Online Markov Decision Processes with Terminal Law Constraints

Bianca Marin Moreno   BIANCA.MARIN-MORENO@INRIA.FR
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
EDF Lab, 7 bd Gaspard Monge, 91120 Palaiseau, France
FiME Laboratoire de Finance des Marches de l'Energie - Dauphine, CREST, EDF R&D

Margaux Brégère   MARGAUX.BREGERE@EDF.FR
EDF Lab, 7 bd Gaspard Monge, 91120 Palaiseau, France
Sorbonne Université LPSM, Paris, France

Pierre Gaillard   PIERRE.GAILLARD@INRIA.FR
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Nadia Oudjane   NADIA.OUDJANE@EDF.FR
EDF Lab, 7 bd Gaspard Monge, 91120 Palaiseau, France
FiME Laboratoire de Finance des Marches de l'Energie - Dauphine, CREST, EDF R&D

Abstract

Traditional reinforcement learning usually assumes either episodic interactions with resets or continuous operation to minimize average or cumulative loss. While episodic settings have many theoretical results, resets are often unrealistic in practice. The infinite-horizon setting avoids this issue but lacks non-asymptotic guarantees in online scenarios with unknown dynamics. In this work, we move towards closing this gap by introducing a reset-free framework called the *periodic* framework, where the goal is to find *periodic policies*: policies that not only minimize cumulative loss but also return the agents to their initial state distribution after a fixed number of steps. We formalize the problem of finding optimal periodic policies and identify sufficient conditions under which it is well-defined for tabular Markov decision processes. To evaluate algorithms in this framework, we introduce the *periodic regret*, a measure that balances cumulative loss with the terminal law constraint. We then propose the first algorithms for computing periodic policies in two multi-agent settings and show they achieve sublinear periodic regret of order $\tilde{O}(T^{3/4})$. This provides the first non-asymptotic guarantees for reset-free learning in the setting of M homogeneous agents, for $M > 1$.

Keywords: Online learning, Markov decision processes, infinite-horizon, terminal law constraint.

1. Introduction

Markov Decision Processes (MDPs) have long been used to model Reinforcement Learning (RL) problems, where one or more agents interact with an environment by taking sequential actions, leading to state transitions and observed losses. Typically, the agents interact with the environment over episodes of fixed length or until a termination condition is met, after which they are reset to an initial state drawn from a fixed initial distribution, which is known as the episodic RL setting (Rosenberg and Mansour, 2019; Jin et al., 2020; Luo et al., 2021). Alternatively, the agents may operate continuously without resets, and the learner aims to minimize either the long-term average loss (infinite-horizon average-reward setting (Jaksch et al., 2008)) or the total loss over a specified number of steps (infinite-horizon setting (Neu et al., 2010)). The episodic framework for online MDPs is well studied, with many algorithms giving theoretical guarantees for adversarial losses and

unknown dynamics, but with the disadvantage of requiring agents to be restarted. On the other hand, the infinite-horizon framework is more realistic but more difficult to solve, as finite-time guarantees for online tasks have only been shown when the dynamics are known (Dai et al., 2022).

In an effort to bridge the gap between the well-studied episodic setting and the more realistic but less theoretically developed infinite-horizon setting, we introduce the *periodic* setting, a reset-free framework where the goal is to learn a policy that minimizes the total loss while ensuring that, after a predefined period of time steps, the agents return to a state distribution equal to their initial distribution. We refer to a policy that satisfies this condition (returning the agents to their initial distribution after a fixed period) as a *periodic policy* (formally defined in Sec. 2). We call each such period an episode, though unlike in episodic RL, there are no resets: learning is continuous, as in infinite-horizon RL, but with the requirement of returning to the initial distribution. This setting allows us to establish the first non-asymptotic guarantees for learning without resets in the case of M homogeneous agents with $M > 1$ (see Table 1 for a comparison with existing approaches).

Motivations. The reset-free periodic framework provides a practical approximation to infinite-horizon problems by considering a restricted, more tractable class of policies, which in turn allows finite-time guarantees in online settings with unknown dynamics. An example of a real-world problem that fits this model is that of energy demand control in smart grids. As storing energy is challenging, thermostatically controlled electrical devices, such as water-heaters, can adjust their consumption to help balance supply and demand (Moreno et al., 2025a). For that, an electricity utility sends signals through smart meters to control devices' On-Off states so that their average consumption follows a target profile consumption, with each device's state defined by its temperature and On-Off status. Daily control is needed, but episodic algorithms fail because devices cannot be reset to the same temperature distribution each day, while infinite-horizon methods fail to handle adversarial losses or unknown transitions caused by supply fluctuations and unknown user behavior. A periodic framework is better suited: devices naturally follow daily cycles, and optimizing over periodic policies enables infinite-horizon control without resets. The periodic framework also provides a natural model for tasks that inherently exhibit periodicity. Consider, for instance, autonomous robots operating in a warehouse, tasked with transporting and sorting goods. Given a list of assignments, their goal is to maximize task completion within a limited time while ensuring they autonomously return to their charging station at the end. The robot's state refers to their position, and all other factors are captured within the objective function.

