lemmas 14 and 8, we get

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathbf{V}}_T^N(x)\right] - \mathbb{E}\left[\widehat{\mathbf{V}}_T(x)\right] &\le (T-2N+2)(N-1)c' + (N-1)U_{\bar{\mathbf{q}}} \\
&= (N-1)\{(T-2N+2)c' + U_{\bar{\mathbf{q}}}\}.
\end{aligned}$$

This, together with (45) gives

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mathbf{V}}_T(x)\right] + (N-1)\left\{(T-2N+2)\,c' + U_{\bar{\mathbf{q}}}\right\} \\
&\ge (1-\gamma)\mathbb{E}\left[\widehat{\mathbf{Q}}_T^N(x,b)\right] - \frac{\ln|\mathcal{A}|}{\eta} - \eta e'' \,\mathbb{E}\left[\mathbf{Q}_{2,T}^N(x)\right].
\end{aligned} \tag{46}$$

By equation (13), we have $\mathbb{E}\left[\widehat{\mathbf{V}}_T(x)\right] = \mathbb{E}\left[\mathbf{V}_T(x)\right]$ and with the definition $\mathbf{Q}_T^N(x,b) = \sum_{t=N}^{T-N+1} \mathbf{q}_t(x,b)$, we also have $\mathbb{E}\left[\widehat{\mathbf{Q}}_T^N(x,b)\right] = \mathbb{E}\left[\mathbf{Q}_T^N(x,b)\right]$. Thus, using (43) again, we get

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{V}_T(x)\right] + (N-1)\left\{(T-2N+2)\,c' + U_{\bar{\mathbf{q}}}\right\} \\
&\quad \ge (1-\gamma)\mathbb{E}\left[\mathbf{Q}_T^N(x,b)\right] - \frac{\ln|\mathcal{A}|}{\eta} \\
&\qquad - \eta e'' \,(T-2N+2)\{|\mathcal{A}|\, U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\, U_{\bar{\mathbf{q}}} + (N-1)\,c'\, V_{\hat{\mathbf{q}}}\}.
\end{aligned}$$

By reordering the terms, this becomes

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{Q}_T^N(x,b) - \mathbf{V}_T(x)\right] \\
&\le \gamma\mathbb{E}\left[\mathbf{Q}_T^N(x,b)\right] + (N-1)\left\{(T-2N+2)\,c' + U_{\bar{\mathbf{q}}}\right\} + \frac{\ln|\mathcal{A}|}{\eta} \\
&\quad + \eta e'' \,(T-2N+2)\{|\mathcal{A}|\, U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\, U_{\bar{\mathbf{q}}} + (N-1)\,c'\, V_{\hat{\mathbf{q}}}\}\,.
\end{aligned} \tag{47}$$

We now lower bound $\mathbf{Q}_T^N(x,a)$ by $\mathbf{Q}_T(x,a)$. Since the rewards $r_t(x,a)$ are bounded between 0 and 1, by Lemma 3 we have

$$\mathbf{q}_t(x,b) \le U_q = 2\tau + 3\,. \tag{48}$$

Therefore,

$$\mathbf{Q}_T(x,a) - \mathbf{Q}_T^N(x,a) = \sum_{t=T-N+2}^{T} \mathbf{q}_t(x,a) \le U_q(N-1), \tag{49}$$

Moreover, (48) also implies that $\mathbb{E}\left[\mathbf{Q}_T^N(x,b)\right] \le U_q(T-2N+2)$. Combining this with (49) and (47), we obtain the desired bound:

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{Q}_T(x,b) - \mathbf{V}_T(x)\right] \le \frac{\ln|\mathcal{A}|}{\eta} + (N-1)(U_{\bar{\mathbf{q}}} + U_q) \\
&+ (T-2N+2)\Big(c'(N-1)(1+\eta\, e'' V_{\hat{\mathbf{q}}}) + \gamma\, U_q + \eta\, e''|\mathcal{A}|\, U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\, U_{\bar{\mathbf{q}}}\Big).
\end{aligned}$$

■

The proof of Proposition 1 is now easy:

Proof of Proposition 1: Under the conditions of the proposition, combining Lemmas 2 and 15, and using that $0 \le \rho_t^\pi, \boldsymbol{\rho}_t \le 1$ yields

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\rho_t^\pi - \boldsymbol{\rho}_t\right] \le N - 1 + \sum_{x,a} \mu_{\mathrm{st}}^\pi(x)\pi(a|x)\,\mathbb{E}\left[\mathbf{Q}_T(x,a) - \mathbf{V}_T(x)\right] \\
&\le (N-1)(U_{\bar{\mathbf{q}}} + U_q + 1) + \frac{\ln|\mathcal{A}|}{\eta} \\
&+ (T-2N+2)\Big(c'(N-1)(1+\eta\, e'' V_{\hat{\mathbf{q}}}) + \gamma\, U_q + \eta\, e''|\mathcal{A}|\, U_{\boldsymbol{\pi}\hat{\mathbf{q}}}\, U_{\bar{\mathbf{q}}}\Big)\,.
\end{aligned}$$

