Online Markov Decision Processes under Bandit Feedback

Gergely Neu[*†]
[*]Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Hungary
neu.gergely@gmail.com

András György
[†]Machine Learning Research Group
MTA SZTAKI Institute for Computer Science and Control, Hungary
gya@szit.bme.hu

Csaba Szepesvári
Department of Computing Science,
University of Alberta, Canada
szepesva@ualberta.ca

András Antos
Machine Learning Research Group
MTA SZTAKI Institute for Computer Science and Control, Hungary
antos@szit.bme.hu

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 terms of the total reward received. 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 a no-regret algorithm. In this paper we propose a new learning algorithm and, assuming that stationary policies mix uniformly fast, we show that the expected regret of the new algorithm in T time steps is $\mathcal{O}\left(T^{2/3}(\ln T)^{1/3}\right)$, giving the first rigorously proved regret bound for the problem.

1 Introduction

The problem that we consider is online learning in finite Markov decision processes (MDPs) with a fixed, known dynamics. The problem is defined formally as follows: An agent navigates in a finite stochastic environment by selecting actions based on the states and rewards experienced previously. At each time instant the agent observes the reward associated with the last transition and the current state, that is, at time $t+1$ the agent observes $r_t(\mathbf{x}_t, \mathbf{a}_t)$, where $\mathbf{x}_t$ is the state visited at time t and $\mathbf{a}_t$ is the action chosen. The agent does not observe the rewards associated with other transitions, that is, the agent faces a *bandit* situation. The goal of the agent is to maximize its total expected reward $\hat{R}_T$ in T steps. As opposed to the standard MDP setting, the reward function at each time step may be different. The only assumption about this sequence of reward functions r_t is that they are chosen ahead of time, independently of how the agent acts. However, no statistical assumptions are made about the choice of this sequence. As usual in such cases, a meaningful performance measure for the agent is how well it can compete with a certain class of reference policies, in our case the set of all stationary policies: If R_T^* denotes the expected total reward in T steps that can be collected by choosing the best stationary policy (this policy can be chosen based on the full knowledge of the sequence r_t), the goal of learning can be expressed as minimizing the total expected regret, $\hat{L}_T = R_T^* - \hat{R}_T$.

In this paper we propose a new algorithm for this setting. Assuming that the stationary distributions underlying stationary policies exist, are unique and they are uniformly bounded away from zero and that these policies mix uniformly fast, our main result shows that the total expected regret of our algorithm in T time steps is $\mathcal{O}\left(T^{2/3}(\ln T)^{1/3}\right)$.

The first work that considered a similar online learning setting is due to Even-Dar et al. (2005, 2009). In fact, this is the work that provides the starting point for our algorithm and analysis. The major difference between our work and that of Even-Dar et al. (2005, 2009) is that they assume that the reward function is fully observed (i.e., in each time step the learning agent observes the whole reward function r_t), whereas we consider the bandit setting. The main result in these works is a bound on the total expected regret, which scales with the square root of the number of time steps under mixing assumptions identical to our assumptions. They propose an algorithm, MDP-E, which is very similar to ours in that it uses some (optimized) expert algorithm in every state which is fed with the action-values of the policy used in the last round. Another work that considered the full information problem is due to Yu et al. (2009) who proposed new algorithms and proved a bound on the expected regret of order $\mathcal{O}\left(T^{3/4+\varepsilon}\right)$ for arbitrary $\varepsilon \in (0, 1/3)$. The algorithm proposed ("Lazy FPL") works with phases of length $m^{1/3-\varepsilon}$ and changes policies only at the end of the phases. At the end of a phase the optimal (differential) value function corresponding to the sum of past reward functions is first found. Within the phase, the action to be followed at some time step is then selected as the one that maximizes the one-step lookahead action value computed with this value function but with the immediate rewards perturbed randomly in an appropriate manner. The advantage of this algorithm to that of Even-Dar et al. (2009) is that it is computationally less expensive, which, however, comes at the price of an increased bound on the regret. Yu et al. (2009) introduced another algorithm ("Q-FPL") and they have shown a sublinear ($o(T)$) almost sure bound on the regret.

All the works reviewed so far considered the full information case. The requirement that the full reward function must be given to the agent at every time step significantly limits their applicability. There are only three papers that we know of where the bandit situation was considered.

The first paper which falls into this category is due to Yu et al. (2009) who proposed an algorithm ("Exploratory FPL") for this setting. This algorithm estimates the immediate rewards by appropriately weighting the rewards received and in a phase either uses a uniformly exploring policy or that of underlying their Lazy FPL algorithm. They prove an $o(T)$ almost sure bound on the regret of this algorithm.

Recently, Neu et al. (2010) gave $\mathcal{O}\left(\sqrt{T}\right)$ regret bounds for a special bandit setting when the agent interacts with a *loop-free episodic* environment. The algorithm and analysis in this work heavily exploits the specifics of these environments (i.e., that in the same episode no state can be visited twice) and so they do not generalize to our setting.

Another closely related work is due to Yu and Mannor (2009a,b) who considered the 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 changed arbitrarily. Accordingly, the algorithms proposed in these papers fail to achieve consistency for this setting. The reason these papers are relevant because the regret bounds are provided in terms of a parameter ε which describe the extent by which the transition probabilities are allowed to vary. By taking the limit $\varepsilon \to 0$, we may obtain results for the case when the transition probabilities are fixed. This way, we can obtain a result from Theorem IV.1 of Yu and Mannor (2009b) for the case which interests us, that is, when rewards are only observed along the trajectory traversed by the agent. However, the result which can be obtained this way seems to be incorrect: If the state space consists of only a single state, the learning problem becomes identical to the non-stochastic multi-armed bandit problem. The above technique then gives a bound of order $\mathcal{O}\left(\sqrt{\ln|\mathcal{A}|T}\right)$ on the expected regret, which contradicts the known $\Omega\left(\sqrt{|\mathcal{A}|T}\right)$ lower bound.[1] It remains for future work to see if the results in this paper can be corrected. Thus, currently, the only result for the case considered in this paper is an asymptotic "no-regret" result.

[1]To show this contradiction, one has to replace, in the bound of Theorem IV.1 of Yu and Mannor (2009b), the condition $T > N$ with an extra $\mathcal{O}(1/T)$ term, and then let ϵ and δ converge to zero at appropriate rates.

The rest of the paper is organized as follows: The problem is laid out in Section 2, which is followed by a section about our assumptions (Section 3). The algorithm and the main result are given in Section 4, while the proof of the latter, with the exception of some technical results given in the Appendix, is presented in Section 5.

2 Problem definition

Formally, a finite Markov Decision Process (MDP) M is defined 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 $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$. In time step $t \in \{1, 2, \ldots\}$, knowing the state $\mathbf{x}_t \in \mathcal{X}$, an agent acting in the MDP M chooses an action $\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)$ to be executed based on $(\mathbf{x}_t, r(\mathbf{a}_{t-1}, \mathbf{x}_{t-1}), \mathbf{a}_{t-1}, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_2, r(\mathbf{a}_1, \mathbf{x}_1), \mathbf{a}_1, \mathbf{x}_1)$.[2] Here $\mathcal{A}(x) \subset \mathcal{A}$ is the set of admissible actions at state x. As a result of executing the chosen action the process moves to state $\mathbf{x}_{t+1} \in \mathcal{X}$ with probability $P(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t)$ and the agent receives reward $r(\mathbf{x}_t, \mathbf{a}_t)$. In the so-called average-reward problem, the goal of the agent is to maximize the average reward received over time. For a more detailed introduction the reader is referred to, for example, Puterman (1994).

2.1 Online learning in MDPs

In this paper we consider the online version of MDPs when the reward function is allowed to change arbitrarily. 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 receives $\mathbf{x}_t$ and then selects an action $\mathbf{a}_t$ 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_0.