Challenges. The main challenge of finding the optimal periodic policy arises when the environment dynamics are unknown, forcing the learner to rely on estimates to compute a policy. As a result, the learner cannot guarantee that the computed policy is truly periodic, *i.e.*, if it returns the agents to their initial distribution in the true environment. This can cause the agents to drift gradually over successive episodes, potentially preventing a return to the initial distribution.

Outline and contributions. In this paper, we introduce and study the problem of computing periodic policies in online MDPs with unknown transitions and adversarial loss functions, *i.e.*, that can vary arbitrarily across episodes and are unknown to the learner in advance (Even-Dar et al., 2009). In Sec. 2 we formally define the problem of finding an optimal periodic policy and identify sufficient conditions on the MDP for this problem to be well-posed. To evaluate the performance of algorithms designed for this setting, we introduce a novel metric, the *periodic regret*. In Sec. 3 and 4 we design algorithms for computing periodic policies in environments with unknown dynamics and adversarial losses, considering the case of $M > 1$ agents under two distinct settings, each based on

Framework	Resets	Transition	Algorithm	Regret
Episodic	✓	known	Zimin and Neu (2013)	$\widetilde{O}(N\sqrt{T})$
		unknown	Neu et al. (2012)	$\widetilde{O}(N^2|\mathcal{X}||\mathcal{A}|\sqrt{T})$
			Rosenberg and Mansour (2019)	$\widetilde{O}(N^2|\mathcal{X}|\sqrt{|\mathcal{A}|T})$
Infinite-horizon	✗	known	Even-Dar et al. (2009)	(ergodic) $\widetilde{O}(\tau^2\sqrt{T})$
			Chandrasekaran and Tewari (2021)	(commu.+ ineff.) $\widetilde{O}(D^2\sqrt{|\mathcal{X}|T})$
			Dai et al. (2022)	(commu. + eff.) $\widetilde{O}(|\mathcal{A}|^{1/2}(|\mathcal{X}|D)^{2/3}T^{5/6})$
			Zhao et al. (2022)	(ergodic) $\widetilde{O}(\sqrt{\tau T})$
		unknown	None	
Periodic (this work)	✗	unknown	**MDPP-K (Alg. 2)**	(Ass. 2) $\widetilde{O}((1-\alpha)^{-1/2}N^3|\mathcal{A}|^{1/4}|\mathcal{X}|^{5/4}T^{3/4})$
			MDPP-U (Alg. 3)	(Ass. 4) $\widetilde{O}((1-\alpha)^{-1}N^3|\mathcal{A}|^{1/4}|\mathcal{X}|^{5/4}T^{3/4})$

Table 1: Online MDP frameworks with theoretical guarantees under adversarial losses and stochastic transitions; all notations are in App. A, $\tau = -(\log(\alpha))^{-1}$, D is the MDP diameter, 'commu.' denotes a communicating MDP, 'ineff.' an inefficient algorithm, and 'eff.' an efficient one.

different underlying assumptions. We show both cases achieve a periodic regret of order $\widetilde{O}(T^{3/4})$ while having low computational complexity. We present empirical results in Sec. 5.

2. The learning problem

Our first contribution is to formalize the task of finding an optimal periodic policy and to establish sufficient conditions on the MDP under which this problem is well-defined. To address a broader range of tasks, such as those explored in our experiments (Sec. 5), we formulate the problem within the convex RL framework (Hazan et al., 2019; Geist et al., 2022), meaning that we consider any convex loss over the state-action distribution induced by the agent's policy, rather than linear losses (classic RL). **Notations:** For any finite set $\mathcal{S}$, we denote by $\Delta_{\mathcal{S}}$ the simplex induced by this set, and by $|\mathcal{S}|$ its cardinality. For all $d \in \mathbb{N}$, let $[d] := \{1, \ldots, d\}$. Let $\|\cdot\|_1$ be the L_1 norm, and for all $v := (v_n)_{n\in[N]}$, such that $v_n \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$, we set $\|v\|_{\infty,1} := \sup_{n\in[N]} \|v_n\|_1$. We denote the index of an episode by t, and a time step within an episode by n. A list of notations is in App. A.

2.1. Preliminaries: finite-horizon (convex) MDP

For clarity, we define here the structure of a Markov decision process (MDP) within an episode without the notation t. We define a MDP with finite state and action spaces $\mathcal{X}$ and $\mathcal{A}$, a horizon of length N, an initial state-action distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{A}}$, and a time-dependent dynamics $p := (p_n)_{n\in[N]}$, with $p_n : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$ the probability transition function at time step n where $p_n(x'|x, a)$ denotes the probability that an agent moves to state x' when performing action a at state x at time step n. The agent chooses an action at time step n by sampling from a non-stationary random Markov policy $\pi := (\pi_n)_{n\in[N]}$, with $\pi_n : \mathcal{X} \to \Delta_{\mathcal{A}}$, and $a_n \sim \pi_n(\cdot|x_n)$. We let $\Pi := (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ denote the space of policies. We define the state-action distribution sequence $\mu^{\pi,\rho} := (\mu_n^{\pi,\rho})_{n\in[N]}$ induced by playing the policy π and starting at the initial distribution ρ recursively through the following forward equation for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, with $\mu_0^{\pi,\rho}(x, a) := \rho(x, a)$, and for all $n \in [N]$,

$$\mu_{n+1}^{\pi,\rho}(x, a) := \sum_{x',a'} \mu_n^{\pi,\rho}(x', a') p_{n+1}(x|x', a') \pi_{n+1}(a|x). \tag{1}$$

