No-Regret Learning with High-Probability in Adversarial Markov Decision Processes (Supplementary Material/Full Version)

Mahsa Ghasemi[*1,2] **Abolfazl Hashemi**[2] **Haris Vikalo**[1] **Ufuk Topcu**[2,3]

[1]Department of Electrical and Computer Engineering, University of Texas at Austin
[2]Oden Institute for Computational Engineering and Sciences, University of Texas at Austin
[3]Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin

Abstract

In a variety of problems, a decision-maker is unaware of the loss function associated with a task, yet it has to minimize this unknown loss in order to accomplish the task. Furthermore, the decision-maker's task may evolve, resulting in a varying loss function. In this setting, we explore sequential decision-making problems modeled by adversarial Markov decision processes, where the loss function may arbitrarily change at every time step. We consider the bandit feedback scenario, where the agent observes only the loss corresponding to its actions. We propose an algorithm, called *online relative-entropy policy search with implicit exploration*, that achieves a sublinear regret not only in expectation but, more importantly, with high probability. In particular, we prove that by employing an *optimistically biased* loss estimator, the proposed algorithm achieves a regret of $\tilde{\mathcal{O}}(T^{\frac{2}{3}}\sqrt{\tau|\mathcal{A}||\mathcal{S}|})$, where $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of actions, τ is the mixing time, and T is the time horizon. To our knowledge, the proposed algorithm is the first scheme that enjoys such high-probability regret bounds for general adversarial Markov decision processes under the presence of bandit feedback.

1 INTRODUCTION

A central notion in the analysis of online and sequential decision-making systems is that of Markov decision processes (MDPs). MDPs enable modeling decision-makers (learners) that need to make a sequence of decisions in the presence of uncertainty in the decision-maker's environment. In this scenario, a loss (or reward) function captures the task expected from the learner. Therefore, the decision-maker's goal is to design a learning algorithm that, despite operating under uncertainty, learns a policy with the lowest cumulative loss (or the highest cumulative reward). In a traditional MDP problem, one assumes that the environment's dynamics and the losses are stationary (i.e., time-invariant) throughout the time horizon of the interaction between the learner and the environment. However, the stationarity assumption does not hold in scenarios where the agent's task evolves over time.

The so-called adversarial MDP (A-MDP) [11] is a new paradigm that enables the study of sequential decision-making problems with evolving tasks. In particular, in an A-MDP, the environment's dynamics remain invariant while the loss function changes arbitrarily over time.[1] The learner aims to follow a policy that minimizes its loss in expectation over the time horizon. A standard metric for evaluating the learner's performance is regret, i.e., the difference between the learner's loss and the loss occurred by the best (stationary stochastic) policy in hindsight.

We focus on the setting of A-MDPs with bandit feedback, i.e., after each action, the learner observes only the corresponding loss but not the entire loss function. We consider the general class of uniformly ergodic adversarial MDPs for which the loss may change at every time step and provide the first no-regret algorithm that achieves a high-probability regret bound in this setting.

The state-of-the-art algorithms for learning in A-MDPs with bandit feedback guarantee sublinear regret of $\tilde{\mathcal{O}}(\sqrt{T})$ in expectation[2], where T denotes the time horizon. However, it remains a challenge to establish algorithms that attain sublinear regrets with high probability. Since in many practical settings, e.g., robotics and recommender systems, a learner may operate only once in an environment, a high-probability regret bound is more desirable than a bound in expectation. Nevertheless, high-probability guarantees are considered to be significantly more difficult to obtain than expected

[*]The first two authors contributed equally.

[1]We assume a setting where the adversarial changes in the loss are *oblivious* to the past actions taken by the learner.

[2]The notation $\tilde{\mathcal{O}}$ hides $\log(.)$ factors.

guarantees in online tasks with bandit feedback [21].

Contribution. We propose a learning algorithm for A-MDPs with bandit feedback that achieves a sublinear regret guarantee with high probability. We consider the linear programming formulation of MDPs where the decision variables are the occupancy measure of state-action pairs. This adaptation enables us to employ online mirror descent as a building block of our proposed scheme to learn low-regret occupancy measures. Furthermore, inspired by the idea of implicit exploration [16, 21] for adversarial bandits, we design a new *optimistically biased* estimator of the loss function for A-MDPs with bandit feedback to establish regret guarantees that hold with arbitrarily high probability. Our specific contributions are as follows:

- We propose a new online learning algorithm, called *online relative-entropy policy search with implicit exploration*, for A-MDPs under bandit feedback that employs a novel optimistically biased loss estimator.
- We design a novel optimistically biased loss estimator which implicitly promotes the learner to explore the action space and learn a sequence of randomized policies by relying on a variant of online mirror descent.
- We prove for uniformly ergodic MDPs that the proposed algorithm achieves a regret bound of $\tilde{\mathcal{O}}((T|\mathcal{A}||\mathcal{S}|)^{\frac{2}{3}}\sqrt{\tau})$, both in expectation and with high probability. Here, $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of actions, and τ is the mixing time of the MDP. To our knowledge, the proposed scheme is the first to achieve the above high probability regret bound for uniformly ergodic MDPs with bandit feedback.

2 RELATED WORK

A number of exact and approximate solutions to the problem of learning optimal policies in MDPs have been proposed in the literature. These include value iteration, policy iteration, and policy gradient techniques (see, e.g. [4, 5, 26] for a detailed discussion). Diverging from these methods, a linear programming (LP) approach has recently gained attention for MDPs with stationary loss functions [1, 28] as well as A-MDPs [31, 27, 6]. Our proposed algorithm relies on the LP formulation for computing an occupancy measure corresponding to the optimal policy. However, we propose a new loss estimator to deal with the bandit feedback setting.

Learning with A-MDPs can be categorized into two cases: episodic MDPs and uniformly ergodic MDPs where the latter is considered more general and challenging [24].

2.1 EPISODIC A-MDPS

In this setting, the loss may change from episode to episode. Jaksch et al. [13] introduced the UCRL-2 algorithm for MDPs with stochastic rewards under the setting of unknown transition functions and full information feedback. UCRL-2 keeps track of confidence sets that, with high probability, contain the true transition function and shrink over time. For episodes of length L, they showed a regret of $\tilde{\mathcal{O}}(L|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ compared to the optimal policy and provided a min-max lower bound, which can be achieved for a sufficiently large T (see the recent work of Azar et al. [3]). Non-adversarial episodic MDPs are studied in [7, 30] and recently [9] establish the state-of-the-art lowerbound for this setting. For the adversarial episodic MDP model, Neu et al. [22] proposed the *follow-the-perturbed-optimistic-policy* algorithm which relies on the follow-the-perturbed-leader method [15] and provides a regret of $\tilde{\mathcal{O}}(L|\mathcal{S}||\mathcal{A}|\sqrt{T})$. Recently, Rosenberg and Mansour [27] extended the results of Jaksch et al. [13] to MDPs with convex loss functions by employing online convex optimization and provided an algorithm achieving a regret of $\tilde{\mathcal{O}}(L|\mathcal{S}|\sqrt{|\mathcal{A}|T})$. However, these results still rely on the availability of full information feedback.

Under bandit feedback, Zimin and Neu [31] introduced the O-REPS algorithm, which employs the online mirror descent algorithm to achieve an expected regret of $\tilde{\mathcal{O}}(\sqrt{LT|\mathcal{A}||\mathcal{S}|})$. Similarly, Dick et al. [8] provided a regret of $\tilde{\mathcal{O}}(\sqrt{LT|\mathcal{S}||\mathcal{A}|})$ but for a computationally improved algorithm. Jin et al. [14] explored the use of an implicit exploration in the loss estimation for the case of *episodic A-MDPs* with unknown transition functions and obtained a high-probability regret of $\tilde{\mathcal{O}}(L|\mathcal{S}|\sqrt{T|\mathcal{A}|})$. In this work, we employ a different loss estimator with implicit exploration property for the general case of *ergodic A-MDPs* where the transition functions are known. We also analyze the regret in terms of both the expectation and high-probability bounds.

2.2 UNIFORMLY ERGODIC A-MDPS

Learning with uniformly ergodic MDPs in the adversarial setting is considerably more complicated than the episodic case. In the former, which is also the focus of this paper, the loss may change in every round, as opposed to the episodic setting where the loss in each episode is fixed. This difficulty renders the task of deriving learning algorithms with sublinear regrets more challenging.

Early work by Even-Dar et al. [11] on A-MDPs is on uniformly ergodic MDPs with known transition function and full information feedback. Considering each action to be an expert, the MDP-E algorithm in [11] employs the weighted majority method [19] in each state and achieves a regret of $\mathcal{O}(\tau^2\sqrt{T\log|\mathcal{A}|})$. Yu et al. [29] improved the computational efficiency utilizing the so-called *follow-the-perturbed-leader* method [15] with a regret of $\mathcal{O}(|\mathcal{A}|^2|\mathcal{S}|\tau T^{3/4})$. More recently, Cardoso et al. [6] provided a regret of $\mathcal{O}(\sqrt{\tau(\log|\mathcal{A}||\mathcal{S}|)T}\log T)$ for the same problem of uniformly ergodic MDPs with known transition function and

Table 1: Overview of theoretical regret guarantees of learning algorithms for uniformly ergodic A-MDPs. Expected stands for in expectation while High Probability stands for with high probability. $\beta \in (0, 1]$ in [23, 24] is a lower bound on all stationary distributions. The result in [24] only holds for $T = \tilde{\Omega}(|\mathcal{A}|\tau\beta^{-3})$.

Reference	Related Setting	Regret Bound	Regret Type
[11]	full feedback	$\mathcal{O}(\tau^2\sqrt{T\log|\mathcal{A}|})$	Expected
[23]	bandit feedback	$\mathcal{O}(T^{2/3}\tau\sqrt[3]{\log(T)|\mathcal{A}|\log(|\mathcal{A}|)/\beta})$	Expected
[24]	bandit feedback	$\mathcal{O}(\sqrt{\tau^3 T\log(T)|\mathcal{A}|}\log(|\mathcal{A}|)/\beta)$	Expected, Only for large T
[6]	full feedback	$\mathcal{O}(\sqrt{\tau(\log|\mathcal{A}||\mathcal{S}|)T}\log T)$	Expected
This work	bandit feedback	$\mathcal{O}((T|\mathcal{A}||\mathcal{S}|)^{2/3}\sqrt{\tau\log(|\mathcal{S}||\mathcal{A}|)\log T\log 1/\delta})$	Expected, High Probability

full information feedback. Compared to this line of work, we consider the bandit setting and propose a new algorithm achieving high probability regret bounds.

For uniformly ergodic MDPs and under bandit feedback, Neu et al. [23] developed an algorithm that obtains $\tilde{\mathcal{O}}(\tau T^{2/3}|\mathcal{A}|^{1/3}\beta^{-1/3})$ regret in expectation; here, β is a lower bound on all stationary distributions, typically satisfying $\beta^{-1} = \mathcal{O}(|\mathcal{S}|)$. The expected regret of this algorithm is furthered improved with respect to T to $\tilde{\mathcal{O}}(\sqrt{\tau^3 T|\mathcal{A}|\beta^{-1}})$ [24]; however, this result is semi-asymptotic, i.e., it holds only for very long time horizons satisfying $T = \tilde{\Omega}(|\mathcal{A}|\tau\beta^{-3})$. We note that all of the aforementioned episodic schemes, as well as the uniformly ergodic results in [11, 23, 24], are guaranteed to achieve certain *expected regrets*. In contrast, in this paper, we propose a new scheme that achieves regret bound of $\tilde{\mathcal{O}}((T|\mathcal{A}||\mathcal{S}|)^{\frac{2}{3}}\sqrt{\tau})$ not only in expectation but also with high probability.

Table 1 summarizes the differences between our approach with the most relevant existing methods.

3 BACKGROUND AND PRELIMINARY

We briefly overview the definitions of Markov decision processes, uniformly ergodicity assumption, random and expected regret, and the occupancy measures.

3.1 MARKOV DECISION PROCESS

Definition 1. *A Markov decision process (MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \ell)$, where $\mathcal{S}$ is a finite discrete state space, $\mathcal{A}$ is a finite discrete action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a probabilistic transition function, and $\ell : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a loss function.*

A sequence of actions by the agent generates a state trajectory over an MDP in the following manner. The agent starts in an initial state $s_{init} \in \mathcal{S}$. At time t, the agent is in state $\mathbf{s}_t$. Upon taking an action $\mathbf{a}_t \in \mathcal{A}$, the environment stochastically selects a next state $\mathbf{s}_{t+1}$ according to $\mathcal{P}(\cdot|\mathbf{s}_t, \mathbf{a}_t)$ and the agent receives a loss $\ell(\mathbf{s}_t, \mathbf{a}_t)$.