The goal of the learning agent is to maximize its expected total reward

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

An equivalent goal is to minimize the regret, that is, to minimize the difference between the expected total reward received by the best algorithm within some reference class and the expected total reward of the learning algorithm. In the case of MDPs a reasonable reference class, used by various previous works (e.g., Even-Dar et al., 2005, 2009; Yu et al., 2009) is the class of stationary stochastic policies.[3] A stationary stochastic policy, π, (or, in short: a policy) is a mapping $\pi : \mathcal{A} \times \mathcal{X} \to [0, 1]$, where $\pi(a|x) \equiv \pi(a, x)$ is the probability of taking action a in state x. We say that a policy π is followed in an MDP if the action at time t is drawn from π, independently of previous states and actions given the current state $\mathbf{x}'_t$: $\mathbf{a}'_t \sim \pi(\cdot|\mathbf{x}'_t)$. The expected total reward while following a policy π is defined as

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

Here $\{(\mathbf{x}'_t, \mathbf{a}'_t)\}$ denotes the trajectory that results from following policy π from $\mathbf{x}'_1 \sim P_0$.

The *expected regret* (or *expected relative loss*) of the learning agent relative to the class of policies (in short, the regret) is defined as

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

where the supremum is taken over all (stochastic stationary) policies. Note that the optimal policy is chosen in hindsight, depending acausally on the reward function. If the regret of an agent grows

[2] We follow the convention that boldface letters denote random variables.

[3] This is a reasonable reference class because for a fixed reward function one can always find a member of it which maximizes the average reward per time step, see Puterman (1994).

sublinearly with T then we can say that in the long run it acts as well as the best (stochastic stationary) policy (i.e., the average expected regret of the agent is asymptotically equal to that of the best policy).

3 Assumptions

In this section we list the assumptions that we make throughout the paper about the transition probability kernel (hence, these assumptions will not be mentioned in the subsequent results). In addition, recall that we assume that the rewards are bound to $[0, 1]$.

Before describing the assumptions, a few more definitions are needed: Let π be a stationary policy. Define

$$P^\pi(x'|x) = \sum_a \pi(a|x)P(x'|x, a).$$

We will also view P^π as a matrix: $(P^\pi)_{x,x'} = P^\pi(x'|x)$, where, without loss of generality, we assume that $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$. In general, distributions will also be treated as *row* vectors. Hence, for a distribution μ, μP^π is the distribution over $\mathcal{X}$ that results from using policy π for one step from μ (i.e., the "next-state distribution" under π). Remember that the stationary distribution of a policy π is a distribution μ which satisfies $\mu P^\pi = \mu$.

Assumption A1 *Every policy π has a well-defined unique stationary distribution μ^π.*

Assumption A2 *The stationary distributions are uniformly bounded away from zero: $\inf_{\pi,x} \mu^\pi(x) \geq \beta$ for some $\beta > 0$.*

Assumption A3 *There exists some fixed positive τ such that for any two arbitrary distributions μ and μ' over $\mathcal{X}$,*

$$\sup_\pi \|(\mu - \mu')P^\pi\|_1 \leq e^{-1/\tau}\|\mu - \mu'\|_1,$$

where $\|\cdot\|_1$ is the 1-norm of vectors: $\|v\|_1 = \sum_i |v_i|$.

Note that Assumption A3 implies Assumption A1. The quantity τ is called the *mixing time* underlying P by Even-Dar et al. (2009) who also assume A3.

4 Learning in online MDPs under bandit feedback

In this section we shall first introduce some additional, standard MDP definitions, which we will be used later. That these are well-defined follows from our assumptions on P and from standard results to be found, for example, in the book by Puterman (1994). Next, we discuss a previous result that motivates our algorithm, which is followed by the definition of our algorithm. We finish by stating our main results concerning the performance of the proposed algorithm.

4.1 Preliminaries

Fix an arbitrary policy π and $t \geq 1$. Let $\{(\mathbf{x}'_s, \mathbf{a}'_s)\}$ be the random trajectory generated by π and the transition probability kernel P. Define the action-value and value functions underlying π and the immediate reward r_t by

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

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

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 computed by

$$\rho_t^\pi = \sum_x \mu^\pi(x) \sum_a \pi(a|x) r_t(x,a),$$

where μ^π is the stationary distribution underlying policy π. These value functions are equivalently defined by the Bellman equations:

$$q_t^\pi(x,a) = r_t(x,a) - \rho_t^\pi + \sum_{x'} P(x'|x,a) v_t^\pi(x')$$
$$v_t^\pi(x) = \sum_a \pi(a|x) q_t^\pi(x,a).$$

Now, consider the trajectory $\{(\mathbf{x}_t, \mathbf{a}_t)\}$ underlying a learning agent, where $\mathbf{x}_1$ is randomly chosen from P_0, and define

$$\mathbf{u}_t = (\, \mathbf{x}_1, \mathbf{a}_1, r_1(\mathbf{x}_1, \mathbf{a}_1),\ \mathbf{x}_2, \mathbf{a}_2, r_2(\mathbf{x}_2, \mathbf{a}_2),\ \ldots,\ \mathbf{x}_t, \mathbf{a}_t, r_t(\mathbf{x}_t, \mathbf{a}_t)\,)$$

and $\boldsymbol{\pi}_t(a|x) = \mathbb{P}[\mathbf{a}_t = a | \mathbf{u}_{t-1}, \mathbf{x}_t = x]$. That is, $\boldsymbol{\pi}_t$ denotes the policy followed by the agent at time step t (which is computed based on past information and is therefore random). We will use the following notation:

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

Thus, the following holds:

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

For reasons to be made clear later in the paper, we shall need the state distribution at time step t given that we start from the state-action pair (x,a) at time $t-N$, conditioned on the policies used between time steps $t-N$ and t:

$$\boldsymbol{\mu}_{t,x,a}^N(x') \stackrel{\text{def}}{=} \mathbb{P}\left[\mathbf{x}_t = x' \,|\, \mathbf{x}_{t-N} = x, \mathbf{a}_{t-N} = a, \boldsymbol{\pi}_{t-N+1}, \ldots, \boldsymbol{\pi}_{t-1}\right], \qquad x, x' \in \mathcal{X}, a \in \mathcal{A}.$$

It will be useful to view $\boldsymbol{\mu}_t^N$ as a matrix of dimensions $|\mathcal{X} \times \mathcal{A}| \times |\mathcal{X}|$. Thus, $\boldsymbol{\mu}_{t,x,a}^N(\cdot)$ will be viewed as one row of this matrix. To emphasize the conditional nature of this distribution, we will also use $\boldsymbol{\mu}_t^N(\cdot|x,a)$ instead of $\boldsymbol{\mu}_{t,x,a}^N(\cdot)$.

4.2 A previous result for the full-information setting

The starting point of our work is the paper of Even-Dar et al. (2009) who proposed an algorithm for the full information online learning problem that uses an "optimized best expert algorithm" to select the actions in each state x. The expert algorithm at state x is fed with the value functions $\mathbf{q}_t(x,\cdot)$.

Even-Dar et al. (2009) decompose the regret relative to a policy π as

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

Note that in the algorithm of Even-Dar et al. (2009), the policies $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T$ are computed deterministically based on the past, just like $\mathbf{q}_1, \ldots, \mathbf{q}_T$ (in particular, they do not depend on the past states and actions visited during learning). That is, each term in the above decomposition is deterministic.

Bounding each term separately Even-Dar et al. (2009) arrive at the following bound (cf. Theorem 5.1 in there):

$$\hat{L}_T \le 2\tau + 2 + \sqrt{(4\tau+6)T \ln|\mathcal{A}|} + \Big((1+3\tau)^2 \sqrt{T \ln|\mathcal{A}|} + 2\tau\Big). \qquad (2)$$

Note that this bound is a slightly refined version of the original one, the difference being that while Even-Dar et al. (2009) claimed that $q_t^\pi(x,a) \in [0, 3\tau]$ for all x, a, π (and $\tau \ge 1$), we prove in Lemma 2 that in fact $q_t^\pi(x,a) \in [-2\tau - 3, 2\tau + 3]$. We also handle some terms more carefully in Lemma 7 and replace the original factor of $4\tau^2$ by $(3\tau+1)^2$ in the third term. The first term in the bound comes from the following standard MDP result (which is a slightly corrected version of the corresponding lemma of Even-Dar et al., 2009):

Lemma 1. *For any $T \geq 1$ and any policy π it holds that*

$$\left(R_T^\pi - \sum_{t=1}^T \rho_t^\pi \right) \leq 2\tau + 2\,.$$

The proof is given for completeness.