3

At each time step, the learner incurs a loss $f_n(\mu_n^{\pi,\rho})$, where $f_n : \mathbb{R}^{\mathcal{X} \times \mathcal{A}} \to \mathbb{R}$ is any convex and ℓ-Lipschitz function with respect to the norm $\|\cdot\|_1$ in the state-action distribution. The learner's objective is then to find a policy π that minimizes the loss $F(\mu^{\pi,\rho}) := \sum_{n=1}^{N} f_n(\mu_n^{\pi,\rho})$. This formulation is known as the convex RL problem (Moreno et al., 2024). In the case where $f_n(\mu) = \langle \ell_n, \mu_n \rangle$ for $\ell_n : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ a loss function, we retrieve the classic RL case (Zimin and Neu, 2013). Considering a general convex function over the state-action distribution instead of a linear one increases the range of applicable scenarios, for example the energy optimization task discussed in the previous section (Coffman et al., 2023; Moreno et al., 2025a) where the loss is typically quadratic on the distribution, as well as the experiments in Sec. 5, where we compare our methods with episodic RL approaches by incorporating a regularizer into the loss function of the episodic algorithm, thereby breaking the linearity of the RL problem.

Note that F is not convex on the policy π. To convexify the problem we introduce the set of sequences of state-action distributions initialized with $\rho \in \Delta_{\mathcal{X} \times \mathcal{A}}$ and satisfying the dynamics p:

$$\mathcal{M}_\rho^p := \left\{ \mu := (\mu_n)_{n \in [N]} \; \middle| \; \mu_0 = \rho \text{ and } \sum_a \mu_{n+1}(x,a) := \sum_{x',a'} \mu_n(x',a') p_{n+1}(x|x',a') \right\}. \quad (2)$$

Since $\mathcal{M}_\rho^p$ is a set of affine constraints, it is convex (Puterman, 2014). For any $\mu \in \mathcal{M}_\rho^p$, there is a policy π such that $\mu^{\pi,\rho} = \mu$. It suffices to take $\pi_n(a|x) = \mu_n(x,a)/\sum_{a'} \mu_n(x,a')$, if the denominator is non-zero, and arbitrarily defined if not. Thus, finding an optimal policy minimizing $F(\mu^{\pi,\rho})$ (a non-convex problem) is equivalent to minimizing $F(\mu)$ over $\mathcal{M}_\rho^p$ (a convex problem).

2.2. Periodic policies

For $n \in [N]$, let $y_n := (x_n, a_n)$, and for all $\pi \in \Pi$, let $P_n^\pi(y_{n-1}, y_n) := p_n(x_n|y_{n-1})\pi_n(a_n|x_n)$ be the probability to move from y_{n-1} to y_n when following π. We introduce for all $y_0, y_N \in \mathcal{X} \times \mathcal{A}$,

$$P_\pi(y_0, y_N) := \sum_{y_1} P_1^\pi(y_0, y_1) \dots \sum_{y_{N-1}} P_{N-1}^\pi(y_{N-2}, y_{N-1}) P_N^\pi(y_{N-1}, y_N), \quad (3)$$

as the probability to move from the initial state y_0 to the final state y_N. Interpreting P_π as a $|\mathcal{X}||\mathcal{A}| \times |\mathcal{X}||\mathcal{A}|$ matrix, for any $\nu \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$, we define νP_π as the vector satisfying for each (x,a), $\nu P_\pi(x,a) = \sum_{x',a'} \nu(x',a') P_\pi(x',a';x,a)$.

We consider agents that interact repeatedly with the finite-horizon MDP over a sequence of T episodes by playing a sequence of policies $(\pi_t)_{t \in [T]}$, with each $\pi_t := (\pi_{t,n})_{n \in [N]}$ a policy over the horizon N. At the first episode, the agents are initialized following the initial state-action distribution ρ. In episodic settings, the system restarts the agents at the distribution ρ after the end of each episode. However, in real-world, resets are often prohibitive. Instead, we consider that the system never resets to the distribution ρ. Thus, at episode $t+1$, the initial distribution, that we denote by ρ_{t+1}, is equal to the final distribution on the previous episode, *i.e.*, $\rho_{t+1} := \rho_t P_{\pi_t} = \mu_N^{\pi_t,\rho_t}$.

Definition 1 (Periodic policy) *We call a policy $\pi := (\pi_n)_{n \in [N]}$ periodic in the MDP with initial state-action distribution ρ, if $\rho P_\pi = \rho$.*

The set of periodic policies can be seen as the set of policies whose stationary distribution under the MDP transition is ρ. Note that for a sequence of periodic policies, $\rho_t = \rho$ for all $t \in [T]$. In this paper, we study the online learning problem of computing a sequence of approximately *periodic* policies $(\pi_t)_{t \in [T]}$ over T episodes, aiming to minimize the total loss $\sum_{t=1}^{T} F_t(\mu^{\pi_t,\rho})$. The loss functions $(F_t)_{t \in [T]}$ are an arbitrary sequence of convex losses over the state-action distribution and

are unknown to the learner in advance. At the end of each episode t, the loss function F_t is fully revealed to the learner (full-information setting). We also assume the probability kernel p is unknown and must be learned. This problem is well-defined only under the following assumption:

Assumption 1 *The set of periodic policies of the MDP with initial distribution ρ is non-empty.*

Although Ass. 1 cannot be verified for an arbitrary ρ, one can construct a suitable ρ that satisfies it and use this distribution to run the proposed algorithms. For example, under an ergodicity assumption, running any policy for sufficiently many episodes yields an approximate stationary distribution, for which the set of periodic policies is non-empty. Moreover, some processes are inherently periodic, with the nominal policy naturally returning the state to a stationary distribution. For instance, in the energy demand control example in Sec. 1, the uncontrolled water temperature distribution in electric water-heaters follows a periodic daily cycle due to the seasonality of user behavior.