A policy π is a mapping from the history of states and actions, i.e., $\mathbf{h}_t = (\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \ldots, \mathbf{s}_t, a_t)$, to the action space. The goal is to find a policy that minimizes the expected cumulative loss

$$\mathbb{E}[\sum_{t=1}^{T} \ell(\mathbf{s}_t, \mathbf{a}_t)]$$

over a horizon of length T. Since MDPs admit optimal stochastic stationary policies [26], it suffices to search for an optimal policy in the family of stochastic stationary policies, i.e., $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Throughout this paper, we use $\pi(a|s)$ to denote the probability of selecting action a in state s and $\mathcal{P}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ to denote the transition probabilities between the states under policy π. Let $\nu_t \in \Delta_{\mathcal{S}}$ denote the state distribution at time t, where $\Delta_{\mathcal{S}}$ is a simplex in $\mathbb{R}^{|\mathcal{S}|}$. Further, let ν^π_{t+1} denote the state distribution at time t under policy π conforming to

$$\nu^\pi_{t+1} = \nu_t \mathcal{P}^\pi.$$

We use $\bar{\nu}^\pi$ to indicate the stationary distribution of policy π satisfying

$$\bar{\nu}^\pi = \bar{\nu}^\pi \mathcal{P}^\pi.$$

We assume that the MDP satisfies the so-called uniformly ergodic property, stated formally next.

Definition 2. *Let $\nu_1, \nu_2 \in \Delta_{\mathcal{S}}$ denote a pair of state distributions. A uniformly ergodic MDP is an MDP for which there exists $\tau \geq 1$ such that, for any policy π it holds that*

$$\|\nu_1 \mathcal{P}^\pi - \nu_2 \mathcal{P}^\pi\|_1 \leq e^{-\frac{1}{\tau}} \|\nu_1 - \nu_2\|_1.$$

Intuitively, for every policy over a uniformly ergodic MDP, the convergence rate of state distributions to a unique stationary distribution is exponentially fast. Similar to [24], we assume that the dynamics of the AMDP, i.e., the probabilistic transition function $\mathcal{P}$ is known. The important and more practical setting of dealing with unknown dynamic is left to future work.

In an A-MDP, the loss function, denoted by ℓ_t, varies over time. We assume that $\ell_t \in [0, 1]$ to remove the dependence of the analysis on the magnitude of the loss. We indicate the long-time average loss of a fixed policy π with respect to a fixed loss function ℓ by

$$\xi_\ell^\pi = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell(\mathbf{s}_t^\pi, \mathbf{a}_t^\pi).$$

One can show that for a fixed loss function ℓ, an optimal policy minimizing ξ_ℓ^π can be computed by solving a linear program (see (2) in [6]). We use ξ_t^π and $\xi_t^{\pi_t}$ to respectively denote the long-time average loss of policy π and policy π_t, with respect to the loss function ℓ_t.

We use the formulation of the regret minimization where the regret is defined with respect to the best policy in hindsight. Define

$$\mathcal{L}_T = \mathbb{E}[\sum_{t=1}^{T} \ell_t(\mathbf{s}_t, \mathbf{a}_t)]$$

as the expected cumulative loss of the learner, where the expectation is with respect to the randomness of the trajectories over the MDP. Also, define

$$\mathcal{L}_T(\pi) = \mathbb{E}[\sum_{t=1}^{T} \ell_t(\mathbf{s}_t, \mathbf{a}_t)|\pi]$$

as the expected cumulative loss under a fixed policy π. Then, the main goal of this paper is to achieve a low random regret

$$\mathcal{R}_T := \max_\pi \mathcal{L}_T - \mathcal{L}_T(\pi), \tag{1}$$

on which we seek a high-probability bound. Note that the randomness of $\mathcal{R}_T$ is injected by the learner and is different from the randomness in the objective function $\mathcal{L}_T$. The expected regret is defined as

$$\bar{\mathcal{R}}_T = \mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\max_\pi \mathcal{L}_T - \mathcal{L}_T(\pi)\right].$$

3.2 OCCUPANCY MEASURE

A (stochastic stationary) policy can equivalently be represented using occupancy measures. The occupancy measure of a policy is defined as the distribution induced by the execution of that policy over the state-action pairs, asymptotically, i.e.,

$$\rho^\pi(s,a) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \Pr(\mathbf{s}_t = s, \mathbf{a}_t = a|\pi).$$

A key property of occupancy measures is that the inflow into a state should be balanced by the outflow from that state. Formally, for every state $s \in \mathcal{S}$,

$$\sum_{a\in\mathcal{A}} \rho^\pi(s,a) = \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} \mathcal{P}(s|s',a')\rho^\pi(s',a').$$

Additionally, the occupancy measures are normalized over the entire state-action space, i.e.,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \rho(s,a) = 1.$$

There is a one-to-one mapping between stochastic stationary policies and occupancy measures. The policy π^ρ corresponding to an occupancy measure ρ can be computed according to

$$\pi^\rho(a|s) = \frac{\rho(s,a)}{\sum_{a'\in\mathcal{A}} \rho(s,a')}\ , \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{2}$$

The mapping between policies and occupancy measures allows one to reformulate a search over the policy space as a search over the occupancy measure space.

4 ONLINE RELATIVE-ENTROPY POLICY SEARCH WITH IMPLICIT EXPLORATION

Our proposed solution to the regret minimization problem is outlined in Algorithm 1. The algorithm builds upon the relative-entropy policy search of Peters et al. [25] and its online variant O-REPS by [31] while employing a novel loss estimator. In particular, Algorithm 1, which we refer to as *online relative-entropy policy search with implicit exploration (O-REPS-IX)*, employs an online mirror descent (OMD) optimization approach in conjunction to an optimistically biased loss estimator. We study these components of O-REPS-IX next.

4.1 ESTIMATING THE LOSS

In the bandit feedback setting, the agent observes only the loss corresponding to its current state and action and consequently has to construct an estimate of the overall loss function. In the episodic setting, the O-REPS algorithm [31] uses the following unbiased estimator to estimate the unobserved part of the loss in each episode and achieves the optimal expected regret:

$$\hat{\ell}_t^{\text{O-REPS}}(x,a) = \frac{\ell_t(s,a)\mathbb{I}\{(s,a)\in\mathbf{e}_t\},}{\rho_t(s,a)} \tag{3}$$

where $\mathbf{e}_t$ is the t^{th} episode. A similar unbiased loss estimator is further adopted in the uniformly ergodic setting [23, 24]. However, different from O-REPS [31] there is no notion of episode in the setting of uniformly ergodic A-MDPs that we consider here. One implication of this non-episodic aspect is that we can no longer limit the position of the agent to a specific layer at a time step. Furthermore, the agent does not restart from the initial state after every episode. Another implication of having no episode is that the agent updates

Algorithm 1 Online Relative-Entropy Policy Search with Implicit Exploration (O-REPS-IX)

Input: An MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$, time horizon T, estimation window N, exploration parameter γ, learning rate η

Output: Occupancy measures $\rho_1, \rho_2, \ldots, \rho_T$ at each time step

Initialize the occupancy measures for the first $2N-1$ time steps

$$\rho_t(s,a) = \frac{1}{|\mathcal{S}||\mathcal{A}|}\,, \qquad \forall t \in [2N-1], \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

Set the initial state $s_1 = s_{init}$ and the initial history $h_1 = (s_1)$

for $t = 1, \ldots, T$ **do**

  Compute the current policy at the current state

$$\pi_t(a|s_t) = \frac{\rho_t(s_t, a)}{\sum_{a' \in \mathcal{A}} \rho_t(s_t, a')}$$

  Draw an action a_t randomly from the distribution $\pi_t(a|s_t)$

  Observe the loss value $\ell_t(s_t, a_t)$ and the next state s_{t+1}

  Update the history,

$$\mathbf{h}_t \leftarrow \mathbf{h}_{t-1} + (a_t, \ell_t(s_t, a_t), s_{t+1})$$

  if $t \geq N$ **then**

    Compute $\nu_{t|t-N}$

    Construct the loss estimator $\hat{\ell}_t$ from the current history h_t

$$\hat{\ell}_t(s,a) = \frac{\ell_t(s,a)}{\nu_{t|t-N}(s)\pi_t(a|s) + \gamma}\mathbb{I}\{s_t = s, a_t = a\}$$

    Compute the optimal value function

$$\hat{v}_t = \arg\min_v \ln Z_t(v)$$

    Compute the solution ρ_{t+N} to the projection step (11)

$$\rho_{t+N}(s,a) = \frac{\rho_{t+N-1}(s,a)e^{\delta(s,a|\hat{v}_t, \hat{\ell}_t)}}{Z_t(\hat{v}_t)}$$

  end if

end for

the policy every time step given that the loss function may change in every time step. Furthermore, a major limitation of the estimators in [31, 23, 24] is that they suffer from a high degree of variance. Consequently, although these schemes achieve the optimal expected regret, it can be shown, using the arguments presented in Remark 1 in Section 11.5 and Exercise 11.5 in [18] and Section 3 in [2], that the random regret will be linear with a nonzero probability due to the high variance of the loss estimator.

Given above challenges, designing a loss estimator for the uniformly ergodic setting requires further considerations.

Let

$$\nu_{t|t-N}(s) = \Pr(\mathbf{s}_t = s|\mathbf{h}_{t-N})$$

denote the probability of being in state s at time t given the history at time $t-N$, $t \geq N+1$. Also, let $\vec{\nu}_{t|t-N}$ represent a vector of dimension $|\mathcal{S}|$, concatenating $\nu_{t|t-N}(s)$ for all $s \in \mathcal{S}$, and $e_{\mathbf{h}_{t-N}}$ represent a unit vector in $\mathbb{R}^{|\mathcal{S}|}$ such that

$$e_{\mathbf{h}_{t-N}}(s) = \begin{cases} 1 & \text{if } \mathbf{s}_{t-N} = s \\ 0 & \text{otherwise.} \end{cases}$$

Then, one can obtain $\vec{\nu}_{t|t-N}$ according to:

$$\vec{\nu}_{t|t-N} = e_{\mathbf{h}_{t-N}} \mathcal{P}^{\mathbf{a}_{t-N}} \mathcal{P}^{\pi_{t-N+1}} \mathcal{P}^{\pi_{t-N+2}} \ldots \mathcal{P}^{\pi_{t-1}},$$

where $\mathcal{P}^{\mathbf{a}_{t-N}}$ denotes the transition probabilities between the states upon taking action $\mathbf{a}_{t-N}$.

We propose the following loss estimator:

$$\hat{\boldsymbol{\ell}}_t(s,a) := \frac{\ell_t(s,a)}{\boldsymbol{\nu}_{t|t-N}(s)\boldsymbol{\pi}_t(a|s) + \gamma}\mathbb{I}\{\boldsymbol{s}_t = s, \boldsymbol{a}_t = a\}, \tag{4}$$

which exploits the bandit observation of the loss in the current time step; here, $\gamma > 0$ is an exploration parameter that induces exploration whose value will be determined in Theorem 1. Intuitively, $\boldsymbol{\nu}_{t|t-N}(s)\boldsymbol{\pi}_t(a|s)$ can be thought of as some form of occupancy measure. Looking at (3), one might be tempted to set $N = 1$ in (4), justified by the fact that in the episodic setting

$$\mathbb{E}[\mathbb{I}\{(s,a) \in \mathbf{e}_t | \mathbf{e}_1, \ldots, \mathbf{e}_{t-1}] = \rho_t(s,a).$$

However, this does not hold in the uniformly ergodic settings as

$$\begin{aligned} \mathbb{E}[\mathbb{I}\{\boldsymbol{s}_t = s, \boldsymbol{a}_t = a\}|t-1] &= \mathcal{P}(s, a|s_{t-1}, a_{t-1}) \\ &\neq \boldsymbol{\nu}_t(s)\boldsymbol{\pi}_t(a|s). \end{aligned} \tag{5}$$

Due to this discrepancy, which is discussed originally by Neu et al. [23], we will consider a sufficiently large N. Larger values of N, which we henceforth refer to as the estimation window, results in a better estimate of the loss function. Intuitively, delaying the policy update leads to lower variance of the random regret, enabling a high-probability analysis since the estimation window N helps to robustify the estimator against the learner's randomness.

Now given a sufficiently large N and an exploration parameter $\gamma > 0$, by taking the expectation of (4) we observe

$$\mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)|t-N] = \frac{\ell_t(s,a)\mathbb{E}[\mathbb{I}\{\boldsymbol{s}_t = s, \boldsymbol{a}_t = a\}|t-N]}{\boldsymbol{\nu}_{t|t-N}(s)\boldsymbol{\pi}_t(a|s)+\gamma} = \frac{\ell_t(s,a)\boldsymbol{\nu}_{t|t-N}(s)\boldsymbol{\pi}_t(a|s)}{\boldsymbol{\nu}_{t|t-N}(s)\boldsymbol{\pi}_t(a|s)+\gamma} \leq \ell_t(s,a). \tag{6}$$

That is, our proposed loss estimator in (4) is *optimistically biased*. This aspect is inline with the optimism principle in online learning [17]. Intuitively, since for a given state and action pair (s,a) the proposed estimator underestimates the true loss, as the agent interacts with the environment the estimated loss of any sub-optimal action will eventually become larger than that of the optimal ones. Furthermore, as we will show in Section 5, the proposed estimator in (4) achieves a variance reducing effect compared to typical unbiased estimators, e.g., (3), thereby enabling a high probability sublinear regret for O-REPS-IX.