Proof. Let $\{(\mathbf{x}_t, \mathbf{a}_t)\}$ be the trajectory when π is followed and let $\nu_t^\pi(x) = \mathbb{P}[\mathbf{x}_t = x]$. Then, using $\nu_t^\pi = \nu_{t-1}^\pi P^\pi$, the stationarity of μ^π and Assumption A3, we have

$$\begin{aligned}\left(R_T^\pi - \sum_{t=1}^T \rho_t^\pi \right) &= \sum_{t=1}^T \sum_x \left(\nu_t^\pi(x) - \mu^\pi(x)\right) \sum_a \pi(a|x) r_t(x,a) \\ &\leq \sum_{t=1}^T 2e^{-(t-1)/\tau} \leq 2 \left(1 + \int_0^\infty e^{-t/\tau}\,dt \right) = 2(\tau+1).\end{aligned}$$

□

The second term of (2) comes from the regret bound available for the expert algorithms sitting in the states. The last term, that compares the sum of average rewards $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_T$ to the actual expected return $\hat{R}_T$, is similar to the first term, just it uses the policies $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T$ instead of a fixed policy. Hence, similarly to the previous case, we can expect this term to stay small as long as the policies change slowly. The algorithms proposed by Even-Dar et al. (2009) produce policies that enjoy this property. In particular, the "total change" in the policies (in an appropriate metric) is bounded by $\mathcal{O}\left(\sqrt{T}\right)$. As a result, we get the bound as shown in the last term of (2).

4.3 The algorithm

Our algorithm is similar to that of Even-Dar et al. (2009) in that we use an expert algorithm in each state. Since in our case the full reward function r_t is not observed, the agent uses an estimate of it. The main difficulty is to come up with an unbiased estimate of r_t with a controlled variance. Here we propose to use the following estimate:

$$\hat{\mathbf{r}}_t(x,a) = \begin{cases} \frac{r_t(x,a)}{\boldsymbol{\pi}_t(a|x)\boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N},\mathbf{a}_{t-N})} & \text{if } (x,a) = (\mathbf{x}_t, \mathbf{a}_t) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $t \geq N+1$. Define $\hat{\mathbf{q}}_t$, $\hat{\mathbf{v}}_t$ and $\hat{\boldsymbol{\rho}}$ as the solution to the Bellman equations underlying the average reward MDP defined by $(P, \boldsymbol{\pi}_t, \hat{\mathbf{r}}_t)$:

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

Note that if N is sufficiently large and $\boldsymbol{\pi}_t$ changes sufficiently slowly then

$$\boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N}, \mathbf{a}_{t-N}) > 0, \tag{5}$$

almost surely, for arbitrary $x \in \mathcal{X}, t \geq N+1$. This fact will be shown in Lemma 9. Now, assume that $\boldsymbol{\pi}_t$ is computed based on $\mathbf{u}_{t-N}$, that is, $\boldsymbol{\pi}_t$ is measurable with respect to the σ-field $\sigma(\mathbf{u}_{t-N})$ generated by the history $\mathbf{u}_{t-N}$:

$$\boldsymbol{\pi}_t \in \sigma(\mathbf{u}_{t-N})\,. \tag{6}$$

Then also $\boldsymbol{\pi}_{t-1}, \ldots, \boldsymbol{\pi}_{t-N} \in \sigma(\mathbf{u}_{t-N})$ and $\boldsymbol{\mu}_t^N$ can be computed using

$$\boldsymbol{\mu}_{t,x,a}^N = e_x P^a P^{\boldsymbol{\pi}_{t-N+1}} \cdots P^{\boldsymbol{\pi}_{t-1}}, \tag{7}$$

where P^a is the transition probability matrix when in every state action a is used and e_x is the unit row vector corresponding to x (and we assumed that $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$). Moreover, a simple but tedious calculation shows that (5) and (6) ensure the conditional unbiasedness of our estimates: First note that

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

where we have exploited that $\boldsymbol{\pi}_t$, $\boldsymbol{\mu}_t^N$, $\mathbf{x}_{t-N}, \mathbf{a}_{t-N} \in \sigma(\mathbf{u}_{t-N})$. Let us now deal with $\mathbb{E}\left[\mathbb{I}_{\{(x,a)=(\mathbf{x}_t,\mathbf{a}_t)\}} \,|\, \mathbf{u}_{t-N}\right] = \mathbb{P}\left[\mathbf{a}_t = a \,|\, \mathbf{x}_t = x, \mathbf{u}_{t-N}\right]\mathbb{P}\left[\mathbf{x}_t = x \,|\, \mathbf{u}_{t-N}\right]$. Since by assumption $\boldsymbol{\pi}_t \in \sigma(\mathbf{u}_{t-N})$, $\mathbb{P}\left[\mathbf{a}_t = a \,|\, \mathbf{x}_t = x, \mathbf{u}_{t-N}\right] = \mathbb{P}\left[\mathbf{a}_t = a \,|\, \mathbf{x}_t = x, \mathbf{u}_{t-1}\right] = \boldsymbol{\pi}_t(a|x)$ holds, where the last equality follows from the definition of $\boldsymbol{\pi}_t$ and $\mathbf{a}_t$. Since $\boldsymbol{\pi}_{t-N+1}, \ldots, \boldsymbol{\pi}_{t-1} \in \sigma(\mathbf{u}_{t-N})$, $\mathbb{P}\left[\mathbf{x}_t = x \,|\, \mathbf{u}_{t-N}\right] = \mathbb{P}\left[\mathbf{x}_t = x \,|\, \mathbf{u}_{t-N}, \boldsymbol{\pi}_{t-N+1}, \ldots, \boldsymbol{\pi}_{t-1}\right] = \boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N},\mathbf{a}_{t-N})$. Combining these identities we get

$$\mathbb{E}\left[\hat{\mathbf{r}}_t(x,a)|\, \mathbf{u}_{t-N}\right] = r_t(x,a). \tag{8}$$

It then follows that

$$\mathbb{E}[\hat{\boldsymbol{\rho}}_t|\mathbf{u}_{t-N}] = \boldsymbol{\rho}_t,$$

and, hence, by the uniqueness of the solutions of the Bellman equations, we have, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{q}}_t(x,a)|\mathbf{u}_{t-N}] &= \mathbf{q}_t(x,a), \\ \mathbb{E}[\hat{\mathbf{v}}_t(x)|\mathbf{u}_{t-N}] &= \mathbf{v}_t(x). \end{aligned} \tag{9}$$

As a consequence, we also have, for all $(x,a) \in \mathcal{X} \times \mathcal{A}, t \geq N+1$,

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\rho}}_t] &= \mathbb{E}\left[\boldsymbol{\rho}_t\right], \\ \mathbb{E}[\hat{\mathbf{q}}_t(x,a)] &= \mathbb{E}\left[\mathbf{q}_t(x,a)\right], \\ \mathbb{E}[\hat{\mathbf{v}}_t(x)] &= \mathbb{E}\left[\mathbf{v}_t(x)\right]. \end{aligned} \tag{10}$$

The bandit algorithm that we propose is shown as Algorithm 1. It follows the approach of Even-Dar et al. (2009) in that a bandit algorithm is used in each state which together determine the policy to be used. These bandit algorithms are fed with estimates of action-values for the current policy and the current reward. In our case these action-value estimates are $\hat{\mathbf{q}}_t$ defined earlier, which are based on the reward estimates $\hat{\mathbf{r}}_t$. A major difference is that the policy computed based on the most recent action-value estimates is used only N steps later. This delay allows us to construct unbiased estimates of the rewards. Its price is that we need to store N policies (or weights, leading to the policies), thus, the memory needed by our algorithm scales with $N\,|\mathcal{A}||\mathcal{X}|$. 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 $\boldsymbol{\mu}_t^N(\cdot|\mathbf{x}_{t-N},\mathbf{a}_{t-N})$). The cost of this is $\mathcal{O}\left(N|\mathcal{A}||\mathcal{X}|^3\right)$. In addition to the need of dealing with the delay, we also need to deal with the fact that in our case $\mathbf{q}_t$ and $\hat{\mathbf{q}}_t$ can be both negative, which must be taken into account in the proper tuning of the algorithm's parameters.