proving Proposition 1. ■

E. Proof of Proposition 2

Let $t \ge N$. First, since $\boldsymbol{\pi}_t$ is $\sigma(\mathbf{u}_{t-N})$-measurable, $\mathbb{E}\left[\boldsymbol{\rho}_t\right] = \mathbb{E}\left[\sum_x \mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x)\mathbb{E}\left[\,r_t(x,\mathbf{a}_t)|\,\mathbf{u}_{t-N}\right]\right]$. We also have

$$\begin{aligned}
\mathbb{E}\left[r_t(\mathbf{x}_t,\mathbf{a}_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\,r_t(\mathbf{x}_t,\mathbf{a}_t)|\,\mathbf{u}_{t-N}\right]\right] \\
&= \mathbb{E}\left[\sum_x \boldsymbol{\mu}_t^N(x)\,\mathbb{E}\left[\,r_t(x,\mathbf{a}_t)|\,\mathbf{u}_{t-N}\right]\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\rho}_t - r_t(\mathbf{x}_t,\mathbf{a}_t)\right] &= \mathbb{E}\left[\sum_x (\mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x) - \boldsymbol{\mu}_t^N(x))\,\mathbb{E}\left[\,r_t(x,\mathbf{a}_t)|\,\mathbf{u}_{t-N}\right]\right] \\
&\le \mathbb{E}\left[\sum_x \Big|\mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x) - \boldsymbol{\mu}_t^N(x)\Big|\right],
\end{aligned}$$

where we have used that $r_t(x,a) \in [0,1]$.

Thanks to Lemma 13, Lemma 10 is applicable. Hence, $\sum_x \left|\mu_{\mathrm{st}}^{\boldsymbol{\pi}_t}(x) - \boldsymbol{\mu}_t^N(x)\right| \le c(\tau+1)^2 + 2e^{-(N-1)/\tau}$, and thus $\mathbb{E}\left[\boldsymbol{\rho}_t - r_t(\mathbf{x}_t,\mathbf{a}_t)\right] \le c(\tau+1)^2 + 2e^{-(N-1)/\tau}$. Summing up these inequalities for $t = N,\ldots,T$, and using the trivial bound $\mathbb{E}\left[\boldsymbol{\rho}_t - r_t(\mathbf{x}_t,\mathbf{a}_t)\right] \le 1$ for the first $N-1$ terms, we get the desired result. ■

ACKNOWLEDGMENTS

This work was supported in part by the Hungarian Scientific Research Fund and the Hungarian National Office for Research and Technology (KTIA-OTKA CNK 77782), the PASCAL2 Network of Excellence under EC grant no. 216886, NSERC, AITF, the Alberta Ingenuity Centre for Machine Learning, the DARPA GALE project (HR0011-08-C-0110) and iCore.

REFERENCES

[1] Antos, A., Grover, V., and Szepesvári, C. (2010). Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29–30):2712–2728.

[2] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77.

[3] Cao, X.-R. (2007). *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer, New York.

[4] Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50:2050–2057.

[5] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA.

[6] Even-Dar, E., Kakade, S. M., and Mansour, Y. (2005). Experts in a Markov decision process. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 401–408, Cambridge, MA, USA. MIT Press.

[7] Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009). Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736.

[8] György, A., Linder, T., Lugosi, G., and Ottucsák, Gy. (2007). The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403.

[9] Ipsen, I. C. F. and Selee, T. M. (2011). Ergodicity coefficients defined by vector norms. *SIAM J. Matrix Analysis Applications*, 32(1):153–200.

[10] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600.

[11] Kakade, S., Sridharan, K., and Tewari, A. (2010). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 793–800. Curran Associates. (December 7–10, 2009).

[12] Lazaric, A. and Munos, R. (2011). Learning with stochastic inputs and adversarial outputs. *Journal of Computer and System Sciences*. To appear.

[13] Neu, G., György, A., and Szepesvári, Cs. (2010). The online loop-free stochastic shortest-path problem. In Kalai, A. and Mohri, M., editors, *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 231–243.

[14] Neu, G., György, A., and Szepesvári, Cs. (2012). The adversarial stochastic shortest path problem with unknown transition probabilities. In Lawrence, N. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR Workshop and Conference Proceedings*, pages 805–813.

[15] Neu, G., György, A., and Szepesvári, Cs. (2013). The online loop-free stochastic shortest-path problem. *In preparation*.

[16] Neu, G., György, A., Szepesvári, Cs., and Antos, A. (2011). Online Markov decision processes under bandit feedback. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1804–1812.

[17] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience.

[18] Rakhlin, A., Sridharan, K., and Tewari, A. (2012). Online learning: Stochastic and constrained adversaries. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, page 2752, Cambridge, MA, USA. Curran Associates. (December 12–15, 2011).

[19] Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIP Press.

[20] Yu, J. Y. and Mannor, S. (2009a). Arbitrarily modulated Markov decision processes. In *Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*, pages 2946–2953. IEEE Press.

[21] Yu, J. Y. and Mannor, S. (2009b). Online learning in Markov decision processes with arbitrarily changing rewards and transitions. In *GameNets'09: Proceedings of the First ICST International Conference on Game Theory for Networks*, pages 314–322, Piscataway, NJ, USA. IEEE Press.

[22] Yu, J. Y., Mannor, S., and Shimkin, N. (2009). Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757.