4.2 POLICY UPDATE VIA OMD

Given an occupancy measure ρ, we use the unnormalized negative entropy as the potential function of OMD, i.e.,

$$R(\rho) = \sum_{\mathbf{s}\in\mathcal{S},a\in\mathcal{A}} \rho(s,a)\log\rho(s,a) - \sum_{\mathbf{s}\in\mathcal{S},a\in\mathcal{A}} \rho(s,a). \tag{7}$$

Given this potential function, the Bregman divergence $D(\rho\|\rho')$ between two occupancy measures ρ and ρ' is the unnormalized Kullback–Leibler divergence:

$$D(\rho\|\rho') = \sum_{\mathbf{s}\in\mathcal{S},a\in\mathcal{A}} \rho(s,a)\log\frac{\rho(s,a)}{\rho'(s,a)} - \sum_{\mathbf{s}\in\mathcal{S},a\in\mathcal{A}} \left(\rho(s,a)-\rho'(s,a)\right). \tag{8}$$

In t^{th} time step, the agent selects an occupancy measure $\boldsymbol{\rho}_{t+N}$ which minimizes a linear combination of the estimated loss $\hat{\boldsymbol{\ell}}_t$ and the divergence from the previous occupancy measure $\boldsymbol{\rho}_{t+N-1}$. Formally, the agent finds a solution to the constrained optimization problem

$$\boldsymbol{\rho}_{t+N} = \underset{\rho\in\Delta(\mathcal{M})}{\arg\min}\left\{\eta\langle\rho,\hat{\boldsymbol{\ell}}_t\rangle + D(\rho\|\boldsymbol{\rho}_{t+N-1})\right\}, \tag{9}$$

where $\Delta(\mathcal{M})$ denotes the set of all occupancy measures over an MDP $\mathcal{M}$ that satisfy the inflow-outflow balancing and normalization constraints, and $\langle\cdot,\cdot\rangle$ is the inner product in the space of $\mathcal{S}\times\mathcal{A}$. Note that as we discussed $\boldsymbol{\rho}_{t+N}$ is entirely determined by the history $\mathbf{h}_t$. Hence, in the first $2N-1$ rounds of learning (see step 3 of Algorithm 1) we initialize the occupancy measure (and consequently the policy π_t) uniformly, i.e., $\boldsymbol{\rho}_t(s,a) = 1/|\mathcal{S}||\mathcal{A}|$.

Similar to the standard mirror descent techniques, the constrained optimization in (9) can be efficiently solved through a two-step procedure. First, an unconstrained version of the problem is solved, i.e., we find

$$\tilde{\boldsymbol{\rho}}_{t+N} = \underset{\rho}{\arg\min}\left\{\eta\langle\rho,\hat{\boldsymbol{\ell}}_t\rangle + D(\rho\|\boldsymbol{\rho}_{t+N-1})\right\}, \tag{10}$$

which admits a closed form solution

$$\tilde{\boldsymbol{\rho}}_{t+N}(s,a) = \boldsymbol{\rho}_{t+N-1}(s,a)e^{-\eta\hat{\boldsymbol{\ell}}_t(s,a)}.$$

Then, $\tilde{\boldsymbol{\rho}}_{t+N}(s,a)$ is projected to the constraint set $\Delta(\mathcal{M})$, i.e., we find

$$\boldsymbol{\rho}_{t+N} = \underset{\rho\in\Delta(\mathcal{M})}{\arg\min}\left\{D(\rho\|\tilde{\boldsymbol{\rho}}_{t+N-1})\right\}. \tag{11}$$

By enforcing constraints of inflow-outflow balancing and normalization on the occupancy measures, the following constrained optimization yields the solution to the projection step:

$$\begin{aligned} \min_{\rho}\quad & D(\rho\|\tilde{\boldsymbol{\rho}}_{t+N-1}) \\ \textit{s.t.}\quad & \sum_{a\in\mathcal{A}}\rho(s,a) = \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}}\mathcal{P}(s|s',a')\rho(s',a') \quad \forall s\in\mathcal{S}, \\ & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\rho(s,a) = 1. \end{aligned} \tag{12}$$

We show in Proposition 1 that the above optimization problem using ideas from [31] can equivalently be written as an unconstrained convex optimization problem.

Proposition 1. *Let $v:\mathcal{S}\to\mathbb{R}$ denote a value function for each state and $\ell:\mathcal{S}\times\mathcal{A}\to[0,1]$ denote a loss function. Define*

$$\delta(s,a|v,\ell) = \nu(s) - \eta\ell(s,a) - \sum_{\mathbf{s}'\in\mathcal{S}} v(s')\mathcal{P}(s'|s,a),$$

a function capturing the notion of Bellman error for the value function v. Furthermore, for $t>1$, define a partition function

$$\mathbf{Z}_t(v) = \sum_{s\in\mathcal{S},a\in\mathcal{A}} \boldsymbol{\rho}_t(s,a)e^{\delta(s,a|v,\hat{\boldsymbol{\ell}}_t)}.$$

The optimal value function (corresponding to the dual problem) is the solution to an unconstrained optimization problem

$$\hat{v}_t = \underset{v}{\arg\min}\ln\mathbf{Z}_t(v).$$

With these definitions in place, the solution to the projection step (11) *is*

$$\boldsymbol{\rho}_{t+N}(s,a) = \frac{\boldsymbol{\rho}_{t+N-1}(s,a)e^{\delta(s,a|\hat{v}_t,\hat{\boldsymbol{\ell}}_t)}}{\mathbf{Z}_t(\hat{v}_t)}.$$

Proof. The projection step of online mirror descent has a closed-form solution for episodic A-MDPs as shown in [31]; it is readily derived by differentiating the Lagrangian with respect to $\rho(s,a)$ and setting the gradient to zero. The solution follows by using the second constraint and solving the dual maximization problem.

The projection step of online mirror descent for uniformly ergodic MDPs in (12) is different from the one for episodic MDPs only in constraints on the occupancy measures. That is, for an episodic MDP, occupancy measures are normalized across each layer while for a uniformly ergodic MDP the normalization is over the entire state space. Therefore, with comparable arguments to those stated in [31], we can derive the presented closed form solution. ∎

5 REGRET ANALYSIS

We now present the theoretical analysis of O-REPS-IX for A-MDPs satisfying the uniform ergodicity assumption. The detailed proofs are deferred to the Appendix.

5.1 BOUND ON THE RANDOM REGRET

We start by stating our main theoretical result, establishing a high-probability bound on the random regret $\mathcal{R}_T$.

Theorem 1. *Let*

$$\begin{aligned}\eta &= (T|\mathcal{S}||\mathcal{A}|)^{-2/3}\sqrt{\log(|\mathcal{S}||\mathcal{A}|)},\\ \gamma &= (T|\mathcal{S}||\mathcal{A}|)^{-1/3}\sqrt{\tau\log T\log\frac{1}{\delta}},\\ N &= 1+\lceil\tau\log T\rceil.\end{aligned} \tag{13}$$

Then, for any $\delta\in(0,1)$, with probability at least $1-4\delta$, it holds that the random regret of Algorithm 1 satisfies

$$\begin{aligned}\mathcal{R}_T \le\, & C\,(T|\mathcal{A}||\mathcal{S}|)^{\frac{2}{3}}\sqrt{\tau\log(|\mathcal{S}||\mathcal{A}|)\log T\log\frac{1}{\delta}}\\ &+C'\tau\log T,\end{aligned}$$

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

Remark 1. *Theorem 1 establishes that Algorithm 1 achieves $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ regret bound with high probability. The pioneering algorithm in [24] achieves an optimal regret of $\tilde{\mathcal{O}}(\sqrt{T})$ only in expectation for sufficiently large T, and under an extra assumption compared to our result (see Assumption A2 there) that bounds all stationary distributions from zero. Algorithm 1 enjoys a high-probability regret bound which is a much stronger type of guarantee. Additionally, the proposed algorithm has a better dependence on τ compared to the algorithm in [24] (that is, $\mathcal{O}(\sqrt{\tau})$ vs. $\mathcal{O}(\tau\sqrt{\tau})$). In our current analysis we employ a relatively large γ to ensure that the proposed estimator is uniformly bounded with probability one. However, this restriction results in a sub-optimal regret bound $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$. Hence, it remains an open problem to see whether the high probability regret of Algorithm 1 can be improved to $\tilde{\mathcal{O}}(\sqrt{T})$, i.e., the optimal regret (with respect to T).*

Proof of Theorem 1. Recall step 3 of O-REPS-IX (see Algorithm 1). Given that by assumption $\ell_t(s,a)\le 1$, the first $2N-1$ terms in the random regret $\mathcal{R}_T$ can be bounded by $2N$. Hence, we study the regret of O-REPS-IX starting $t=2N$. To obtain the regret bound, we first decompose the random regret:

$$\begin{aligned}\mathcal{R}_T = \mathcal{O}(N) + &\underbrace{\max_{\rho\in\Delta(\mathcal{M})}\sum_{t=2N}^{T}\xi_t^{\pi} - \mathbb{E}\left[\sum_{t=2N}^{T}\ell_t(s_t^{\pi},a_t^{\pi})\right]}_{\text{I}}\\ &+\underbrace{\sum_{t=2N}^{T}\xi_t^{\pi_t}-\sum_{t=2N}^{T}\xi_t^{\pi}}_{\text{II}}\\ &+\underbrace{\mathbb{E}\left[\sum_{t=2N}^{T}\ell_t(s_t^{\pi_t},a_t^{\pi_t})\right]-\sum_{t=2N}^{T}\xi_t^{\pi_t}}_{\text{III}}.\end{aligned}$$

Next, we decompose the second term according to

$$\begin{aligned}\sum_{t=2N}^{T}\xi_t^{\pi_t}-\sum_{t=2N}^{T}\xi_t^{\pi} &= \sum_{t=2N}^{T}\langle\boldsymbol{\rho}_t-\rho,\ell_t\rangle\\ &=\underbrace{\sum_{t=2N}^{T}\langle\boldsymbol{\rho}_t,\ell_t-\hat{\boldsymbol{\ell}}_t\rangle}_{\text{II-I}}+\underbrace{\sum_{t=2N}^{T}\langle\rho,\hat{\boldsymbol{\ell}}_t-\ell_t\rangle}_{\text{II-II}}\\ &+\underbrace{\sum_{t=2N}^{T}\langle\boldsymbol{\rho}_t-\rho,\hat{\boldsymbol{\ell}}_t\rangle}_{\text{II-III}},\end{aligned}$$

by recalling definition of ξ^{π} and ξ^{π_t}:

$$\begin{aligned}\xi^{\pi} &= \lim_{T'\to\infty}\sum_{t=2N}^{T'}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Pr(s_{t'}^{\pi}=s,a_{t'}^{\pi}=a)\ell_t(s,a)\\ &=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\rho(s,a)\ell_t(s,a),\end{aligned} \tag{14}$$

and

$$\begin{aligned}\xi^{\pi_t} &= \lim_{T'\to\infty}\sum_{t=2N}^{T'}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Pr(s_{t'}^{\pi_t}=s,a_{t'}^{\pi_t}=a)\ell_t(s,a)\\ &=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\boldsymbol{\rho}_t(s,a)\ell_t(s,a).\end{aligned} \tag{15}$$

Now, we bound each term individually. For a fixed policy π over a finite time horizon T, term I measures the difference between the expected reward starting from the initial distribution $\boldsymbol{\nu}_1$ and the expected reward starting from the stationary distribution $\bar{\boldsymbol{\nu}}^{\pi}$. Note that this term is deterministic and we can bound this difference by a factor of τ due to the uniform ergodicity assumption which ensures fast mixing time (Lemma 1 and Appendix B).