4.4 Main result

Our main result is the following bound concerning the performance of Algorithm 1.

Theorem 1. *Let* $N = \lceil \tau \ln T \rceil$,

$$\eta = T^{-2/3} \cdot (\ln|\mathcal{A}|)^{2/3} \cdot \left(\frac{4\tau+8}{\beta}\left((2\tau+4)\tau|\mathcal{A}|\ln T + (3\tau+1)^2\right)\right)^{-1/3},$$

$$\gamma = T^{-1/3} \cdot (2\tau+4)^{-2/3} \cdot \left(\frac{2\ln|\mathcal{A}|}{\beta}\left((2\tau+4)\tau|\mathcal{A}|\ln T + (3\tau+1)^2\right)\right)^{1/3}.$$

Then the regret can be bounded as

$$\hat{L}_T \leq 3\,T^{2/3} \cdot \left(\frac{(4\tau+8)\ln|\mathcal{A}|}{\beta}\left((2\tau+4)\tau|\mathcal{A}|\ln T + (3\tau+1)^2\right)\right)^{1/3} + \mathcal{O}\left(T^{1/3}\right).$$

Algorithm 1 Algorithm for the online bandit MDP.

Set $N \geq 1$, $\mathbf{w}_1(x,a) = \mathbf{w}_2(x,a) = \cdots = \mathbf{w}_{2N}(x,a) = 1$, $\gamma \in (0,1)$, $\eta \in (0,\gamma]$.
For $t = 1, 2, \ldots, T$, 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 $\mathbf{a}_t$ randomly, according to the policy $\boldsymbol{\pi}_t(\cdot|\mathbf{x}_t)$.
3. Receive reward $r_t(\mathbf{x}_t, \mathbf{a}_t)$ and observe $\mathbf{x}_{t+1}$.
4. If $t \geq N+1$
   (a) Compute $\boldsymbol{\mu}_t^N(x|\mathbf{x}_{t-N}, \mathbf{a}_{t-N})$ for all $x \in \mathcal{X}$ using (7).
   (b) Construct estimates $\hat{\mathbf{r}}_t$ using (3) and compute $\hat{\mathbf{q}}_t$ using (4).
   (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 $(x,a) \in \mathcal{X} \times \mathcal{A}$.

It is interesting to note that, similarly to the regret bound of Even-Dar et al. (2009), the main term of the regret bound does not directly depend on the size of the state space, but it depends on it only through β and the mixing time τ, defined in Assumptions A2 and A3, respectively; however, we also need to note that $\beta > 1/|\mathcal{X}|$. While the theorem provides the first rigorously proved finite sample regret bound for the online bandit MDP problem, we suspect that the given convergence rate is not sharp in the sense that it may be possible, in agreement with the standard bandit lower bound of Auer et al. (2002), to give an algorithm with an $\mathcal{O}\left(\sqrt{T}\right)$ regret (up to some logarithmic factors).

The proof of the theorem is similar to the proof of a similar bound done for the full-information case by Even-Dar et al. (2009). We use the decomposition of the regret given in (1). Since the first term is bounded by $2(\tau+1)$ as before (Lemma 1), it remains to bound the expectation of the other terms. This is done in the following two propositions.

Proposition 1. *Let* $c = \frac{2\eta}{\beta}\left(\frac{1}{\gamma} + 4\tau + 6\right)$ *and assume that*
$$c(3\tau+1)^2 < \beta/2, \tag{11}$$
$$N \geq \max\left\{\tau \ln\left(\frac{4}{\beta - 2c(3\tau+1)^2}\right), \tau \ln T\right\}, \tag{12}$$
and
$$0 < \eta < \frac{\beta}{2(1/\gamma + 2\tau + 3)}$$
Then, for any policy π*, we have*
$$\sum_{t=1}^{T} \mathbb{E}\left[\rho_t^{\pi} - \boldsymbol{\rho}_t\right]$$
$$\leq (4\tau+10)N + \frac{\ln|\mathcal{A}|}{\eta} + (2\tau+4)\,T\left(\gamma + \frac{2\eta}{\beta}|\mathcal{A}|\Big(N\left(1/\gamma + 4\tau + 6\right) + (e-2)(2\tau+4)\Big)\right).$$

Proposition 2. *Assume that* (11) *and* (12) *hold. Then*
$$\sum_{t=1}^{T} \mathbb{E}\left[\boldsymbol{\rho}_t\right] - \hat{R}_T \leq T\,\frac{2\eta}{\beta}\left(\frac{1}{\gamma} + 4\tau + 6\right)(3\tau+1)^2 + 2Te^{-N/\tau} + 2N. \tag{13}$$

Note that setting
$$N \geq \tau \ln T,$$
the second term in (13) becomes $\mathcal{O}(1)$. Also, if T is large enough, the choices of N, η and γ in Theorem 1 will satisfy the conditions of Proposition 1. That the conclusion of the theorem holds can be verified by plugging in the definitions of N, η and γ in the bounds of the two propositions.

The proofs are broken into a number of statements presented in the next section.

5 Analysis

5.1 General tools

We proceed with a series of lemmas to control the rate of change of the policies generated by **Exp3**.

Lemma 2. *Pick any policy π. If $|\sum_a \pi(a|x) r(x,a)| \le R$ 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 $(x,a) \in \mathcal{X} \times \mathcal{A}$, $|q^\pi(x,a)| \le 2R(\tau+1) + R + |r(x,a)|$.*

Proof. As it is well known, the differential value of policy π at state x can be written as

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

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

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

where in the second inequality we used $\|\nu_{s,x}^\pi - \mu^\pi\|_1 \le 2e^{-(s-1)/\tau}$ 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 second inequality follows from the first part and the Bellman equation:

$$|q^\pi(x,a)| \le |r(x,a)| + |\rho^\pi| + \sum_{x'} P(x'|x,a) |v^\pi(x')| \le |r(x,a)| + R + 2R(\tau+1).$$

Here, we used that $\rho^\pi = \sum_x \mu^\pi(x) \sum_a \pi(a|x) r(x,a)$ and the bound on $\sum_a \pi(a|x) r(x,a)$. □

Lemma 3. *Let $N < t \le T$ and assume that $\boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x') \ge \beta/2$ holds for all states x'. Then, for any $x \in \mathcal{X}$, we have*

$$|\hat{\mathbf{v}}_t(x)| \le \frac{4\tau+4}{\beta}\ .$$

Proof. The result follows because $\hat{\mathbf{v}}_t = v^{\boldsymbol{\pi}_t}$ and thus one can apply Lemma 2. The proof is finished by noting that $|\sum_a \boldsymbol{\pi}_t(a|x) \hat{\mathbf{r}}_t(x,a)| \le \frac{1}{\boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x')} \le \frac{2}{\beta}$, by assumption. □

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 4. *Let $N < t \le T$ and assume that $\boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x') > 0$ 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 2 and then taking expectations, we get

$$\mathbb{E}\left[|\hat{\mathbf{v}}_t(x)| \,|\, \mathbf{u}_{t-N}\right] \le \sum_{s=1}^{\infty} \sum_{x'} |\nu_{s,x}^\pi(x') - \mu^\pi(x')|\, \mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x') \hat{\mathbf{r}}_t(x',a) \,\middle|\, \mathbf{u}_{t-N} \right],$$

where we have exploited that $\hat{\mathbf{r}}_t$ takes only nonnegative values. Now, by (6) and (8),