Due to uncertainties in the unknown dynamics, the learner cannot guarantee that the computed policy sequence is strictly periodic. As a result, in practice, ρ_t may deviate from ρ. To justify that we can compute a policy that induces a state-action distribution at the end of an episode close enough to ρ when starting from $\rho_t \neq \rho$ (*i.e.*, $\rho_t P_\pi \approx \rho$), we impose the following contraction assumption:

Assumption 2 *There exists some $0 \leq \alpha < 1$ such that for all periodic policies π and distributions $\nu, \nu' \in \Delta_{\mathcal{X} \times \mathcal{A}}$, $\|\nu P_\pi - \nu' P_\pi\|_1 \leq \alpha \|\nu - \nu'\|_1$.*

Ass. 2 can be understood as requiring that the Markov chain induced by any periodic policy be ergodic, *i.e.*, irreducible and aperiodic (Levin et al., 2006). Consequently, the Markov chain under a periodic policy converges to a unique stationary distribution (Ass. 2), which in this case is ρ (Ass. 1), independent of the initial state. This type of assumption is common in infinite-horizon MDPs (Even-Dar et al., 2009; Neu et al., 2010; Chen et al., 2022).

We evaluate the learner's performance through a new regret that we call the *periodic regret*, which compares the learner's total loss to that of any periodic policy π in the MDP with the initial state-action distribution ρ, and also penalizes the distance between the final distribution in episode $t-1$, ρ_t, for all $t \in \{2, \ldots, T+1\}$, and ρ, scaled by a factor of $\gamma > 0$, as follows:

$$R_T(\pi) := \sum_{t=1}^{T} \left[F_t(\mu^{\pi_t, \rho_t}) - F_t(\mu^{\pi, \rho}) \right] + \gamma \sum_{t=1}^{T} \|\rho_t - \rho\|_1. \tag{4}$$

Our aim is to find a near-optimal periodic policy, one that returns the agents to the prescribed initial distribution at the end of each episode while minimizing the cumulative loss. Accordingly, it is natural to define a static regret that compares the learner's performance to that of the best periodic policy in hindsight, augmented with a penalty term that captures deviations from the desired initial distribution. However, since a non-periodic policy may achieve a lower total loss than the optimal periodic policy, the penalty term must be scaled by a constant γ.

2.3. Learning frameworks

We consider $M > 1$ homogeneous agents and observe an independent trajectory from one of them per episode to estimate the dynamics. This independence is essential for applying standard martingale concentration bounds. Unlike the episodic setting, where restarts ensure independence, here it is obtained through trajectories observed from different agents every episode. A sufficiently large M is needed to guarantee independence across sampled trajectories. When $M = 1$, the conditional expectation of the initial distribution inducing the occupancy measure in the current episode given

the past trajectories collapses to a Dirac mass at the last observed state at time step N, instead of the true initial distribution ρ_t (see the proof of Prop. 12 for more details). This prevents the use of martingale difference sequence arguments typically employed to control the regret term associated with learning the dynamics. Introducing a second independent agent resolves this issue starting from the second episode, and the same reasoning then applies inductively across episodes.

In practice, under standard ergodicity assumptions, the system can be expected to eventually "forget" its past, making it reasonable to reuse agents in the estimation procedure after a finite number of episodes. The multi-agent assumption is therefore justified both as a natural first step toward estimating dynamics in the online infinite-horizon setting, and by the fact that it already encompasses many practical applications, including those discussed in the introduction. We leave the single-agent case for future work.

We examine two frameworks in this paper. In the first, we assume that the learner has access to the initial state-action distribution ρ_t of the agents at the beginning of each episode t. We introduce an algorithm for this setting in Sec. 3 and establish its periodic regret guarantee.

While observing ρ_t can be justified for large agent populations, it may not always be feasible in practice. To address this limitation, our second framework eliminates the assumption of direct access to ρ_t. Instead, we introduce an estimated distribution $\tilde{\rho}_t$ and apply the same algorithm from Sec. 3, replacing ρ_t with its estimate. However, this introduces additional challenges, as errors now accumulate across episodes due to both unknown transition probabilities and unknown initial distributions, motivating the need for additional assumptions. First, that the MDP must satisfy the contraction property of Ass. 2 for all policies π, not just periodic ones. Additionally, we assume that one randomly selected agent can be reset at the start of each episode, which we use *exclusively* to estimate the initial distribution ρ_t for all $t \in [T]$. Note that this is still a weaker requirement than resetting all agents to the initial distribution as is done in the episodic case. The assumption is also realistic in certain applications. For example, in the energy demand control scenario from Sec. 1, some consumers may have special electricity contracts allowing the utility to reset their thermic devices to a desired temperature distribution at the end of each day, whereas resetting all devices for all consumers would be impractical. In Sec. 4, we provide a detailed discussion of this framework.