Lemma 1 (Bounding term I [11])**.** *For any $T \geq 1$ and any policy π, it holds that*

$$\sum_{t=2N}^{T} \xi_t^{\pi} - \mathbb{E}\left[\sum_{t=2N}^{T} \ell_t(s_t^{\pi}, a_t^{\pi})\right] \leq 2(1+\tau). \tag{16}$$

Term II – which is random – is studied in Lemma 2 (see Appendix C for the unabridged statement). To analyze term II-I, we show in Lemma 9 the fact that the evolving policies from the mirror descent algorithm do not change much between consecutive time steps as long as N satisfies the condition given in Theorem 1. We further need to show that $\hat{\boldsymbol{\ell}}_t$ is a good estimate for ℓ_t, which we prove in two steps. First, we show that ℓ_t is close to $\mathbb{E}[\hat{\boldsymbol{\ell}}_t]$ by a factor directly proportional to γ. Note that with $\gamma = 0$, $\hat{\boldsymbol{\ell}}_t$ becomes an unbiased estimator of ℓ_t. Second, $\hat{\ell}_t$ concentrates around its mean $\mathbb{E}[\hat{\boldsymbol{\ell}}_t]$. We bound term II-II in Lemma 10 by relying on closeness of $\hat{\boldsymbol{\ell}}_t$ to ℓ_t. In particular, the optimistic bias of $\hat{\boldsymbol{\ell}}_t$ allows us to use a concentration result based on the Cramer-Chernoff method (Lemma 4). Analysis of term II-III also has two components, where one of them depends on the regret of the mirror descent algorithm and the other one depends on the fact that the iterates of the OMD do not change too rapidly as long as η and γ satisfy the conditions of Theorem 1.

Lemma 2 (Bounding term II)**.** *With η, γ, and N given in* (13)*, it holds, with probability exceeding $1-4\delta$, that*

$$\sum_{t=2N}^{T} \xi^{\pi_t} - \sum_{t=2N}^{T} \xi^{\pi} = \mathcal{O}\left((T|\mathcal{A}||\mathcal{S}|)^{\frac{2}{3}} \sqrt{\tau \log(|\mathcal{S}||\mathcal{A}|) \log T \log \frac{1}{\delta}} \right).$$

Term III, similar to term II, is random and captures the difference between the expected reward actually obtained by the agent and the expected reward obtained by the agent had it been in the stationary distribution $\bar{\boldsymbol{\nu}}^{\pi_t}$ at each time step t. By using the fact that the evolving policies from the mirror descent algorithm do not change much between consecutive time steps, in Lemma 3 we establish an upper bound on term III (see Appendix D for the proof).

Lemma 3 (Bounding term III)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\mathbb{E}\left[\sum_{t=2N}^{T} \ell_t(s_t, a_t)\right] - \sum_{t=2N}^{T} \xi_t^{\pi_t} \leq 2(1+\tau) + 2\eta(1+\tau)\left(\frac{\log\frac{1}{\delta}}{2\gamma} + T|\mathcal{S}||\mathcal{A}|\right).$$

Lastly, by adding the bounds from each term and properly selecting the values of γ and η, we obtain the desired result stated in Theorem 1. The details of the regret bound is provided in Appendix E. ■

5.2 BOUND ON THE EXPECTED REGRET

An upper bound on the expected regret of Algorithm 1 can also be obtained by integrating the tail of the high-probability regret bound provided in Theorem 1. The result is formalized in the following theorem.

Theorem 2. *With η, γ, and N given in* (13)*, the expected regret of Algorithm 1 satisfies*

$$\bar{\mathcal{R}}_T \leq C\,(T|\mathcal{A}||\mathcal{S}|)^{\frac{2}{3}} \sqrt{\tau \log(|\mathcal{S}||\mathcal{A}|) \log T \log \frac{1}{\delta}} + C'\tau \log T, \tag{17}$$

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

Proof of Theorem 2. First, we relate the expected regret to the random regret on which we have derived a high-probability bound. In particular, defining $\mathcal{R}_T^{+} := \max\{\mathcal{R}_T, 0\}$, we have

$$\bar{\mathcal{R}}_T \leq \mathbb{E}\left[\mathcal{R}_T^{+}\right] \overset{(a)}{=} \int_0^{\infty} \Pr\left(\mathcal{R}_T^{+} \geq u\right) du = \int_0^{\infty} \Pr\left(\mathcal{R}_T \geq u\right) du,$$

where (a) is due to the fact that for a non-negative random variable X it holds that $\mathbb{E}[X] = \int_0^{\infty} \Pr(X > x)dx$. We can evaluate the integral using the high-probability bound of Theorem 1 and a change of variables. Assume $\tau > 1$ and let $B = (T|\mathcal{A}||\mathcal{S}|)^{\frac{2}{3}} \sqrt{\tau \log(|\mathcal{S}||\mathcal{A}|) \log T}$. Then, it is apparent that with probability at most 4δ, the following lower bound holds on the random regret:

$$\mathcal{R}_T \geq CB \log \frac{1}{\delta} + C'\tau \log T, \tag{18}$$

for some $C, C' > 0$. Note that the second term is deterministic. Next, let $u = CB\log\left(\frac{1}{\delta}\right)$ and thus $\delta = \exp\left(-u/CB\right)$. If $\delta \to 0^{+}$, then $u \to \infty$, while if $\delta \to \left(\frac{1}{4}\right)^{-}$, then $u \to (CB\log 4)^{+}$. Then,

$$\bar{\mathcal{R}}_T \overset{(b)}{\leq} C'\tau \log T + \int_{CB\log 4}^{\infty} \Pr\left(\mathcal{R}_T \geq CB\log\frac{L}{\delta}\right) du \overset{(c)}{\leq} C'\tau \log T + \int_{CB\log 4}^{\infty} 4\exp\left(-\frac{u}{CB}\right) du,$$

where (b) is due to the nonnegativity of the integrand and (c) corresponds to the simplified high-probability bound in (18). Lastly, a simple integration yields the desired result. ■

6 CONCLUSION AND FUTURE WORK

We considered the general class of uniformly ergodic A-MDPs whose loss functions may change arbitrarily over time. By relying on an optimistically biased loss estimator and online linear optimization techniques, we proposed O-REPS-IX that finds a policy achieving sublinear regret bounds both with high probability and in expectation. In particular, the algorithm achieves the regret of $\tilde{\mathcal{O}}(T^{\frac{2}{3}}\sqrt{\tau|\mathcal{A}||\mathcal{S}|})$ with respect to the best stationary policy in hindsight. The proposed scheme is the first algorithm achieving a high probability sublinear regret bound in the setting of learning with uniformly ergodic A-MDPs and bandit feedback.

As a future research direction, it is important to establish whether the high-probability regret of O-REPS-IX can be improved to $\tilde{\mathcal{O}}(\sqrt{T})$, i.e., the optimal regret. Furthermore, we would like to explore the potential of using the proposed algorithm for learning in safety-critical scenarios. In these scenarios, the high-probability guarantees of O-REPS-IX can be employed to provide desirable safety assurances. Finally, it is valuable to extend our results to the class of risk-aware MDPs.

ACKNOWLEDGMENTS

This work was supported partially by AFRL grant FA9550-19-1-0169, DARPA grant D19AP00004, and NSF ECCS grant 1809327.

REFERENCES

[1] ABBASI-YADKORI, Y., BARTLETT, P. L., CHEN, X., AND MALEK, A. Large-scale Markov decision problems via the linear programming dual. *arXiv preprint arXiv:1901.01992* (2019).

[2] ABERNETHY, J., AND RAKHLIN, A. Beating the adaptive bandit with high probability. In *2009 Information Theory and Applications Workshop* (2009), IEEE, pp. 280–289.

[3] AZAR, M. G., OSBAND, I., AND MUNOS, R. Minimax regret bounds for reinforcement learning. In *Proceedings of International Conference on Machine Learning (ICML)* (2017), vol. 70, JMLR. org, pp. 263–272.

[4] BERTSEKAS, D. P., BERTSEKAS, D. P., BERTSEKAS, D. P., AND BERTSEKAS, D. P. *Dynamic programming and optimal control*, vol. 1. Athena scientific Belmont, MA, 1995.

[5] BERTSEKAS, D. P., AND TSITSIKLIS, J. N. *Neuro-dynamic programming*, vol. 5. Athena Scientific Belmont, MA, 1996.

[6] CARDOSO, A. R., WANG, H., AND XU, H. Large scale Markov decision processes with changing rewards. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (2019), pp. 2337–2347.

[7] DANN, C., LI, L., WEI, W., AND BRUNSKILL, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning* (2019), PMLR, pp. 1507–1516.

[8] DICK, T., GYORGY, A., AND SZEPESVARI, C. Online learning in Markov decision processes with changing cost sequences. In *Proceedings of International Conference on Machine Learning (ICML)* (2014), pp. 512–520.

[9] DOMINGUES, O. D., MÉNARD, P., KAUFMANN, E., AND VALKO, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory* (2021), PMLR, pp. 578–598.

[10] DUCHI, J. C., SHALEV-SHWARTZ, S., SINGER, Y., AND TEWARI, A. Composite objective mirror descent. In *COLT* (2010), Citeseer, pp. 14–26.

[11] EVEN-DAR, E., KAKADE, S. M., AND MANSOUR, Y. Online Markov decision processes. *Mathematics of Operations Research 34*, 3 (2009), 726–736.

[12] HOEFFDING, W. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*. Springer, 1994, pp. 409–426.

[13] JAKSCH, T., ORTNER, R., AND AUER, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research (JMLR) 11*, Apr (2010), 1563–1600.

[14] JIN, C., JIN, T., LUO, H., SRA, S., AND YU, T. Learning adversarial MDPs with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192* (2019).

[15] KALAI, A., AND VEMPALA, S. Efficient algorithms for online decision problems. In *Learning Theory and Kernel Machines*. Springer, 2003, pp. 26–40.

[16] KOCÁK, T., NEU, G., VALKO, M., AND MUNOS, R. Efficient learning by implicit exploration in bandit problems with side observations. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (2014), pp. 613–621.

[17] LAI, T. L., AND ROBBINS, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*, 1 (1985), 4–22.

[18] LATTIMORE, T., AND SZEPESVÁRI, C. Bandit algorithms. *preprint* (2018).

[19] LITTLESTONE, N., AND WARMUTH, M. K. The weighted majority algorithm. *Information and computation 108*, 2 (1994), 212–261.

[20] NESTEROV, Y., ET AL. *Lectures on convex optimization*, vol. 137. Springer, 2018.

[21] NEU, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (2015), pp. 3168–3176.

[22] NEU, G., GYORGY, A., AND SZEPESVÁRI, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)* (2012), pp. 805–813.

[23] NEU, G., GYÖRGY, A., SZEPESVÁRI, C., AND ANTOS, A. Online markov decision processes under bandit feedback. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2* (2010), pp. 1804–1812.

[24] NEU, G., GYORGY, A., SZEPESVARI, C., AND ANTOS, A. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control (TAC) 3*, 59 (2014), 676–691.

[25] PETERS, J., MÜLLING, K., AND ALTÜN, Y. Relative entropy policy search. In *Proceedings of AAAI Conference on Artificial Intelligence* (2010), pp. 1607–1612.

[26] PUTERMAN, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[27] ROSENBERG, A., AND MANSOUR, Y. Online convex optimization in adversarialMarkov decision processes. In *Proceedings of International Conference on Machine Learning (ICML)* (2019), pp. 5478–5486.

[28] WANG, M. Primal-dual π learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *arXiv preprint arXiv:1710.06100* (2017).

[29] YU, J. Y., MANNOR, S., AND SHIMKIN, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research 34*, 3 (2009), 737–757.

[30] ZANETTE, A., AND BRUNSKILL, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning* (2019), PMLR, pp. 7304–7312.

[31] ZIMIN, A., AND NEU, G. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Proceedings of Advances in neural information processing systems (NeurIPS)* (2013), pp. 1583–1591.

A INTERMEDIATE LEMMAS

We first state a concentration lemma from [18] (Lemma 12.2 there) which was originally stated in [21].

Lemma 4. *Let $\mathbb{F} = (\mathcal{F}_t)_{t=1}^T$ to be a filtration and for $i \in [k]$ let $(\tilde{\mathbf{Y}}_{ti})_t$ be $\mathbb{F}$-adapted such that the following conditions hold:*

1. *For any $S \subset [k]$ with $|S| > 1$, $\mathbb{E}\left[\prod_{i\in S} \tilde{\mathbf{Y}}_{ti} | \mathcal{F}_{t-1}\right] \leq 0$.*
2. *$\mathbb{E}\left[\tilde{\mathbf{Y}}_{ti} | \mathcal{F}_{t-1}\right] = y_{ti}$ for all $t \in [T]$ and $i \in [k]$.*

Furthermore, let $(\alpha_{ti})_{ti}$ and $(\lambda_{ti})_{ti}$ be real-valued $\mathbb{F}$-predictable random sequences such that for all t, i it holds that $0 \leq \alpha_{ti}\tilde{\mathbf{Y}}_{ti} \leq 2\lambda_{ti}$. Then for any $\delta \in (0, 1)$,

$$\Pr\left(\sum_{t=1}^{T}\sum_{i=1}^{k} \alpha_{ti}\left(\frac{\tilde{\mathbf{Y}}_{ti}}{1+\lambda_{ti}} - y_{ti}\right) \geq \log\frac{1}{\delta}\right) \leq \delta. \tag{19}$$

We use this Lemma frequently in establishing high-probability bounds on terms involved in the regret bound of the proposed algorithm with i going over $\mathcal{S} \times \mathcal{A}$ and $k = |\mathcal{S}||\mathcal{A}|$, and judicious choices for $(\tilde{\mathbf{Y}}_{ti})_t$, $(\alpha_{ti})_{ti}$, $(\lambda_{ti})_{ti}$, and y_{ti}. Furthermore, we define the filtration $\mathbb{F}$ to capture all sources of randomness up to time $t - N + 1$, including $t - N + 1$ itself where $N > 1$. That is, $\mathbb{F} = (\mathcal{F}_t)_{t=2N}^T$, where $\mathcal{F}_t = \sigma(\mathbf{h}_{t-N+1})$ such that for any random variable $\mathbf{z} \in \mathcal{F}_t$, $\mathbf{z}$ is measurable by the history $\mathbf{h}_{t-N+1}$.