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

which is bounded between 0 and 1. Hence,

$$\mathbb{E}\left[|\hat{\mathbf{v}}_t(x)| \,|\, \mathbf{u}_{t-N}\right] \le \sum_{s=1}^{\infty} \sum_{x'} |\nu_{s,x}^\pi(x') - \mu^\pi(x')|.$$

Now, finishing as before we get the statement. □

We shall also need a bound on the expected value of $\mathbb{E}\left[|\hat{\mathbf{q}}_t(x,a)| \,|\, \mathbf{u}_{t-N}\right]$. This is bounded as follows:

Lemma 5. *Let $N < t \le T$ and assume that $\boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x') > 0$ holds for all states x'. Then, for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have*

$$\mathbb{E}\left[|\hat{\mathbf{q}}_t(x,a)| \,|\, \mathbf{u}_{t-N}\right] \le 2(\tau+2)\ .$$

Proof. By the Bellman equations (4),

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

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 desired statement. □

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 6. *Let $N < t \le T$ and assume that $\boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x') > \beta/2$ 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 assumption and the construction of $\hat{\mathbf{r}}_t(x,a)$,

$$\boldsymbol{\pi}_t(x,a)|\hat{\mathbf{r}}_t(x,a)| \le \frac{2}{\beta}. \tag{14}$$

Thus, we can apply Lemma 2 with $R = 2/\beta$ to obtain $|\hat{\mathbf{q}}_t(x,a)| \le \frac{2}{\beta}\left(2(\tau+1)+1\right) + |\hat{\mathbf{r}}_t(x,a)|$. Multiplying both sides by $\boldsymbol{\pi}_t(x,a)$ and using (14) again finishes the proof. □

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

Lemma 7. *Let $1 \le N < t \le T$ and $c > 0$ be such that $\max_x \sum_a |\boldsymbol{\pi}_{s+1}(a|x) - \boldsymbol{\pi}_s(a|x)| \le c$ holds for $1 \le s \le t-1$. Then we have*

$$\max_{x,a} \sum_{x'} \left| \boldsymbol{\mu}^N_{t,x,a}(x') - \mu^{\boldsymbol{\pi}_t}(x') \right| \le c\,(3\tau+1)^2 + 2e^{-N/\tau}.$$

Proof. The proof is similar to that of Lemma 5.2 of Even-Dar et al. (2009). Again, we will treat distributions as row vectors. In particular, $\boldsymbol{\mu}^N_{t,x,a}$ will denote the row vector whose ${x'}^{\text{th}}$ element is $\boldsymbol{\mu}^N_{t,x,a}(x')$. First, note that

$$\begin{aligned}
\|\boldsymbol{\mu}^N_{t,x} - \mu^{\boldsymbol{\pi}_t}\|_1 &= \|\boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_{t-1}} - \boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_t} + \boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_t} - \mu^{\boldsymbol{\pi}_t}\|_1 \\
&\le \|\boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_{t-1}} - \boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_t}\|_1 + \|\boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_t} - \mu^{\boldsymbol{\pi}_t}P^{\boldsymbol{\pi}_t}\|_1 \\
&\le c + e^{-1/\tau}\|\boldsymbol{\mu}^{N-1}_{t-1,x} - \mu^{\boldsymbol{\pi}_t}\|_1.
\end{aligned}$$

Here we used that

$$\|\boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_{t-1}} - \boldsymbol{\mu}^{N-1}_{t-1,x}P^{\boldsymbol{\pi}_t}\|_1 \le \max_x \sum_a |\boldsymbol{\pi}_{t-1}(a|x) - \boldsymbol{\pi}_t(a|x)| \le c$$

(this follows by elementary algebra; for the actual proof see Lemma 5.1 in Even-Dar et al., 2009). Similarly, we get

$$\|\boldsymbol{\mu}^{N-k}_{t-k,x} - \mu^{\boldsymbol{\pi}_t}\|_1 \le (k+1)\,c + e^{-1/\tau}\|\boldsymbol{\mu}^{N-(k+1)}_{t-(k+1),x} - \mu^{\boldsymbol{\pi}_t}\|_1, \quad 1 \le k \le N.$$

Thus,

$$\|\boldsymbol{\mu}^N_{t,x} - \mu^{\boldsymbol{\pi}_t}\|_1 \le c(1 + 2e^{-1/\tau} + 3e^{-2/\tau} + \ldots + Ne^{-(N-1)/\tau}) + 2e^{-N/\tau}.$$

Now, $S_N \stackrel{\text{def}}{=} 1+2e^{-1/\tau}+3e^{-2/\tau}+\ldots+Ne^{-(N-1)/\tau} \le e^{1/\tau}\left(e^{-1/\tau}+2e^{-2/\tau}+3e^{-3/\tau}+\ldots\right)$. We bound the term in the bracket as follows: First, assume that $\tau \ge 1$ and let $n = \lfloor \tau \rfloor$. Then $t\exp(-t/\tau)$ is increasing up to n and it is decreasing from $t = n+1$. Hence,

$$\begin{aligned} e^{-1/\tau}+2e^{-2/\tau}+3e^{-3/\tau}+\ldots &\le e^{-1/\tau}+2e^{-2/\tau}+3e^{-3/\tau}+\ldots+(n+1)e^{-(n+1)/\tau} \\ &\quad + (n+2)e^{-(n+2)/\tau}+(n+3)e^{-(n+3)/\tau}+\ldots \\ &\le \frac{(n+1)(n+2)}{2e^{1/\tau}} + \int_{n+1}^{\infty} te^{-t/\tau}dt \\ &\le \tfrac{3}{e^{1/\tau}}\,\tau^2 + \int_0^{\infty} te^{-t/\tau}dt \\ &\le \left(\tfrac{3}{e^{1/\tau}}+1\right)\tau^2. \end{aligned}$$

Therefore, we get $S_N \le (3+e)\tau^2 < 6\tau^2$. When $\tau < 1$, $S_N \le 1 + e^{1/\tau}\int_1^\infty t\exp(-t/\tau)dt \le 1 + e^{1/\tau}e^{-1/\tau}\tau(\tau+1) = 1+\tau(\tau+1) = 1+\tau+\tau^2$. Hence, $S_N \le 1+\tau+6\tau^2 \le (1+3\tau)^2$.

Plugging in this bound into the bound derived for $\|\boldsymbol{\mu}_{t,x}^N - \mu^{\boldsymbol{\pi}_t}\|_1$ gives the desired result. □

In the next two lemmas we compute the rate of change of the policies produced by **Exp3** and show that for a large enough value of N, $\boldsymbol{\mu}_{t,x,a}^N$ can be uniformly bounded form below by $\beta/2$.

Lemma 8. *Assume that for some $N+1 \le t \le T$, $\boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x') \ge \beta/2$ holds for all states x'. Let*

$$c = \frac{2\eta}{\beta}\left(\frac{1}{\gamma}+4\tau+6\right).$$

Then,

$$\max_x \sum_a |\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)| \le c. \tag{15}$$

Proof. We have $\hat{\boldsymbol{\rho}}_t = \sum_{x,a}\mu^{\boldsymbol{\pi}_t}(x)\,\boldsymbol{\pi}_t(a|x)\mathbf{r}_t(x,a)$ and thus from the definition of $\mathbf{r}_t(x,a)$ we get $\hat{\boldsymbol{\rho}}_t \le 2/\beta$. Hence, by Lemma 3 and the Bellman equations,

$$\hat{\mathbf{q}}_t(x,a) \ge -\frac{2}{\beta} - \frac{4\tau+4}{\beta} = -\frac{2}{\beta}\,(2\tau+3). \tag{16}$$

Using $\mathbf{r}_t(x,a) \le \frac{2}{\gamma\beta}$ and $\hat{\boldsymbol{\rho}}_t \ge 0$, we also get

$$\hat{\mathbf{q}}_t(x,a) \le \frac{2}{\gamma\beta} + \frac{4\tau+6}{\beta} = \frac{2}{\beta}\left(\frac{1}{\gamma}+(2\tau+3)\right). \tag{17}$$