Online protocol (Alg. 1). The learning process proceeds in episodes, but unlike episodic RL, the agents are not reset between episodes. We consider $M > 1$ independent agents that do no interact with each other. At the start of each episode t, the learner selects a policy π_t, sends it to all agents, and then observes the trajectory of an agent uniformly sampled from the M available, independently of previous observations. Using the independent trajectories observed up to episode t, the learner computes $\hat{p}_{t+1}$, an estimate of the dynamics p. For the first episode, the initial state-action pair for each agent $j \in [M]$, $(x_{1,0}^j, a_{1,0}^j)$, is drawn from the initial distribution $\rho \sim \Delta_{\mathcal{X} \times \mathcal{A}}$. In subsequent episodes, the initial state-action pair is carried over from the final step of the previous episode: $(x_{t+1,0}^j, a_{t+1,0}^j) = (x_{t,N}^j, a_{t,N}^j)$. The initial distribution at episode $t+1$ evolves according to $\rho_{t+1} = \rho_t P_{\pi_t}$. At the end of each episode, the learner observes the full loss function F_t, and computes the next policy π_{t+1} using different information depending on the framework, as illustrated in the learner protocol in Alg. 1.

3. Framework 1: known initial distributions

We first study the case where the learner observes the initial distribution at the start of each episode.

Algorithm 1 Multi-agent Online Protocol

1: **Setting:** initial policy π_1, M agents
2: **for** $t = 1, \ldots, T$ **do**
3:      All agents follow policy π_t for an episode
4:      Observe the trajectory of an agent independently sampled from the population of M agents
5:      Use the trajectory to update the estimate $\hat{p}_{t+1}$ of p

     Framework 1: known ρ_{t+1} **Framework 2: unknown** ρ_{t+1}

6:      Observe ρ_{t+1} and F_t 8:   Compute $\tilde{\rho}_{t+1}$ estimate of ρ_{t+1}, observes F_t
7:      Compute π_{t+1} using ρ_{t+1}, F_t and $\hat{p}_{t+1}$ 9:   Compute π_{t+1} using $\tilde{\rho}_{t+1}$, F_t and $\hat{p}_{t+1}$

10: **end for**

3.1. Preliminaries

We aim to formulate the learning problem as the one of finding the optimal policy per episode within a constrained MDP (C-MDP), where the final state-action distribution must satisfy an equality constraint. Leveraging the equivalence between the policy space and the space of state-action distributions satisfying the MDP dynamics as in Eq. 2, we would like to solve at episode t

$$\min_{\mu \in \mathcal{M}_{\rho_t}^p} \quad F_t(\mu), \quad \text{s.t.} \quad \mu_N := \rho_t P_\pi = \rho, \tag{5}$$

where π is the policy corresponding to the distribution μ in the sense of Eq. (1).

However, in practice, we encounter several challenges when considering Eq. (5): (1) The sequence of objective functions is adversarial. (2) The probability kernel is unknown. (3) The problem may be ill-posed because it is not necessarily feasible. Ass. 1 ensures the existence of a policy satisfying $\rho P_\pi = \rho$, but due to the uncertainties on the probability kernel, the distribution ρ_t at the end of episode $t-1$ may not exactly match ρ. Moreover, there is no guarantee that there is a policy π such that $\rho_t P_\pi = \rho$.

While numerous works in RL have addressed challenges 1 and 2 (see Table 1), they do not consider constrained problems or periodic regret. Although we model our algorithm as a C-MDP, previous works (Efroni et al., 2020; Qiu et al., 2020; Castiglioni et al., 2024; Ding and Lavaei, 2023; Stradi et al., 2024; Müller et al., 2024) differ on that they focus on either episodic or average-reward infinite-horizon tasks rather than on bounding periodic regret. Some works on the literature of stochastic control investigate constraints on the probability distribution of the terminal state (Daudin, 2020; Bouchard et al., 2010; Föllmer and Leukert, 1999; Pfeiffer, 2020; Pfeiffer et al., 2021). However, these studies primarily focus on deriving necessary optimality conditions under known dynamics. As such, they do not address challenge 1, nor 2. We present the first approach to address the three challenges with provable non-asymptotic guarantees.

3.2. Algorithm

At each episode, the algorithm uniformly and independently selects one of the M agents, and observe its trajectory that we denote by $(x_{t,n}, a_{t,n})_{n \in [N]}$. Since agents are not reset between episodes, it is essential to ensure that observations remain independent across episodes. We define $N_{t,n}(x, a) = \sum_{s=1}^{t-1} \mathbb{1}_{\{x_{s,n}=x, a_{s,n}=a\}}$, $M_{t,n}(x'|x, a) = \sum_{s=1}^{t-1} \mathbb{1}_{\{x_{s,n+1}=x', x_{s,n}=x, a_{s,n}=a\}}$, as the counters of the number of visits of the pair (x, a), and the triple (x, a, x'). The learner's estimate for

the transition kernel at the end of episode $t - 1$, to be used in episode t, is

$$\hat{p}_{t,n+1}(x'|x,a) := \frac{M_{t,n}(x'|x,a)}{\max\{1, N_{t,n}(x,a)\}}. \qquad (6)$$