A.1 AUXILIARY LEMMAS

Next, we state four auxiliary lemmas that will be used multiple times in the proofs of the main lemmas.

Lemma 5. *For any $\delta_1 \in (0, 1)$, with probability at least $1 - \delta_1$, it holds that*

$$\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \hat{\boldsymbol{\ell}}_t(s,a) - \ell_t(s,a) \leq \frac{\log\frac{1}{\delta_1}}{2\gamma}. \tag{20}$$

Proof. Define the following random sequences

$$\begin{aligned}
\alpha_t(s,a) &= 2\gamma, \\
\lambda_t(s,a) &= \frac{\gamma}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}, \\
\tilde{\mathbf{Y}}_t(s,a) &= \frac{\ell_t(s,a)\mathbb{I}\{s_t = s, a_t = a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}.
\end{aligned} \tag{21}$$

Note that $\alpha_t(s,a)$ and $\lambda_t(s,a)$ are real-valued $\mathbb{F}$-predictable random sequences such that

$$0 \leq \alpha_t(s,a)\tilde{\mathbf{Y}}_t(s,a) \leq 2\lambda_t(s,a). \tag{22}$$

due to non-negativity of $\alpha_t(s,a)$ and $\tilde{\mathbf{Y}}_t(s,a)$ and the bounded loss assumption:

$$\ell_t(s,a)\mathbb{I}\{s_t = s, a_t = a\} \leq 1. \tag{23}$$

Additionally, $\tilde{\mathbf{Y}}_t(s,a)$ is $\mathbb{F}$-adapted and the two conditions stated in Lemma 4 hold since

$$\mathbb{E}\left[\prod_{(s,a)\in S} \tilde{\mathbf{Y}}_t(s,a)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\prod_{(s,a)\in S} \frac{\ell_t(s,a)\mathbb{I}\{s_t = s, a_t = a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}|\mathcal{F}_{t-1}\right] = 0, \tag{24}$$

for any set $S \subset \mathcal{S} \times \mathcal{A}$ having at least two elements, and

$$\mathbb{E}\left[\tilde{\mathbf{Y}}_t(s,a)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{\ell_t(s,a)\mathbb{I}\{s_t = s, a_t = a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}|\mathcal{F}_{t-1}\right] = \ell_t(s,a). \tag{25}$$

Therefore we can can use the concentration result of Lemma 4 to prove the gap between the estimated loss and the expected loss are bounded with high probability according to:

$$
\begin{aligned}
&\Pr\left(\sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha_t(s,a) \left(\frac{\tilde{\mathbf{Y}}_t(s,a)}{1+\lambda_t(s,a)} - \ell_t(s,a) \right) \geq \log\frac{1}{\delta_1} \right) \leq \delta_1 \\
\Leftrightarrow \quad &\Pr\left(\sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \hat{\boldsymbol{\ell}}_t(s,a) - \ell_t(s,a) \geq \frac{\log\frac{1}{\delta_1}}{2\gamma} \right) \leq \delta_1 .
\end{aligned} \tag{26}
$$

■

Lemma 6. *For any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, it holds that*

$$
\sum_{t=2N}^{T} \|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1 \leq 2(1+\tau) + 2\eta(1+\tau) \left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}| \right). \tag{27}
$$

Proof. By triangle inequality we have that

$$
\|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1 \leq \|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_{t-1}}\|_1 + \|\bar{\boldsymbol{\nu}}^{\pi_{t-1}} - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1. \tag{28}
$$

By a simple recursive application of the τ-mixing assumption the first term in RHS of (28) can be bounded as

$$
\begin{aligned}
\|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_{t-1}}\|_1 &= \|\boldsymbol{\nu}_{t-1}\mathcal{P}^{\pi_{t-1}} - \bar{\boldsymbol{\nu}}^{\pi_{t-1}}\mathcal{P}^{\pi_{t-1}}\|_1 \\
&\leq e^{-\frac{1}{\tau}} \|\boldsymbol{\nu}_{t-1} - \bar{\boldsymbol{\nu}}^{\pi_{t-1}}\|_1 \\
&\leq e^{-\frac{1}{\tau}} \|\boldsymbol{\nu}_{t-1} - \bar{\boldsymbol{\nu}}^{\pi_{t-2}}\|_1 + e^{-\frac{1}{\tau}} \|\bar{\boldsymbol{\nu}}^{\pi_{t-1}} - \bar{\boldsymbol{\nu}}^{\pi_{t-2}}\|_1.
\end{aligned} \tag{29}
$$

Regarding the second term in RHS of (28), using the definition of the occupancy measure we have

$$
\begin{aligned}
\sum_{s\in\mathcal{S}} |\bar{\nu}^{\pi_{t-1}}(s) - \bar{\nu}^{\pi_t}(s)| &= \sum_{s\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}} \rho_{t-1}(s,a) - \rho_t(s,a) \right| \\
&\leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\rho_{t-1}(s,a) - \rho_t(s,a)| \\
&\leq \|\boldsymbol{\rho}_{t-1} - \boldsymbol{\rho}_t\|_1.
\end{aligned} \tag{30}
$$

Thus, by unrolling (29) to $t = 2N$ we can bound $\|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1$ according to

$$
\|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1 \leq e^{-\frac{t-1}{\tau}} \|\boldsymbol{\nu}_{2N} - \bar{\boldsymbol{\nu}}^{\pi_1}\|_1 + \sum_{q=1}^{t-1} e^{-\frac{q-1}{\tau}} \|\boldsymbol{\rho}_{t-q} - \boldsymbol{\rho}_{t-q+1}\|_1. \tag{31}
$$

Summing the above result over t yields

$$
\sum_{t=2N}^{T} \|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1 \leq \sum_{t=2N}^{T} e^{-\frac{t-1}{\tau}} \|\boldsymbol{\nu}_{2N} - \bar{\boldsymbol{\nu}}^{\pi_1}\|_1 + \sum_{t=2N}^{T} \sum_{q=1}^{t-1} e^{-\frac{q-1}{\tau}} \|\boldsymbol{\rho}^{\pi_{t-q}} - \boldsymbol{\rho}^{\pi_{t-q+1}}\|_1. \tag{32}
$$

The first term in the RHS of (32) can be bounded by $2(1+\tau)$ since $\|\boldsymbol{\nu}_1 - \bar{\boldsymbol{\nu}}^{\pi_1}\|_1 \leq 2$ and

$$
2 \sum_{t=2N}^{T} e^{-\frac{t-1}{\tau}} \leq 2(1 + \int_0^\infty e^{-\frac{t}{\tau}}) = 2(1+\tau). \tag{33}
$$

The starting point of our method to bound the second term in RHS of (32) is similar to the proof of Lemma 6 in [6]. However, the main steps are different due to the bandit feedback setting that we consider.

First note that since $R(.)$ is a barrier function and its domain is the probability simplex we can equivalently express the update rule of $\boldsymbol{\rho}_t$ for $t \geq 2N-1$ according to (see section 28.1 in [18] for more detail)

$$
\boldsymbol{\rho}_{t+1} = \arg\min_{\rho\in\Delta} \left[J^t(\rho) := \sum_{i=2N-1}^{t} \langle \rho, \hat{\boldsymbol{\ell}}_{i-N+1} \rangle + \frac{1}{\eta} R(\rho) \right]. \tag{34}
$$

Since R is 1-strongly convex w.r.t. $\|.\|_1$, $J^t(.)$ is $1/\eta$-strongly convex. Thus, we can establish by strong convexity and the optimality condition for $\boldsymbol{\rho}_{t+1}$ (see, e.g., Theorem 2.2.9 in [20]) that

$$
\begin{aligned}
\frac{1}{2\eta}\|\boldsymbol{\rho}_{t+1}-\boldsymbol{\rho}_t\|_1^2 &\leq J^t(\boldsymbol{\rho}_t)-J^t(\boldsymbol{\rho}_{t+1})+\langle \nabla J^t(\boldsymbol{\rho}_{t+1}),\boldsymbol{\rho}_{t+1}-\boldsymbol{\rho}_t\rangle \\
&\leq \left[\sum_{i=2N-1}^{t}\langle \boldsymbol{\rho}_t,\hat{\boldsymbol{\ell}}_{i-N+1}\rangle+\frac{1}{\eta}R(\boldsymbol{\rho}_t)\right]-\left[\sum_{i=2N-1}^{t}\langle \boldsymbol{\rho}_{t+1},\hat{\boldsymbol{\ell}}_{i-N+1}\rangle+\frac{1}{\eta}R(\boldsymbol{\rho}_{t+1})\right] \\
&\leq \left[\sum_{i=2N-1}^{t-1}\langle \boldsymbol{\rho}_t,\hat{\boldsymbol{\ell}}_{i-N+1}\rangle+\frac{1}{\eta}R(\boldsymbol{\rho}_t)\right]-\left[\sum_{i=2N-1}^{t-1}\langle \boldsymbol{\rho}_{t+1},\hat{\boldsymbol{\ell}}_{i-N+1}\rangle+\frac{1}{\eta}R(\boldsymbol{\rho}_{t+1})\right] \\
&\quad+\langle \boldsymbol{\rho}_t,\hat{\boldsymbol{\ell}}_{t-N+1}\rangle-\langle \boldsymbol{\rho}_{t+1},\hat{\boldsymbol{\ell}}_{t-N+1}\rangle \\
&\leq \langle \boldsymbol{\rho}_t-\boldsymbol{\rho}_{t+1},\hat{\boldsymbol{\ell}}_{t-N+1}\rangle \\
&\leq \|\boldsymbol{\rho}_{t+1}-\boldsymbol{\rho}_t\|_1\|\hat{\boldsymbol{\ell}}_{t-N+1}\|_\infty,
\end{aligned}
\tag{35}
$$

where we used Hölder's inequality and the fact that by the update rule and optimality of $\boldsymbol{\rho}_t$, the term

$$
\left[\sum_{i=2N-1}^{t-1}\langle \boldsymbol{\rho}_t,\hat{\boldsymbol{\ell}}_{i-N+1}\rangle+\frac{1}{\eta}R(\boldsymbol{\rho}_t)\right]
$$

can be bounded by

$$
\left[\sum_{i=2N-1}^{t-1}\langle \boldsymbol{\rho}_{t+1},\hat{\boldsymbol{\ell}}_{i-N+1}\rangle+\frac{1}{\eta}R(\boldsymbol{\rho}_{t+1})\right]
$$

which in turn results in cancellation of the two. Thus,

$$
\|\boldsymbol{\rho}_{t+1}-\boldsymbol{\rho}_t\|_1 \leq 2\eta\|\hat{\boldsymbol{\ell}}_{t-N+1}\|_\infty, \tag{36}
$$

and in turn

$$
\sum_{t=2N}^{T}\|\boldsymbol{\nu}_t-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1 \leq 2(1+\tau)+2\eta\sum_{t=2N}^{T}\sum_{q=1}^{t-1}e^{-\frac{q-1}{\tau}}\|\hat{\boldsymbol{\ell}}_{t-N-q+1}\|_\infty. \tag{37}
$$

The last result explains a challenge of the bandit-feedback setting. In the full-information setting $\|\hat{\boldsymbol{\ell}}_t\|_\infty=\|\ell_t\|_\infty\leq 1$ and we could simply bound the difference of two consecutive occupancy measures [6]. However, this does hold in the bandit-feedback setting. To overcome this challenge, we provide a new and different analysis from [6] to show that using the judicious choice of the loss estimator $\hat{\boldsymbol{\ell}}_t$, the difference $\|\boldsymbol{\rho}_{t+1}-\boldsymbol{\rho}_t\|_1$ is bounded with high probability. To this end, we use the result of Lemma 5 to bound the second term on the RHS of (37):

$$
\begin{aligned}
\sum_{t=2N}^{T}\sum_{q=1}^{t-1} e^{-\frac{q-1}{\tau}}\|\hat{\boldsymbol{\ell}}_{t-q}\|_\infty &= e^{-\frac{0}{\tau}}\sum_{t=2N}^{T}\|\hat{\boldsymbol{\ell}}_{t-N}\|_\infty + \cdots + e^{-\frac{T-2}{\tau}}\sum_{t=T}^{T}\|\hat{\boldsymbol{\ell}}_{t-T-N}\|_\infty \\
&= e^{-\frac{0}{\tau}}\sum_{t=2N}^{T-1}\|\hat{\boldsymbol{\ell}}_{t-N}\|_\infty + \cdots + e^{-\frac{T-2}{\tau}}\sum_{t=1}^{1}\|\hat{\boldsymbol{\ell}}_{t-N}\|_\infty \\
&\le \left(\sum_{q=2N}^{T-1} e^{-\frac{q-1}{\tau}}\right)\left(\sum_{t=2N}^{T-1}\|\hat{\boldsymbol{\ell}}_t\|_\infty\right) \\
&= \left(\sum_{q=2N}^{T-1} e^{-\frac{q-1}{\tau}}\right)\left(\sum_{t=2N}^{T-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\hat{\boldsymbol{\ell}}_t(s,a)\right) \\
&\le (1+\tau)\left(\sum_{t=2N}^{T-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\hat{\boldsymbol{\ell}}_t(s,a)\right) \\
&\le (1+\tau)\left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + \sum_{t=2N}^{T-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\ell_t(s,a)\right) \\
&\le (1+\tau)\left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}|\right)
\end{aligned} \tag{38}
$$

with probability at least $1-\delta_1$, where we used the τ-mixing assumption, the fact that by the definition of our loss estimator we can replace the ℓ_∞ norm with the sum over states and actions (since only one component of the estimator is nonzero), the result of Lemma 5, and the assumption that $\ell_t(s,a)\le 1$.