Hence, $\eta\,\hat{\mathbf{q}}_t(x,a)$ is contained in an interval of length at most $\frac{2\eta}{\beta}\left(\frac{1}{\gamma}+4\tau+6\right) = c$. Define

$$\boldsymbol{\pi}'_s(a|x) = \frac{\mathbf{w}_s(x,a)}{\sum_b \mathbf{w}_s(x,b)}$$

for $s \in \{t+N-1, t+N\}$. Clearly, $|\boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_{t+N}(a|x)| \le |\boldsymbol{\pi}'_{t+N-1}(a|x) - \boldsymbol{\pi}'_{t+N}(a|x)|$ for all x and a, and hence it is sufficient to bound $\max_x \sum_a |\boldsymbol{\pi}'_{t+N-1}(a|x) - \boldsymbol{\pi}'_{t+N}(a|x)|$. First we bound the Kullback-Leibler divergence $D(\boldsymbol{\pi}'_{t+N-1}(\cdot|x)||\boldsymbol{\pi}'_{t+N}(\cdot|x))$ between the action distributions resulting from $\boldsymbol{\pi}'_{t+N-1}$ and $\boldsymbol{\pi}'_{t+N}$ when these policies are used in state x:

$$
\begin{aligned}
& D(\boldsymbol{\pi}'_{t+N-1}(\cdot|x)||\boldsymbol{\pi}'_{t+N}(\cdot|x)) \\
&= \sum_a \boldsymbol{\pi}'_{t+N-1}(a|x) \ln \frac{\boldsymbol{\pi}'_{t+N-1}(a|x)}{\boldsymbol{\pi}'_{t+N-1}(a|x) e^{\eta \hat{\mathbf{q}}_t(x,a)} / \sum_b \boldsymbol{\pi}'_{t+N-1}(b|x) e^{\eta \hat{\mathbf{q}}_t(x,b)}} \\
&= \sum_a \boldsymbol{\pi}'_{t+N-1}(a|x) \ln \frac{\sum_b \boldsymbol{\pi}'_{t+N-1}(b|x) e^{\eta \hat{\mathbf{q}}_t(x,b)}}{e^{\eta \hat{\mathbf{q}}_t(x,a)}} \\
&= \ln \sum_b \boldsymbol{\pi}'_{t+N-1}(b|x) e^{\eta \hat{\mathbf{q}}_t(x,b)} - \eta \sum_a \boldsymbol{\pi}'_{t+N-1}(a|x) \hat{\mathbf{q}}_t(x,a) \\
&\le \eta \sum_a \boldsymbol{\pi}'_{t+N-1}(a|x) \hat{\mathbf{q}}_t(x,a) + \frac{c^2}{8} - \eta \sum_a \boldsymbol{\pi}'_{t+N-1}(a|x) \hat{\mathbf{q}}_t(x,a) \\
&= \frac{c^2}{8},
\end{aligned}
\tag{18}
$$

where we have used Hoeffding's Lemma (cf. Lemma 2.2 in Cesa-Bianchi and Lugosi, 2006) and that $\hat{\mathbf{q}}_t(x,a)$ is contained in an interval of length at most c. Now, by Pinsker's inequality (see, e.g., Cesa-Bianchi and Lugosi, 2006), we have

$$
\sum_a |\boldsymbol{\pi}'_{t+N-1}(a|x) - \boldsymbol{\pi}'_{t+N}(a|x)| \le \sqrt{8 D(\boldsymbol{\pi}'_{t+N-1}(\cdot|x)||\boldsymbol{\pi}'_{t+N}(\cdot|x))} \le c.
$$

□

Lemma 9. *Let c be as in Lemma 8. Assume that $c(3\tau+1)^2 < \beta/2$, and let*

$$
N \ge \left\lceil \tau \ln \left(\frac{4}{\beta - 2c(3\tau+1)^2} \right) \right\rceil . \tag{19}
$$

Then, for all $N < t \le T$, $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$, we have $\boldsymbol{\mu}^N_{t,x,a}(x') \ge \beta/2$ and $\max_{x'} \sum_{a'} |\boldsymbol{\pi}_{t+1}(a'|x') - \boldsymbol{\pi}_t(a'|x')| \le c$.

Proof. We prove the lemma by induction on t. The induction hypothesis is that for $N+1 \le t \le T$, $\min_{x,x',a} \boldsymbol{\mu}^N_{s,x,a}(x') \ge \beta/2$ and $\max_{x'} \sum_{a'} |\boldsymbol{\pi}_{s+1}(a'|x') - \boldsymbol{\pi}_s(a'|x')| \le c$ hold for all $N+1 \le s \le t$.

Let us first show that this hypothesis holds when $N+1 \le t \le 2N-1$. Fix some state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$. 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 $1 \le t \le 2N-1$. Thus, by Lemma 7, we get that

$$
\|\boldsymbol{\mu}^N_{t,x,a} - \mu^{\boldsymbol{\pi}_t}\|_1 \le c(3\tau+1)^2 + 2e^{-N/\tau}
$$

holds for all $N+1 \le t \le 2N-1$. By our assumption about N, we have

$$
c(3\tau+1)^2 + 2e^{-N/\tau} \le \beta/2, \tag{20}
$$

thus for any $N+1 \le t \le 2N-1$, $x \in \mathcal{X}$, $a \in \mathcal{A}$,

$$
\|\boldsymbol{\mu}^N_{t,x,a} - \mu^{\boldsymbol{\pi}_t}\|_\infty \le \|\boldsymbol{\mu}^N_{t,x,a} - \mu^{\boldsymbol{\pi}_t}\|_1 \le \beta/2. \tag{21}
$$

Since, by assumption, $\mu^\pi(x') \ge \beta$ holds for any stationary policy π, we also have $\mu^{\boldsymbol{\pi}_t}(x') \ge \beta$. This, together with (21) gives that $\boldsymbol{\mu}^N_{t,x,a}(x') \ge \beta/2$ holds for any $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$.

Now, fix a time index $2N \le t \le T$ and assume that the induction hypothesis holds for time $t-1$. Then, thanks to $\min_{x,x',a} \boldsymbol{\mu}^N_{t-N+1,x,a}(x') \ge \beta/2$, Lemma 8 implies

$$
\max_{x'} \sum_{a'} |\boldsymbol{\pi}_{t+1}(a'|x') - \boldsymbol{\pi}_t(a'|x')| \le c.
$$

Now, by Lemma 7, we have for any $x \in \mathcal{X}$, $a \in \mathcal{A}$,

$$
\|\boldsymbol{\mu}^N_{t,x,a} - \mu^{\boldsymbol{\pi}_t}\|_1 \le c(3\tau+1)^2 + 2e^{-N/\tau}.
$$

Using the same reasoning as above, we finish the inductive step and thus the proof. □

5.2 Proof of Proposition 1

The statement is trivial for $T \leq N$. The following simple result is the first step in proving Proposition 1 for $T > N$.