From Eq. (2), we let $\mathcal{M}_t := \mathcal{M}_{\rho_t}^{\hat{p}_t}$. For any initial distribution $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, we define $\hat{\mu}_t^{\pi,\nu} := (\hat{\mu}_{t,n}^{\pi,\nu})_{n \in [N]}$ the sequence of state-action distributions induced by the policy π in the estimated MDP $\hat{p}_t$ at episode t, where for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\hat{\mu}_{t,0}^{\pi,\nu}(x,a) := \nu(x,a)$, and for all $n \in [N]$,

$$\hat{\mu}_{t,n+1}^{\pi,\nu}(x,a) := \sum_{x',a'} \hat{\mu}_{t,n}^{\pi,\nu}(x',a')\hat{p}_{t,n+1}(x|x',a')\pi_{n+1}(a|x).$$

We present a bonus-based exploration iterative algorithm for adversarial convex constrained MDPs in Eq. (8). To address the challenge of handling convex adversarial objective functions, the learner performs a Mirror Descent-like (MD) (Beck and Teboulle, 2003) iteration in each episode t. Given that the true transition kernel is unknown, the learner solves the optimization problem over the set of state-action distributions induced by the current estimated MDP $\hat{p}_t$.

To ensure accurate estimation of p, we need to properly explore the environment. To achieve this, we adjust the gradient of the objective function in MD by subtracting a sequence of vectors $\bar{b}_t := (\bar{b}_{t,n})_{n \in [N]}$, which we denote by bonus vectors, where $\bar{b}_{t,n} \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ is defined in Eq. (7). To ensure the feasibility of the problem, we must carefully construct the constraint set to consider at each episode. Ass. 1 guarantees the existence of a periodic policy in the true MDP, starting from the distribution ρ. However, this guarantee does not extend to the estimated MDP, as we cannot ensure that the same policy remains periodic there. To address this, we transform the initial equality constraint in the terminal law in an inequality constraint using a second bonus vector $b_t := (b_{t,n})_{n \in [N]}$ also defined in Eq. (7). Recall that ℓ is the Lipschitz constant of $f_{t,n}$ with respect to $\|\cdot\|_1$ for all $t \in [T]$ and $n \in [N]$. We set

$$\bar{b}_{t,n}(x,a) = \frac{\ell(N-n)C_\delta}{\sqrt{\max\{1, N_{t,n}(x,a)\}}}, \quad \text{and} \quad b_{t,n}(x,a) = \frac{C_\delta}{\sqrt{\max\{1, N_{t,n}(x,a)\}}}, \qquad (7)$$

where $C_\delta = (2|\mathcal{X}| \log(|\mathcal{X}||\mathcal{A}|NT/\delta))^{1/2}$, for $\delta \in (0,1)$.

The structure of the bonus terms can be interpreted as follows. The term $\bar{b}$ is subtracted from the MD subgradient in Eq. (8) to encourage exploration: intuitively, a state-action pair (x,a) that has been visited only a few times (and thus has a large $\bar{b}_n(x,a)$) is treated as having a lower loss, which promotes further sampling. The specific form of $\bar{b}$ follows from the concentration bound on the L_1 distance between the true and estimated transition dynamics established in Lemma 9, and the underlying intuition is analogous to the role of UCB bonuses in multi-armed bandits (Lattimore and Szepesvári, 2020). The bonus term b is added to the constraints of the MD scheme to ensure feasibility of the optimization problem with respect to the estimated MDP. It explicitly accounts for the uncertainty in the transition model p. More details on the structure of both bonus terms becomes clear through the regret analysis.

Adding a bonus to the initial constraint raises another feasibility issue, as we cannot guarantee that the policy computed satisfies the equality constraint on the terminal distribution when applied in practice. This implies that the initial distribution at episode t, ρ_t, may differ from ρ, and we cannot ensure the existence of a policy π that brings ρ_t back to ρ, *i.e.*, $\rho_t P_\pi = \rho$. To tackle this issue, we make use of Ass. 2 to build a feasible constraint set as we later show in Lemma 2.

Let π_t be the policy played at episode t. For short let $\mu_t := \hat{\mu}_t^{\pi_t, \rho_t}$, and $\ell_t := \nabla F_t(\mu_t)$. Let D_ψ be any Bregman divergence over a sequence of state-action distributions, with ψ the function

inducing it. Let $0 \leq \bar{\alpha} < 1$. At episode $t+1$, with some $\eta > 0$ to be tuned later, the learner computes

$$
\begin{aligned}
\mu_{t+1} \in \underset{\mu \in \mathcal{M}_{t+1}}{\arg\min} \quad & \left\{ \eta \langle \ell_t - \bar{b}_t, \mu \rangle + D_\psi(\mu, \mu_t) \right\} \\
\text{s.t.} \quad & \|\mu_N - \rho\|_1 \leq \langle \mu, b_t \rangle + \bar{\alpha} \|\rho_t - \rho\|_1,
\end{aligned} \tag{8}
$$

and sends the policy π_{t+1} associated with μ_{t+1} to all agents. We show that the constraint set is sufficiently large to ensure that the problem solved at each iteration is feasible (see Lemma 2), yet sufficiently small to ensure that the cumulative distance between the initial distribution at each episode, ρ_t, and the desired distribution, ρ, remains small (see Lemma 3). This balance is crucial for analyzing the periodic regret in Sec. 3.3. The proof of both Lemmas is in App. E. We define MDPP-K (Mirror Descent for Periodic Policies - Known initial distributions) in Alg. 2 at App. C as a method that solves one iteration of Eq. (8) at each episode t.