Finally, putting together the result established in (38) in (37) yields the stated result. ■

Lemma 7. *For any $\delta_2\in(0,1)$, with probability at least $1-\delta_2$, it holds that*

$$
\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\pi_t(a|s)\left[\hat{\boldsymbol{\ell}}_t(s,a)-\ell_t(s,a)\right] \le \frac{\log\frac{1}{\delta_2}}{2\gamma} \tag{39}
$$

Proof. Define

$$
\begin{aligned}
\alpha_t(s,a) &= 2\gamma|\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\pi_t(a|s), \\
\lambda_t(s,a) &= \frac{\gamma}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}, \\
\tilde{\mathbf{Y}}_t(s,a) &= \frac{\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}.
\end{aligned} \tag{40}
$$

for which it holds

$$
0\le \alpha_t(s,a)\tilde{\mathbf{Y}}_t(s,a)\le 2\lambda_t(s,a), \tag{41}
$$

due to non-negativity of $\alpha_t(s,a)$ and $\tilde{\mathbf{Y}}_t(s,a)$, bounded loss:

$$
\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}\le 1, \tag{42}
$$

and the fact that $0\le|\boldsymbol{\nu}_t(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\le 1$, and $0\le\pi_t(a|s)\le 1$. Furthermore, $\tilde{\mathbf{Y}}_t(s,a)$ is $\mathbb{F}$-adapted and satisfies the conditions stated in Lemma 4. Recall that the first property is that for any set $S\subset\mathcal{S}\times\mathcal{A}$ having at least two elements,

$$
\mathbb{E}\left[\prod_{(s,a)\in S}\tilde{\mathbf{Y}}_t(s,a)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\prod_{(s,a)\in S}\frac{\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}|\mathcal{F}_{t-1}\right] = 0. \tag{43}
$$

The second property states the unbiasedness of $\tilde{\mathbf{Y}}_t(s,a)$, i.e.,

$$
\mathbb{E}\left[\tilde{\mathbf{Y}}_t(s,a)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}|\mathcal{F}_{t-1}\right] = \ell_t(s,a). \tag{44}
$$

Therefore we can can use the concentration result Lemma 4 to prove the stated result, i.e.

$$\begin{aligned} &\Pr\Bigg(\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha_t(s,a)\left(\frac{\tilde{\mathbf{Y}}_t(s,a)}{1+\lambda_t(s,a)}-\ell_t(s,a)\right)\geq\log\frac{1}{\delta_2}\Bigg)\leq\delta_2 \\ \Leftrightarrow\quad &\Pr\Bigg(\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\pi_t(a|s)\left[\hat{\boldsymbol{\ell}}_t(s,a)-\ell_t(s,a)\right]\geq\frac{\log\frac{1}{\delta_2}}{2\gamma}\Bigg)\leq\delta_2. \end{aligned} \tag{45}$$

■

Lemma 8. *For any $\delta_1\in(0,1)$, with probability at least $1-\delta_1$, it holds that*

$$\sum_{t=2N}^{T}\|\boldsymbol{\nu}_{t|t-N}-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1\leq 2e^{-\frac{N-1}{\tau}}(T-2N+1)+2\eta(1+\tau)\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right). \tag{46}$$

Proof. By triangle inequality we have that

$$\|\boldsymbol{\nu}_{t|t-N}-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1\leq\|\boldsymbol{\nu}_{t|t-N}-\bar{\boldsymbol{\nu}}^{\pi_{t-1}}\|_1+\|\bar{\boldsymbol{\nu}}^{\pi_{t-1}}-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1. \tag{47}$$

By a simple recursive application of the τ-mixing assumption the first term in RHS of (47) can be bounded as

$$\begin{aligned} \|\boldsymbol{\nu}_{t|t-N}-\bar{\boldsymbol{\nu}}^{\pi_{t-1}}\|_1 &= \|\boldsymbol{\nu}_{t-1|t-N}\mathcal{P}^{\pi_{t-1}}-\bar{\boldsymbol{\nu}}^{\pi_{t-1}}\mathcal{P}^{\pi_{t-1}}\|_1 \\ &\leq e^{-\frac{1}{\tau}}\|\boldsymbol{\nu}_{t-1|t-N}-\bar{\boldsymbol{\nu}}^{\pi_{t-1}}\|_1 \\ &\leq e^{-\frac{1}{\tau}}\|\boldsymbol{\nu}_{t-1|t-N}-\bar{\boldsymbol{\nu}}^{\pi_{t-2}}\|_1+e^{-\frac{1}{\tau}}\|\bar{\boldsymbol{\nu}}^{\pi_{t-1}}-\bar{\boldsymbol{\nu}}^{\pi_{t-2}}\|_1 \end{aligned} \tag{48}$$

Regarding the second term in RHS of (47), identical to the proof of Lemma 6, using the definition of the occupancy measure we have

$$\begin{aligned} \sum_{s\in\mathcal{S}}|\bar{\nu}^{\pi_{t-1}}(s)-\bar{\nu}^{\pi_t}(s)| &= \sum_{s\in\mathcal{S}}\left|\sum_{a\in\mathcal{A}}\rho_{t-1}(s,a)-\rho_t(s,a)\right| \\ &\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\rho_{t-1}(s,a)-\rho_t(s,a)| \\ &\leq\|\boldsymbol{\rho}_{t-1}-\boldsymbol{\rho}_t\|_1. \end{aligned} \tag{49}$$

Thus, by unrolling (48) to $t-N$ we can bound $\|\boldsymbol{\nu}_t-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1$ according to

$$\|\boldsymbol{\nu}_{t|t-N}-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1\leq e^{-\frac{N-1}{\tau}}\|\boldsymbol{\nu}_{t-N|t-N}-\bar{\boldsymbol{\nu}}^{\pi_{t-N}}\|_1+\sum_{q=1}^{N-1}e^{-\frac{q-1}{\tau}}\|\boldsymbol{\rho}_{t-q}-\boldsymbol{\rho}_{t-q+1}\|_1. \tag{50}$$

Summing the above result over t yields

$$\begin{aligned} \sum_{t=2N}^{T}\|\boldsymbol{\nu}_{t|t-N}-\bar{\boldsymbol{\nu}}^{\pi_t}\|_1 &\leq\sum_{t=2N}^{T}e^{-\frac{N-1}{\tau}}\|\boldsymbol{\nu}_{t-N|t-N}-\bar{\boldsymbol{\nu}}^{\pi_1}\|_1 \\ &\quad+\sum_{t=2N}^{T}\sum_{q=1}^{N-1}e^{-\frac{q-1}{\tau}}\|\boldsymbol{\rho}^{\pi_{t-q}}-\boldsymbol{\rho}^{\pi_{t-q+1}}\|_1 \\ &\leq\sum_{t=2N}^{T}e^{-\frac{N-1}{\tau}}\|\boldsymbol{\nu}_{t-N|t-N}-\bar{\boldsymbol{\nu}}^{\pi_1}\|_1 \\ &\quad+\sum_{t=2N}^{T}\sum_{q=1}^{t-1}e^{-\frac{q-1}{\tau}}\|\boldsymbol{\rho}^{\pi_{t-q}}-\boldsymbol{\rho}^{\pi_{t-q+1}}\|_1. \end{aligned} \tag{51}$$

Since $\|\boldsymbol{\nu}_{t-N|t-N}-\bar{\boldsymbol{\nu}}^{\pi_1}\|_1\leq 2$, the first term in the RHS of (51) can be bounded by

$$2\sum_{t=2N}^{T}e^{-\frac{N-1}{\tau}}\leq 2e^{-\frac{N-1}{\tau}}(T-2N+1). \tag{52}$$

Note that to ensure this last bound is sublinear, we will later impose the requirement that $N = 1 + \lceil \tau \log T \rceil$.

Finally, bounding the second term in RHS of (51) is identical to our analysis in Lemma 8 (see (35) and the proceeding argument). $\blacksquare$

B BOUNDING TERM I

Lemma 1 (Bounding term I)**.** *For any $T \geq 1$ and any policy π, it holds that*

$$\sum_{t=1}^{T} \xi_t^\pi - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t^\pi, a_t^\pi)\right] \leq 2(1+\tau). \tag{53}$$

Proof. The Lemma, which is based on the existing result in [11], relies on the τ-mixing property of uniformly ergodic A-MDPs and holds for any $t \geq 1$.

It holds by definition of ξ_t^π that

$$\begin{aligned}
&\sum_{t=1}^{T} \xi_t^\pi - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t^\pi, a_t^\pi)\right] \\
&\quad = \sum_{t=1}^{T} \lim_{T' \to \infty} \sum_{t=1}^{T'} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \Pr\left(s_{t'}^\pi = s, a_{t'}^\pi = a\right) \ell_t(s,a) \\
&\qquad\qquad - \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \Pr\left(s_t^\pi = s, a_t^\pi = a\right) \ell_t(s,a) \\
&\quad = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \rho^\pi(s,a) \ell_t(s,a) - \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \nu_t^\pi(s) \pi(a|s) \ell_t(s,a).
\end{aligned} \tag{54}$$

Now, since by the definition of the occupancy measure and the stationary distribution $\bar{\nu}$ we have $\boldsymbol{\rho}^\pi(s,a) = \bar{\nu}^\pi(s)\pi(a|s)$,

$$\begin{aligned}
&\sum_{t=1}^{T} \xi_t^\pi - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t^\pi, a_t^\pi)\right] \\
&\quad = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \bar{\nu}^\pi(s) \pi(a|s) \ell_t(s,a) - \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \nu_t^\pi(s) \pi(a|s) \ell_t(s,a) \\
&\quad = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \left(\bar{\nu}^\pi(s) - \nu_t^\pi(s)\right) \sum_{a \in \mathcal{A}} \pi(a|s) \ell_t(s,a) \\
&\quad \leq \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \left(\bar{\nu}^\pi(s) - \nu_t^\pi(s)\right),
\end{aligned} \tag{55}$$

where we used the fact that $\sum_{a \in \mathcal{A}} \pi(a|s) = 1$ and $0 \leq \ell_t(s,a) \leq 1$. Thus, exploiting the vector notation $\boldsymbol{\nu} \in \mathbb{R}^{|\mathcal{S}|}$ and the definition of ℓ_1-norm we have

$$\begin{aligned}
\sum_{t=1}^{T} \xi_t^\pi - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t^\pi, a_t^\pi)\right] &\leq \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} |\bar{\nu}^\pi(s) - \nu_t^\pi(s)| = \sum_{t=1}^{T} \|\bar{\boldsymbol{\nu}}^\pi - \boldsymbol{\nu}_t^\pi\|_1 \\
&= \sum_{t=1}^{T} \|\bar{\nu}^\pi - \boldsymbol{\nu}_{t-1}^\pi \mathcal{P}^\pi\|_1 \leq \sum_{t=1}^{T} e^{-\frac{1}{\tau}} \|\bar{\nu}^\pi - \boldsymbol{\nu}_{t-1}^\pi\|_1,
\end{aligned} \tag{56}$$

where we used the τ-mixing assumption. Recursively repeating the above argument yields

$$\begin{aligned}
\sum_{t=1}^{T} \rho_t^\pi - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t^\pi, a_t^\pi)\right] &\leq \sum_{t=1}^{T} e^{-\frac{t-1}{\tau}} \|\bar{\boldsymbol{\nu}}^\pi - \boldsymbol{\nu}_1^\pi\|_1 \\
&\leq 2 \sum_{t=1}^{T} e^{-\frac{t-1}{\tau}} \leq 2(1 + \int_0^\infty e^{-\frac{t}{\tau}}) = 2(1+\tau).
\end{aligned} \tag{57}$$