Lemma 10. *(cf. Lemma 4.1 in Even-Dar et al., 2009) For any policy π and $t \geq 1$,*

$$\rho_t^\pi - \boldsymbol{\rho}_t = \sum_{x,a} \mu^\pi(x)\pi(a|x)\left[\mathbf{q}_t(x,a) - \mathbf{v}_t(x)\right] \ .$$

Proof. We have

$$\begin{aligned}\sum_{x,a} \mu^\pi(x)\pi(a|x)\mathbf{q}_t(x,a) &= \sum_{x,a} \mu^\pi(x)\pi(a|x)\left[r_t(x,a) - \boldsymbol{\rho}_t + \sum_{x'} P(x'|x,a)\mathbf{v}_t(x')\right] \\ &= \rho_t^\pi - \boldsymbol{\rho}_t + \sum_x \mu^\pi(x)\mathbf{v}_t(x).\end{aligned}$$

Reordering the terms gives the result. □

For every x, a define $\mathbf{Q}_T(x,a) = \sum_{t=N+1}^T \mathbf{q}_t(x,a)$ and $\mathbf{V}_T(x) = \sum_{t=N+1}^T \mathbf{v}_t(x)$. The preceding lemma 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 11. *Let c be as in Lemma 8. Assume that $\gamma \in (0,1)$, $c(3\tau+1)^2 < \beta/2$, $N \geq \left\lceil \tau \ln\left(\frac{4}{\beta - 2c(3\tau+1)^2}\right)\right\rceil$, $0 < \eta \leq \frac{\beta}{2(1/\gamma + 2\tau + 3)}$, and $T > N$ hold. Then, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$\begin{aligned}&\mathbb{E}\left[\mathbf{Q}_T(x,a) - \mathbf{V}_T(x)\right] \\ &\quad\leq (4\tau+8)N + \frac{\ln|\mathcal{A}|}{\eta} + (2\tau+4)\,T\left(\gamma + \frac{2\eta}{\beta}|\mathcal{A}|\Big(N\left(1/\gamma + 4\tau + 6\right) + (e-2)(2\tau+4)\Big)\right).\end{aligned}$$

The proof is presented in Appendix A.

Proof of Proposition 1. Under the conditions of the proposition, combining Lemmas 10-11 yields

$$\begin{aligned}&\sum_{t=1}^T \mathbb{E}\left[\rho_t^\pi - \boldsymbol{\rho}_t\right] \\ &\quad\leq 2N + \sum_{x,a} \mu^\pi(x)\pi(a|x)\,\mathbb{E}\left[\mathbf{Q}_T(x,a) - \mathbf{V}_T(x)\right] \\ &\quad\leq (4\tau+10)N + \frac{\ln|\mathcal{A}|}{\eta} + (2\tau+4)\,T\left(\gamma + \frac{2\eta}{\beta}|\mathcal{A}|\Big(N\left(1/\gamma + 4\tau + 6\right) + (e-2)(2\tau+4)\Big)\right),\end{aligned}$$

proving Proposition 1. □

5.3 Proof of Proposition 2

Let $t > N$. First, since $\boldsymbol{\pi}_t$ is $\sigma(\mathbf{u}_{t-N})$-measurable,

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

We also have

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

Hence,

$$\mathbb{E}\left[\boldsymbol{\rho}_t - r_t(\mathbf{x}_t, \mathbf{a}_t)\right] = \mathbb{E}\left[\sum_x (\mu^{\boldsymbol{\pi}_t}(x) - \boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x))\, \mathbb{E}\left[\, r_t(x, \mathbf{a}_t) \middle|\, \mathbf{u}_{t-N}\right]\right]$$

$$\leq \mathbb{E}\left[\sum_x \left|\mu^{\boldsymbol{\pi}_t}(x) - \boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x)\right|\right],$$

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

Thanks to Lemma 9, Lemma 7 is applicable. Hence,

$$\sum_x \left|\mu^{\boldsymbol{\pi}_t}(x) - \boldsymbol{\mu}^N_{t,\mathbf{x}_{t-N},\mathbf{a}_{t-N}}(x)\right| \leq c(3\tau + 1)^2 + 2e^{-N/\tau},$$

and thus $\mathbb{E}\left[\boldsymbol{\rho}_t - r_t(\mathbf{x}_t, \mathbf{a}_t)\right] \leq c(3\tau + 1)^2 + 2e^{-N/\tau}$. Summing up these inequalities for $t = N + 1, \ldots, T$, we get

$$\left(\sum_{t=N+1}^{T} \mathbb{E}\left[\boldsymbol{\rho}_t\right]\right) - \hat{R}_T \leq T\, c(3\tau + 1)^2 + 2Te^{-N/\tau}.$$

Using the trivial bound $\mathbb{E}\left[\boldsymbol{\rho}_t - r_t(\mathbf{x}_t, \mathbf{a}_t)\right] \leq 2$ for the first N terms, we get the desired result. □

A Proof of Lemma 11

We follow the steps of the proof in Auer et al. (2002). Fix $(x, a) \in \mathcal{X} \times \mathcal{A}$. For any $1 \leq t \leq T$ define $\mathbf{W}_t(x) = \sum_a \mathbf{w}_t(x, a)$. First, note that since the conditions of Lemma 9 are satisfied, hence, the conclusions of this lemma, as well as those of Lemmas 3–6 hold. In particular, by Lemma 6, $\boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x, a)| \leq B_q \stackrel{\text{def}}{=} \frac{4}{\beta}(\tau + 2)$ holds for any $s \geq N + 1$. Now, using Lemma 2 and the Bellman equations, we get that $\hat{\mathbf{q}}_t(x, a) \leq \frac{2}{\beta}(1/\gamma + 2\tau + 3)$, thus by the constraint on η, $\eta\hat{\mathbf{q}}_t(x, a) \leq 1$.

Fix $2N \leq t \leq T - 1$. We have the following:

$$\begin{aligned}
\frac{\mathbf{W}_{t+1}(x)}{\mathbf{W}_t(x)} &= \sum_a \frac{\mathbf{w}_{t+1}(x, a)}{\mathbf{W}_t(x)} \\
&= \sum_a \frac{\mathbf{w}_t(x, a)}{\mathbf{W}_t(x)} e^{\eta\hat{\mathbf{q}}_{t-N+1}(x,a)} \\
&= \sum_a \frac{\boldsymbol{\pi}_t(a|x) - \gamma/|\mathcal{A}|}{1 - \gamma} e^{\eta\hat{\mathbf{q}}_{t-N+1}(x,a)} \\
&\leq \sum_a \frac{\boldsymbol{\pi}_t(a|x) - \gamma/|\mathcal{A}|}{1 - \gamma} \left(1 + \eta\hat{\mathbf{q}}_{t-N+1}(x, a) + (e - 2)\left(\eta\hat{\mathbf{q}}_{t-N+1}(x, a)\right)^2\right) \\
&\qquad \text{(as } \eta\hat{\mathbf{q}}_{t-N+1}(x, a) \leq 1\text{)} \\
&\leq 1 + \frac{\eta}{1 - \gamma}\sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_{t-N+1}(x, a) + \frac{\eta^2(e - 2)}{1 - \gamma}\sum_a \boldsymbol{\pi}_t(a|x)(\hat{\mathbf{q}}_{t-N+1}(x, a))^2 \\
&\leq 1 + \frac{\eta}{1 - \gamma}\hat{\mathbf{v}}^N_t(x) + \frac{B_q\, \eta^2(e - 2)}{(1 - \gamma)}\sum_a |\hat{\mathbf{q}}_{t-N+1}(x, a)|,
\end{aligned}$$

where $\hat{\mathbf{v}}^N_t(x) = \sum_a \boldsymbol{\pi}_t(a|x)\hat{\mathbf{q}}_{t-N+1}(x, a)$. Using $1 + x \leq e^x$ and then taking logarithms gives

$$\ln\frac{\mathbf{W}_{t+1}(x)}{\mathbf{W}_t(x)} \leq \frac{\eta}{1 - \gamma}\hat{\mathbf{v}}^N_t(x) + \frac{B_q\, \eta^2(e - 2)}{(1 - \gamma)}\sum_a |\hat{\mathbf{q}}_{t-N+1}(x, a)|.$$

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

$$\ln\frac{\mathbf{W}_{T+1}(x)}{\mathbf{W}_{2N}(x)} \leq \frac{\eta}{1 - \gamma}\hat{\mathbf{V}}^N_T(x) + \frac{B_q\, \eta^2(e - 2)}{(1 - \gamma)}\sum_{t=N+1}^{T-N+1}\sum_a |\hat{\mathbf{q}}_t(x, a)| \tag{22}$$

with $\hat{\mathbf{V}}_T^N(x) = \sum_{t=2N}^T \hat{\mathbf{v}}_t^N(x)$. On the other hand, for any action b we have