Lemma 2 *For any $\bar{\alpha} \geq \alpha$, for α as in Ass. 2, the problem in Eq. (8) is feasible with high-probability.*

Lemma 3 *Let π_t be the policy obtained from μ_t solution of Eq. (8) at episode t. Recall that $\rho_1 = \rho$, and for all $t \in [T]$, $\rho_{t+1} = \rho_t P_{\pi_t}$. Then, with high probability,*

$$
\sum_{t=1}^{T} \|\rho_{t+1} - \rho\|_1 \leq \tilde{O}\left(\frac{N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right).
$$

Practical solution of Problem (8). Our theoretical analysis assumes that Problem (8) can be solved optimally, which we can do in certain special cases of the Bregman divergence. We outline the key idea behind the practical solution, with full details provided in App. H. This approach is used in the experiments presented in Sec. 5. We adopt a Lagrangian approach that dualizes the constraint on the final distribution, leading to the following Lagrangian formulation:

$$
\mathcal{L}_t(\mu, \lambda) := \langle \ell_t - \bar{b}_t, \mu \rangle + \tfrac{1}{\eta} D_\psi(\mu, \mu_t) + \lambda \left[\|\mu_N - \rho\|_1 - \langle \mu, b_t \rangle - \bar{\alpha} \|\rho_t - \rho\|_1 \right],
$$

where $\lambda \in \mathbb{R}_+$ is the Lagrange multiplier. From the feasibility result of Lemma 2, we know that the optimal solutions to $\max_{\lambda > 0} \min_{\mu \in \mathcal{M}_{t+1}} \mathcal{L}_t(\mu, \lambda)$ are a pair of optimal primal and dual variables. For a fixed $\lambda > 0$ and for a specific Bregman divergence introduced in Eq. (16), it can be shown that the policy inducing the optimal distribution in the unconstrained problem $\min_{\mu \in \mathcal{M}_{t+1}} \mathcal{L}_t(\mu, \lambda)$ has a closed-form solution stated in Eq. (24) (Neu et al., 2017). Thus, we adopt a min-max approach: we initialize the Lagrange multiplier, compute the corresponding distribution by solving the associated unconstrained minimization problem, and then update the Lagrange multiplier using a gradient ascent step. This process continues iteratively until the resulting distribution satisfies the constraints.

Lemma 2 established that the problem in Eq. (8) is feasible for any $\bar{\alpha} \geq \alpha$. In practice, the contraction parameter α may be unknown. Nonetheless, since the algorithm only requires a value $\bar{\alpha}$ that guarantees feasibility, a practical solution is to construct a grid over the interval $(0, 1)$ and choose the smallest $\bar{\alpha}$ for which the problem is feasible. To do this, we can search over an exponentially spaced grid $\{\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \ldots\}$, and stop once the problem becomes feasible for the current value. This procedure yields a computational cost of $\log\left(\frac{1}{1-\alpha}\right)$ per episode. To test feasibility, the L_1 constraint can be reformulated as a set of linear inequalities, allowing a solver to efficiently check feasibility. This introduces only a factor of 2 in the regret, and in practice may perform even better as we may find a value $\bar{\alpha} < \alpha$ for which the problem is already feasible.

3.3. Periodic regret analysis

The main result is in Thm. 4, with its proof in App. G. We highlight here the novelties of our analysis.

Theorem 4 *Running MDPP-K (Alg. 2) for T episodes in an MDP with unknown transition kernel p and initial state-action distribution ρ, against adversarial objectives $F_t := \sum_{n=1}^{N} f_{t,n}$, where each $f_{t,n}$ is convex and ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$, with parameter α, the optimal choice of η, $\|\nabla\psi(\mu_t)\|_{1,\infty} \le \Psi$, guarantees that, for any periodic policy π, with high probability,*

$$R_T(\pi) \le \tilde{O}\big(\ell N^3 |\mathcal{X}|^{5/4} |\mathcal{A}|^{1/4} \sqrt{\Psi(1-\alpha)^{-1}} T^{3/4} + (1-\alpha)^{-1} \ell N^4 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\big).$$

To obtain this result, we begin by decomposing the periodic regret into two components:

$$R_T(\pi) = \underbrace{\sum_{t=1}^{T} F_t(\mu^{\pi_t,\rho_t}) - F_t(\hat{\mu}_t^{\pi_t,\rho_t})}_{R_T^{\text{MDP}}} + \underbrace{\sum_{t=1}^{T} F_t(\hat{\mu}_t^{\pi_t,\rho_t}) - F_t(\mu^{\pi,\rho})}_{R_T^{\text{policy}}} + \gamma \sum_{t=1}^{T} \|\rho - \rho_t\|_1.$$