■

C BOUNDING TERM II

Lemma 2 (Bounding term II)**.** *For any* $\delta_1, \delta_2, \delta_3, \delta_4 \in (0,1)$*, with probability at least* $1-(\delta_1+\delta_2+\delta_3+\delta_4)$*, it holds that*

$$\begin{aligned}\sum_{t=2N}^{T} \xi^{\pi_t} - \sum_{t=2N}^{T} \xi^{\pi} \leq & 4e^{-\frac{N-1}{\tau}} T + 4(1+\tau)\eta \left(\frac{\log \frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}| \right) + \\ & \frac{\log \frac{1}{\delta_2}}{2\gamma} + \gamma T |\mathcal{S}||\mathcal{A}| + \sqrt{\frac{T}{2} \log \frac{1}{\delta_3}} + \frac{\log \frac{1}{\delta_4}}{2\gamma} + \frac{1}{\eta} \log(|\mathcal{S}||\mathcal{A}|) \\ & + \frac{\eta T}{2\gamma^2} + 2\eta \frac{NT}{\gamma^2}. \end{aligned} \tag{58}$$

Proof. Recall from the definition of ξ^π and ξ^{π_t}:

$$\xi^\pi = \lim_{T' \to \infty} \sum_{t=1}^{T'} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \Pr(s_{t'}^\pi = s, a_{t'}^\pi = a) \ell_t(s,a) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \rho(s,a) \ell_t(s,a), \tag{59}$$

and

$$\xi^{\pi_t} = \lim_{T' \to \infty} \sum_{t=1}^{T'} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \Pr(s_{t'}^{\pi_t} = s, a_{t'}^{\pi_t} = a) \ell_t(s,a) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \boldsymbol{\rho}_t(s,a) \ell_t(s,a). \tag{60}$$

Thus, we can establish the following decomposition

$$\begin{aligned} \sum_{t=1}^{T} \xi^{\pi_t} - \sum_{t=1}^{T} \xi^{\pi} &= \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \boldsymbol{\rho}_t(s,a) \ell_t(s,a) - \rho(s,a) \ell_t(s,a) \\ &= \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \boldsymbol{\rho}_t(s,a) \left[\ell_t(s,a) - \hat{\boldsymbol{\ell}}_t(s,a) \right] \\ &\quad + \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \rho^\pi(s,a) \left[\hat{\boldsymbol{\ell}}_t(s,a) - \ell_t(s,a) \right] \\ &\quad + \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left[\boldsymbol{\rho}_t(s,a) - \rho^\pi(s,a) \right] \hat{\boldsymbol{\ell}}_t(s,a), \end{aligned} \tag{61}$$

by adding and subtracting $\boldsymbol{\rho}_t(s,a)\hat{\boldsymbol{\ell}}_t(s,a)$ and $\boldsymbol{\rho}^\pi(s,a)\hat{\boldsymbol{\ell}}_t(s,a)$. Now, we bound each of the terms on the RHS of (61) in the next three lemmas. Before proceeding, we further recall

$$\boldsymbol{\rho}_t(s,a) = \bar{\boldsymbol{\nu}}^{\pi_t}(s) \pi_t(a|s), \quad \bar{\boldsymbol{\nu}}^{\pi_t}(s) = \sum_{a \in \mathcal{A}} \boldsymbol{\rho}_t(s,a). \tag{62}$$

∎

C.1 BOUNDING TERM II-I

Lemma 9 (Bounding term II-I)**.** *For any* $\delta_1, \delta_2, \delta_3 \in (0,1)$*, with probability at least* $1-(\delta_1+\delta_2+\delta_3)$*, it holds that*

$$\begin{aligned} \sum_{t=2N}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \boldsymbol{\rho}_t(s,a) \left[\ell_t(s,a) - \hat{\boldsymbol{\ell}}_t(s,a) \right] \leq & 4e^{-\frac{N-1}{\tau}} T + 4(1+\tau)\eta \left(\frac{\log \frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}| \right) \\ & + \frac{\log \frac{1}{\delta_2}}{2\gamma} + \gamma T|\mathcal{S}||\mathcal{A}| + \sqrt{\frac{T}{2} \log \frac{1}{\delta_3}}. \end{aligned} \tag{63}$$

Proof. We start by adding and subtracting $\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)$ to obtain

$$\begin{aligned}
&\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\bar{\boldsymbol{\nu}}^{\pi_t}(s)\pi_t(a|s)\left[\ell_t(s,a)-\hat{\boldsymbol{\ell}}_t(s,a)\right] \\
&\qquad= \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}(\bar{\boldsymbol{\nu}}^{\pi_t}(s)-\boldsymbol{\nu}_{t|t-N}(s))\pi_t(a|s)\left[\ell_t(s,a)-\hat{\boldsymbol{\ell}}_t(s,a)\right] \\
&\qquad+ \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\left[\ell_t(s,a)-\hat{\boldsymbol{\ell}}_t(s,a)\right] := A+B
\end{aligned} \tag{64}$$

We first bound A and then we proceed to bound B.

○ **Bounding** A**.** It holds that

$$\begin{aligned}
A &= \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}(\bar{\boldsymbol{\nu}}^{\pi_t}(s)-\boldsymbol{\nu}_{t|t-N}(s))\pi_t(a|s)\ell_t(s,a) \\
&\quad+ \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}(\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s))\pi_t(a|s)\hat{\boldsymbol{\ell}}_t(s,a) \\
&\le \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\bar{\boldsymbol{\nu}}^{\pi_t}(s)-\boldsymbol{\nu}_{t|t-N}(s)|\pi_t(a|s)\ell_t(s,a) \\
&\quad+ \sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\pi_t(a|s)\hat{\boldsymbol{\ell}}_t(s,a) \\
&= 2\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\bar{\boldsymbol{\nu}}^{\pi_t}(s)-\boldsymbol{\nu}_{t|t-N}(s)|\pi_t(a|s)\ell_t(s,a) \\
&\quad+ \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\pi_t(a|s)\left[\hat{\boldsymbol{\ell}}_t(s,a)-\ell_t(s,a)\right].
\end{aligned} \tag{65}$$

The first term in (65) can be bounded by noting $0\le \ell_t(s,a)\le 1$, $\pi_t(a|s)\le 1$, and using the result of Lemma 8. Thus, it holds with probability at least $1-\delta_1$ that

$$\begin{aligned}
2\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\bar{\boldsymbol{\nu}}^{\pi_t}(s)-\boldsymbol{\nu}_{t|t-N}(s)|\pi_t(a|s)\ell_t(s,a) &\le 4e^{-\frac{N-1}{\tau}}(T-2N+1) \\
&+ 4\eta(1+\tau)\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right).
\end{aligned} \tag{66}$$

The second term in (65) can be bounded by using the concentration result established in Lemma 7. That is, with probability at least $1-\delta_2$, we have

$$\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|\boldsymbol{\nu}_{t|t-N}(s)-\bar{\boldsymbol{\nu}}^{\pi_t}(s)|\pi_t(a|s)\left[\hat{\boldsymbol{\ell}}_t(s,a)-\ell_t(s,a)\right]\le\frac{\log\frac{1}{\delta_2}}{2\gamma}. \tag{67}$$

Thus, with probability at least $1-(\delta_1+\delta_2)$ the following bound on A holds

$$A\le 4e^{-\frac{N-1}{\tau}}(T-2N+1)+4\eta(1+\tau)\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right)+\frac{\log\frac{1}{\delta_2}}{2\gamma}. \tag{68}$$

○ **Bounding** B**.** We now turn to bounding the second term in (64). By adding and subtracting $\mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)|\mathbf{h}_{t-N}]$ we can express B equivalently as

$$
\begin{aligned}
B = &\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\left[\ell_t(s,a) - \mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)|\mathbf{h}_{t-N}]\right] \\
&+ \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\left[\mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)|\mathbf{h}_{t-N}] - \hat{\boldsymbol{\ell}}_t(s,a)\right]
\end{aligned}
\tag{69}
$$

The first term in (69) can be bounded according to

$$
\begin{aligned}
&\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\left[\ell_t(s,a) - \mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)|\mathbf{h}_{t-N}]\right] \\
&\qquad= \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\ell_t(s,a)\left[1 - \frac{\mathbb{E}[\mathbb{I}\{s_t=s,a_t=a\}|\mathbf{h}_{t-N}]}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)+\gamma}\right] \\
&\qquad= \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\ell_t(s,a)\frac{\gamma}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)+\gamma} \\
&\qquad\leq \sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\ell_t(s,a)\frac{\gamma}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)} \\
&\qquad\leq \gamma(T-2N+1)|\mathcal{S}||\mathcal{A}|
\end{aligned}
\tag{70}
$$

where we used the fact that $\mathbb{E}[\mathbb{I}\{s_t=s,a_t=a\}|\mathbf{h}_{t-N}] = \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)$, $\gamma>0$, and $0\leq\ell_t(s,a)\leq 1$.

To bound the second term in (69) we will use the Azuma-Hoeffding's inequality [12] as follows. First, since we assume an oblivious adversary, it holds that $\mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)] = \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)|\mathbf{h}_{t-N}]]$. Also, note that

$$
\begin{aligned}
\mathbf{x}_t := \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\hat{\boldsymbol{\ell}}_t(s,a) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)+\gamma}\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\} \\
&\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\} \leq 1.
\end{aligned}
\tag{71}
$$

Thus, $0\leq\mathbf{x}_t\leq 1$, and in turn $|\mathbf{S}_t - \mathbf{S}_{t-1}|\leq 1$ where $\mathbf{S}_t = \sum_{i=2N}^{t}\mathbf{x}_i$. Consequently, by Azuma-Hoeffding's inequality for martingales, we can bound the difference $\mathbf{S}_T - \mathbf{S}_{2N}$ such that with probability exceeding $1-\delta_3$,

$$
\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)\left[\mathbb{E}[\hat{\boldsymbol{\ell}}_t(s,a)] - \hat{\boldsymbol{\ell}}_t(s,a)\right] \leq \sqrt{\frac{T-2N+1}{2}\log\frac{1}{\delta_3}}.
\tag{72}
$$

Therefore, we established that with probability at least $1-\delta_3$, the following bound on B holds

$$
B \leq \gamma(T-2N+1)|\mathcal{S}||\mathcal{A}| + \sqrt{\frac{T-2N+1}{2}\log\frac{1}{\delta_3}}.
\tag{73}
$$

Combining (68) and (73) to bound (64) and using $T-2N+1\leq T$ establishes the stated result. ■

C.2 BOUNDING TERM II-II

Lemma 10 (Bounding term II-II). *For any $\delta_4\in(0,1)$, with probability at least $1-\delta_4$, it holds that*

$$
\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \rho^{\pi}(s,a)\left[\hat{\boldsymbol{\ell}}_t(s,a) - \ell_t(s,a)\right] \leq \frac{\log\frac{1}{\delta_4}}{2\gamma}.
\tag{74}
$$

Proof. We resort to the concentration of $\hat{\boldsymbol{\ell}}_t(s,a)$ around $\ell_t(s,a)$. Define

$$\begin{aligned}
\alpha_t(s,a) &= 2\gamma\rho^\pi(s,a), \\
\lambda_t(s,a) &= \frac{\gamma}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}, \\
\tilde{\mathbf{Y}}_t(s,a) &= \frac{\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)},
\end{aligned} \tag{75}$$

and note $\alpha_t(s,a)$ and $\lambda_t(s,a)$ are real-valued $\mathbb{F}$-predictable random sequences satisfying

$$0 \le \alpha_t(s,a)\tilde{\mathbf{Y}}_t(s,a) \le 2\lambda_t(s,a), \tag{76}$$

because of non-negativity of $\alpha_t(s,a)$ and $\tilde{\mathbf{Y}}_t(s,a)$, the bounded loss assumption of

$$\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\} \le 1, \tag{77}$$

and $\rho^\pi(s,a) \le 1$. Furthermore, $\tilde{\mathbf{Y}}_t(s,a)$ is $\mathbb{F}$-adapted and satisfies the two conditions stated in Lemma 4, i.e.

$$\mathbb{E}\left[\prod_{(s,a)\in S}\tilde{\mathbf{Y}}_t(s,a)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\prod_{(s,a)\in S}\frac{\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}|\mathcal{F}_{t-1}\right] = 0. \tag{78}$$

and

$$\mathbb{E}\left[\tilde{\mathbf{Y}}_t(s,a)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{\ell_t(s,a)\mathbb{I}\{s_t=s,a_t=a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s)}|\mathcal{F}_{t-1}\right] = \ell_t(s,a). \tag{79}$$