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

where we used that $\mathbf{w}_{2N}(x,a) = 1$ holds for all $a \in \mathcal{A}$. Combining with (22), we get

$$\hat{\mathbf{V}}_T^N(x) \geq (1-\gamma)\hat{\mathbf{Q}}_T^N(x,b) - \frac{\ln |\mathcal{A}|}{\eta} - B_q\, \eta(e-2) \sum_{t=N+1}^{T-N+1} \sum_a |\hat{\mathbf{q}}_t(x,a)| \tag{23}$$

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

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

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

Note that

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

Therefore,

$$\begin{aligned}
&\hat{\mathbf{V}}_T^N(x) - \hat{\mathbf{V}}_T(x) \\
&\leq \sum_{t=N+1}^{T-N+1} \sum_a |\hat{\mathbf{q}}_t(x,a)| \Big| \boldsymbol{\pi}_{t+N-1}(a|x) - \boldsymbol{\pi}_t(a|x) \Big| + \sum_{t=T-N+2}^{T} \sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)| \\
&\leq \sum_{t=N+1}^{T-N+1} \|\, \boldsymbol{\pi}_{t+N-1}(\cdot|x) - \boldsymbol{\pi}_t(\cdot|x) \,\|_1 \; \|\, \hat{\mathbf{q}}_t(x,\cdot) \,\|_\infty + \sum_{t=T-N+2}^{T} \sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)| \\
&\leq N\, c \sum_{t=N+1}^{T-N+1} \|\, \hat{\mathbf{q}}_t(x,\cdot) \,\|_\infty + \sum_{t=T-N+2}^{T} \sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|,
\end{aligned}$$

where we have used that by Lemma 8, $\|\, \boldsymbol{\pi}_{t+N-1}(\cdot|x) - \boldsymbol{\pi}_t(\cdot|x) \,\|_1 \leq Nc$, that is, the policies change slowly. Taking the expectation of both sides, we get

$$\mathbb{E}\left[\hat{\mathbf{V}}_T^N(x)\right] - \mathbb{E}\left[\hat{\mathbf{V}}_T(x)\right] \leq N\, c \sum_{t=N+1}^{T-N+1} \mathbb{E}\left[\|\, \hat{\mathbf{q}}_t(x,\cdot) \,\|_\infty\right] + \sum_{t=T-N+2}^{T} \mathbb{E}\left[\sum_a \boldsymbol{\pi}_t(a|x)|\hat{\mathbf{q}}_t(x,a)|\right].$$

By Lemma 5,

$$\mathbb{E}\left[|\hat{\mathbf{q}}_t(x,a)|\right] \leq B'_q \stackrel{\text{def}}{=} 2(\tau+2) \tag{24}$$

holds for any $(x,a) \in \mathcal{X} \times \mathcal{A}$. Therefore, $\mathbb{E}\left[\|\, \hat{\mathbf{q}}_t(x,\cdot) \,\|_\infty\right] \leq |\mathcal{A}| B'_q$.

Using again (24),

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

Hence,

$$\mathbb{E}\left[\hat{\mathbf{V}}_T^N(x)\right] \leq \mathbb{E}\left[\hat{\mathbf{V}}_T(x)\right] + N\, B'_q\, (c\, T|\mathcal{A}| + 1).$$

This, together with (23) gives

$$\mathbb{E}\left[\hat{\mathbf{V}}_T(x)\right]+N\,B'_q\,(c\,T|\mathcal{A}|+1) \geq (1-\gamma)\mathbb{E}\left[\hat{\mathbf{Q}}^N_T(x,b)\right]-\frac{\ln|\mathcal{A}|}{\eta}-B_q\,\eta(e-2)\sum_{t=N+1}^{T-N+1}\sum_a \mathbb{E}\left[|\hat{\mathbf{q}}_t(x,a)|\right]. \tag{25}$$

By equation (10), we have

$$\mathbb{E}\left[\hat{\mathbf{V}}_T(x)\right] = \mathbb{E}\left[\mathbf{V}_T(x)\right]$$

and with the definition $\mathbf{Q}^N_T(x,b) = \sum_{t=N+1}^{T-N+1} \mathbf{q}_t(x,b)$, we also have

$$\mathbb{E}\left[\hat{\mathbf{Q}}^N_T(x,b)\right] = \mathbb{E}\left[\mathbf{Q}^N_T(x,b)\right].$$

Thus, using (24) again, we get

$$\mathbb{E}\left[\mathbf{V}_T(x)\right] + N\,B'_q\,(c\,T|\mathcal{A}|+1) \geq (1-\gamma)\mathbb{E}\left[\mathbf{Q}^N_T(x,b)\right] - \frac{\ln|\mathcal{A}|}{\eta} - \eta(e-2)\,B_q\,B'_q\,T|\mathcal{A}|.$$

By reordering the terms and noticing that $B_q = \frac{2}{\beta}B'_q$, we get

$$\mathbb{E}\left[\mathbf{Q}^N_T(x,b) - \mathbf{V}_T(x)\right] \leq \gamma\mathbb{E}\left[\mathbf{Q}^N_T(x,b)\right] + N\,B'_q\,(c\,T|\mathcal{A}|+1) + \frac{\ln|\mathcal{A}|}{\eta} + \frac{2}{\beta}\eta(e-2)\,B'^2_q\,T|\mathcal{A}|. \tag{26}$$

We now lower bound $\mathbf{Q}^N_T(x,a)$ by $\mathbf{Q}_T(x,a)$:

$$\mathbf{Q}_T(x,b) - \mathbf{Q}^N_T(x,b) = \sum_{t=T-N+2}^{T} \mathbf{q}_t(x,b) \leq B'_q N, \tag{27}$$

where we used that by Lemma 2,

$$\mathbf{q}_t(x,b) \leq B'_q = 2(\tau+2) \tag{28}$$

since the rewards are bounded between 0 and 1.

Combining (27) with (26), we obtain

$$\mathbb{E}\left[\mathbf{Q}_T(x,b) - \mathbf{V}_T(x)\right] \leq \gamma\mathbb{E}\left[\mathbf{Q}^N_T(x,b)\right]+N\,B'_q\,(c\,T|\mathcal{A}|+1)+\frac{\ln|\mathcal{A}|}{\eta}+B'_q\,\left(\frac{2}{\beta}\eta(e-2)\,B'_q\,T|\mathcal{A}|+N\right).$$

Using (28) again, we get $\mathbb{E}\left[\mathbf{Q}^N_T(x,b)\right] \leq B'_q T$. Thus,

$$\mathbb{E}\left[\mathbf{Q}_T(x,b) - \mathbf{V}_T(x)\right] \leq 2B'_q N + \frac{\ln|\mathcal{A}|}{\eta} + TB'_q\,\left(\gamma + N|\mathcal{A}|c + \frac{2}{\beta}\eta(e-2)\,B'_q\,|\mathcal{A}|\right).$$

Using that $c = \frac{2\eta}{\beta}\left(\frac{1}{\gamma}+4\tau+6\right)$ and $B'_q = 2(\tau+2)$, we arrive at the final result:

$$\begin{aligned}\mathbb{E}\left[\mathbf{Q}_T(x,b) - \mathbf{V}_T(x)\right] &\leq 2\,\cdot\,2(\tau+2)N + \frac{\ln|\mathcal{A}|}{\eta} \\ &+ 2(\tau+2)\,T\left(\gamma + N|\mathcal{A}|\frac{2\eta}{\beta}\left(\frac{1}{\gamma}+4\tau+6\right) + \eta(e-2)\,\frac{4}{\beta}\,(\tau+2)\,|\mathcal{A}|\right).\end{aligned}$$

□

Acknowledgments

This work was supported in part by the Hungarian Scientific Research Fund and the Hungarian National Office for Research and Technology (OTKA-NKTH 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

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

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

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.

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

Neu, G., György, A., and Szepesvári, C. (2010). The online loop-free stochastic shortest-path problem. In *COLT-10*.

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

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*. IEEE Press.

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.

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