Therefore we can can use Lemma 4 to prove

$$\begin{aligned}
&\Pr\left(\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha_t(s,a)\left(\frac{\tilde{\mathbf{Y}}_t(s,a)}{1+\lambda_t(s,a)}-\ell_t(s,a)\right) \ge \log\frac{1}{\delta_4}\right) \le \delta_4 \\
\Leftrightarrow\quad &\Pr\left(\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\rho^\pi(s,a)\left[\hat{\boldsymbol{\ell}}_t(s,a)-\ell_t(s,a)\right] \ge \frac{\log\frac{1}{\delta_4}}{2\gamma}\right) \le \delta_4.
\end{aligned} \tag{80}$$

■

C.3 BOUNDING TERM II-III

Lemma 11 (Bounding term II-III)**.** *It holds that with probability exceeding* $1-\delta_1$

$$\begin{aligned}
\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\boldsymbol{\rho}_t(s,a)-\rho^\pi(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a) \le{}& \frac{1}{\eta}\log(|\mathcal{S}||\mathcal{A}|) + \frac{\eta}{2\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right) \\
&+ 2\eta\frac{N}{\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right).
\end{aligned} \tag{81}$$

Proof. First, note that we can write

$$\begin{aligned}
\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\boldsymbol{\rho}_t(s,a)-\rho^\pi(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a) \le{}& \sum_{t=N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\boldsymbol{\rho}_{t+N}(s,a)-\rho^\pi(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a) \\
&+\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\boldsymbol{\rho}_t(s,a)-\boldsymbol{\rho}_{t+N}(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a).
\end{aligned} \tag{82}$$

Given the update of O-REPS-IX (see step 14 of Algorithm 1) the first term in (82) is exactly the regret of OMD with respect to a fixed occupancy measure $\rho^\pi(s,a)$. Hence, using the analysis of OMD (see, e.g., Theorem 2 in [10]),

$$\sum_{t=N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\boldsymbol{\rho}_{t+N}(s,a)-\rho^\pi(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a) \le \frac{D(\rho^\pi\|\boldsymbol{\rho}_{2N-1})}{\eta} + \frac{\eta}{2}\sum_{t=N}^{T}\|\hat{\boldsymbol{\ell}}_t\|_\infty^2. \tag{83}$$

Since

$$\hat{\boldsymbol{\ell}}_t(s,a) = \frac{\ell_t(s,a)\mathbb{I}\{s_t = s, a_t = a\}}{\boldsymbol{\nu}_{t|t-N}(s)\pi_t(a|s) + \gamma} \leq \frac{1}{\gamma}, \tag{84}$$

using Lemma 5, it holds with probability larger than $1 - \delta_1$

$$\begin{aligned}\sum_{t=N}^{T} \|\hat{\boldsymbol{\ell}}_t\|_\infty^2 &\leq \frac{1}{\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + \sum_{t=2N}^{T-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \ell_t(s,a)\right) \\ &\leq \frac{1}{\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}|\right).\end{aligned} \tag{85}$$

Additionally, since we initialize the occupancy measures uniformly in O-REPS-IX, $D(\rho^\pi\|\boldsymbol{\rho}_{2N-1})$ can be easily bounded by the definition of the Bregman divergence according to

$$\begin{aligned}\frac{1}{\eta}D(\rho^\pi\|\boldsymbol{\rho}_{2N-1}) &\leq \frac{1}{\eta}\left(R(\rho^\pi) - R(\boldsymbol{\rho}_{2N-1})\right) \leq -\frac{1}{\eta}R(\boldsymbol{\rho}_{2N-1}) \\ &\leq \frac{1}{\eta}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \boldsymbol{\rho}_{2N-1}(s,a)\log\frac{1}{\boldsymbol{\rho}_{2N-1}(s,a)} \leq \frac{1}{\eta}\log(|\mathcal{S}||\mathcal{A}|),\end{aligned} \tag{86}$$

where we exploited the fact that $R(\rho^\pi) \leq 0$ by definition. Therefore,

$$\sum_{t=N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left[\boldsymbol{\rho}_{t+N}(s,a) - \rho^\pi(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a) \leq \frac{1}{\eta}\log(|\mathcal{S}||\mathcal{A}|) + \frac{\eta}{2\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}|\right) \tag{87}$$

We now bound the second term in (82). Using a similar argument as the one used to derive (35) and (36) in Lemma 6, we can write

$$\begin{aligned}\frac{1}{2\eta}\|\boldsymbol{\rho}_{t+N} - \boldsymbol{\rho}_t\|_1^2 &\leq J^t(\boldsymbol{\rho}_t) - J^t(\boldsymbol{\rho}_{t+N}) + \langle\nabla J^t(\boldsymbol{\rho}_{t+N}), \boldsymbol{\rho}_{t+N} - \boldsymbol{\rho}_t\rangle \\ &\leq \left[\sum_{i=2N-1}^{t}\langle\boldsymbol{\rho}_t, \hat{\boldsymbol{\ell}}_i\rangle + \frac{1}{\eta}R(\boldsymbol{\rho}_t)\right] - \left[\sum_{i=2N-1}^{t}\langle\boldsymbol{\rho}_{t+N}, \hat{\boldsymbol{\ell}}_i\rangle + \frac{1}{\eta}R(\boldsymbol{\rho}_{t+N})\right] \\ &\leq \left[\sum_{i=2N-1}^{t-N}\langle\boldsymbol{\rho}_t, \hat{\boldsymbol{\ell}}_i\rangle + \frac{1}{\eta}R(\boldsymbol{\rho}_t)\right] - \left[\sum_{i=2N-1}^{t-N}\langle\boldsymbol{\rho}_{t+N}, \hat{\boldsymbol{\ell}}_i\rangle + \frac{1}{\eta}R(\boldsymbol{\rho}_{t+1})\right] \\ &\quad + \sum_{i=2N-1}^{t-N}\langle\boldsymbol{\rho}_t, \hat{\boldsymbol{\ell}}_i\rangle - \sum_{i=2N-1}^{t-N}\langle\boldsymbol{\rho}_{t+N}, \hat{\boldsymbol{\ell}}_i\rangle \\ &\leq \langle\boldsymbol{\rho}_t - \boldsymbol{\rho}_{t+N}, \sum_{i=2N-1}^{t-N}\hat{\boldsymbol{\ell}}_i\rangle \\ &\leq \|\boldsymbol{\rho}_{t+N} - \boldsymbol{\rho}_t\|_1 \|\sum_{i=2N-1}^{t-N}\hat{\boldsymbol{\ell}}_i\|_\infty.\end{aligned} \tag{88}$$

Therefore,

$$\|\boldsymbol{\rho}_{t+N} - \boldsymbol{\rho}_t\|_1 \leq 2\eta\frac{N}{\gamma}. \tag{89}$$

Thus, the second term in (82) is bounded by

$$\begin{aligned}
\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\boldsymbol{\rho}_t(s,a)-\boldsymbol{\rho}_{t+N}(s,a)\right]\hat{\boldsymbol{\ell}}_t(s,a) &\leq \sum_{t=2N}^{T}\|\boldsymbol{\rho}_{t+N}-\boldsymbol{\rho}_t\|_1\|\hat{\boldsymbol{\ell}}_t\|_\infty \\
&\leq 2\eta\frac{N}{\gamma}\sum_{t=2N}^{T}\|\hat{\boldsymbol{\ell}}_t\|_\infty \\
&\leq 2\eta\frac{N}{\gamma}\left(\sum_{t=2N}^{T}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\hat{\boldsymbol{\ell}}_t(s,a)\right) \\
&\leq 2\eta\frac{N}{\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+\sum_{t=2N}^{T-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\ell_t(s,a)\right) \\
&\leq 2\eta\frac{N}{\gamma}\left(\frac{\log\frac{1}{\delta_1}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right),
\end{aligned} \tag{90}$$

where we used Lemma 5 and the fact that by the definition of our loss estimator we can replace the ℓ_∞ norm with the sum over states and actions (since only one component of the estimator is nonzero). Therefore, combining (87) and (90) furnishes the lemma. ■

D BOUNDING TERM III

Lemma 3 (Bounding term III)**.** *For any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, it holds that*

$$\mathbb{E}\left[\sum_{t=2N}^{T} \ell_t(s_t, a_t)\right] - \sum_{t=2N}^{T} \xi_t^{\pi_t} \leq 2(1+\tau) + 2\eta(1+\tau)\left(\frac{\log\frac{1}{\delta_1}}{2\gamma} + T|\mathcal{S}||\mathcal{A}|\right). \tag{91}$$

Proof. Similar to the proof of Lemma 1, we start by expanding using the definition of $\xi_t^{\pi_t}$ according to

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2N}^{T} \ell_t(s_t, a_t)\right] - \sum_{t=2N}^{T} \xi_t^{\pi_t} \\
&\quad= \sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \Pr(s_t = s, a_t = a)\ell_t(s,a) \\
&\qquad - \sum_{t=2N}^{T} \lim_{T'\to\infty} \sum_{t=2N}^{T'} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \Pr(s_{t'}^{\pi_t} = s, a_{t'}^{\pi_t} = a)\ell_t(s,a) \\
&\quad= \sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \nu_t(s)\pi_t(a|s)\ell_t(s,a) - \sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \rho_t(s,a)\ell_t(s,a) \\
&\quad= \sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} (\nu_t(s) - \bar{\nu}^{\pi_t}(s))\pi_t(a|s)\ell_t(s,a) \\
&\quad\leq \sum_{t=2N}^{T} \sum_{s\in\mathcal{S}} (\nu_t(s) - \bar{\nu}^{\pi_t}(s)) \leq \sum_{t=2N}^{T} \|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}^{\pi_t}\|_1,
\end{aligned} \tag{92}$$

where we used the fact that $\sum_{a\in\mathcal{A}} \pi(a|s) = 1$ and $0 \leq \ell_t(s,a) \leq 1$, as well as the vector notation $\boldsymbol{\nu} \in \mathbb{R}^{|\mathcal{S}|}$ and the definition of ℓ_1-norm. The proof is then completed using the result of Lemma 6. ∎

E PROOF OF THEOREM 1

Theorem 1 (High-Probability Regret Bound)**.** *Let*

$$\eta=(T|\mathcal{S}||\mathcal{A}|)^{-2/3}\sqrt{\log(|\mathcal{S}||\mathcal{A}|)},\quad \gamma=(T|\mathcal{S}||\mathcal{A}|)^{-1/3}\sqrt{\tau\log T\log\frac{1}{\delta}},\quad \textit{and}\quad N=1+\lceil\tau\log T\rceil. \tag{93}$$

Then, for any $\delta\in(0,1)$, with probability at least $1-4\delta$, it holds that the random regret of Algorithm 1 satisfies

$$\mathcal{R}_T\le C\,(T|\mathcal{S}||\mathcal{A}|)^{2/3}\sqrt{\tau\log(|\mathcal{S}||\mathcal{A}|)\log T\log\frac{1}{\delta}}+C'\tau\log T, \tag{94}$$

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

Proof. Let $\delta=\max\{\delta_1,\delta_2,\delta_3,\delta_4\}$. Summing and rearranging the bounds over the terms in the regret decomposition, according to Lemmas 1, 2, and 3, we conclude that, with probability at least $1-4\delta$, it holds that

$$\begin{aligned}\mathcal{R}_T\le\ & 2N+2(1+\tau)+4e^{-\frac{N-1}{\tau}}T+4(1+\tau)\eta\left(\frac{\log\frac{1}{\delta}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right)+\\ & \frac{\log\frac{1}{\delta}}{2\gamma}+\gamma T|\mathcal{S}||\mathcal{A}|+\sqrt{\frac{T}{2}\log\frac{1}{\delta}}+\frac{\log\frac{1}{\delta}}{2\gamma}+\frac{1}{\eta}\log(|\mathcal{S}||\mathcal{A}|)\\ & +\frac{\eta}{2\gamma}\left(\frac{\log\frac{1}{\delta}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right)\\ & +2\eta\frac{N}{\gamma}\left(\frac{\log\frac{1}{\delta}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right)\\ & +2(1+\tau)+2\eta(1+\tau)\left(\frac{\log\frac{1}{\delta}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right).\end{aligned} \tag{95}$$

Note that $\gamma T|\mathcal{S}||\mathcal{A}|$, $\frac{1}{\eta}\log(|\mathcal{S}||\mathcal{A}|)$, and $2\eta\frac{N}{\gamma}\left(\frac{\log\frac{1}{\delta}}{2\gamma}+T|\mathcal{S}||\mathcal{A}|\right)$ are the dominant terms in the regret in terms of T, $|\mathcal{S}|$, and $|\mathcal{A}|$. Using the specified values of γ, η, and N in (13) and separating the deterministic terms $2N+4(1+\tau)$, we can bound other terms in (95) by multiples of $(T|\mathcal{S}||\mathcal{A}|)^{\frac{2}{3}}\sqrt{\tau\log(|\mathcal{S}||\mathcal{A}|)\log T\log\frac{1}{\delta}}$, thereby establishing the proof of the theorem. ■