We show that with high probability $R_T^{\text{MDP}} = \tilde{O}(\sqrt{T})$. As this analysis generalizes conventional techniques (Rosenberg and Mansour, 2019), we leave the details in the appendix. Recall that $\ell_t := \nabla F_t(\mu_t)$ and $\mu_t := \hat{\mu}_t^{\pi_t,\rho_t}$. Using the convexity of F_t, we further upper bound R_T^{policy} with

$$\underbrace{\sum_{t=1}^{T} \langle \ell_t - \bar{b}_t, \mu_t - \hat{\mu}_t^{\pi,\rho_t} \rangle}_{R_T^{\text{MD}}} + \underbrace{\sum_{t=1}^{T} \langle \ell_t - \bar{b}_t, \hat{\mu}_t^{\pi,\rho_t} - \hat{\mu}_t^{\pi,\rho} \rangle}_{R_T^{\text{diff. } \rho_t}} + \underbrace{\sum_{t=1}^{T} \langle \bar{b}_t, \mu_t - \hat{\mu}_t^{\pi,\rho} \rangle + \langle \ell_t, \hat{\mu}_t^{\pi,\rho} - \mu^{\pi,\rho} \rangle}_{R_T^{\text{bonus}}}.$$

The bonus vector added to the gradient term in Eq. (8) ensures that the term $R_T^{\text{bonus}} = \tilde{O}(\sqrt{T})$, and its analysis is generalized from the one in (Moreno et al., 2025b), hence the details are left in App. G. The novelties of our analysis lie in the way we treat the term R_T^{MD} and the ρ_t variation in $R_T^{\text{diff. } \rho_t}$.

$R_T^{\textbf{diff. } \rho_t}$ **term.** This term accounts for the deviation of the initial distribution in each episode from the original target. In the episodic setting, where all agents reset at the end of each episode, this term does not appear. We upper bound it in Prop. 5, with the proof in App. G.3.

Proposition 5 *With high probability we have that*

$$R_T^{\text{diff. } \rho_t} \le \tilde{O}\big((1-\alpha)^{-1} \ell N^4 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\big).$$

$R_T^{\textbf{MD}}$ **term.** This term measures the quality of our computations using an MD scheme, with an upper bound in Prop. 6 and proof in App. F. Unlike in standard MD, this term is more challenging because each iteration uses an MDP induced by a different transition estimate $\hat{p}_t$, starting from a different initial distribution ρ_t. We assume $\|\nabla\psi(\mu_t)\|_{1,\infty} \le \Psi$. This can be guaranteed in practice by mixing the output policy with a uniform policy, adding only logarithmic terms on T to the regret. We do not elaborate on it further in this paper, as the technique is already described in works such as Moreno et al. (2025b) and requires only straightforward additional calculations.

Proposition 6 *For each episode $t \in [T]$, let μ_t denote the solution of the optimization problem in Eq. (8). Suppose further that $\|\nabla\psi(\mu_t)\|_{1,\infty} \le \Psi$, and let ℓ be the Lipschitz constant of $f_{t,n}$ with respect to $\|\cdot\|_1$ for all $t \in [T]$ and $n \in [N]$. Then, with high probability,*

$$R_T^{\text{MD}} \le \tilde{O}\big(\ell N^3 |\mathcal{X}|^{5/4} |\mathcal{A}|^{1/4} \sqrt{\Psi(1-\alpha)^{-1}} T^{3/4}\big).$$

Discussion. We remark that the factor $(1 - \alpha)^{-1}$ scales polynomially with the mixing time τ, commonly used to bound regret in infinite-horizon settings (see Table 1). The $T^{3/4}$ dependency of MDPP-K comes from applying OMD over state-action distribution sequences initialized with varying distributions ρ_t per episode. Lower bounds for the adversarial infinite-horizon problem are known only in the case of known dynamics, yielding $O(\sqrt{T})$ in the full information setting (Chandrasekaran and Tewari, 2021). For unknown dynamics, establishing a lower bound remains an open problem.

Regarding the dependence of the regret on the MDP parameters, it is known that exploration bonuses lead to an upper bound with an N^3 dependence (Luo et al., 2021), introducing an extra factor of N compared to the episodic RL lower bound. In our setting, we show that MDPP-K exhibits the same dependence. Concerning the dependence on the number of states, prior work (Moreno et al., 2025b) shows that, in episodic RL, combining exploration bonuses with OMD in the occupancy-measure space yields a $|\mathcal{X}|^{3/2}$ dependence, an additional $\sqrt{|\mathcal{X}|}$ over the lower bound arising from the constants in the bonus terms (see Eq. (7)). In our setting, OMD must additionally handle changing initial state distributions, and the exploration bonuses appearing in the constraints influence these variations. As a result, the same constant is incurred once more, leading to an overall $|\mathcal{X}|^{5/4}$ dependence.

The computational complexity of MDPP-K increases linearly with $|\mathcal{X}|$, $|\mathcal{A}|$, and N, due to the structure of the closed-form output policy in Eq. (24). When we refer to the method as having low computational complexity, we mean that it admits a closed-form solution, as detailed in App. H, so no additional optimization step or external solver is required.

4. Framework 2: unknown initial distributions

In this setting, the initial distribution ρ_t is unknown. Thus, we approximate it by $\tilde{\rho}_t$. We then apply the iterative scheme from Eq. (8) as if $\tilde{\rho}_t$ were the true initial distribution. As $\rho_t = \rho_{t-1} P_{\pi_{t-1}}$, an accurate estimation relies on a good transition kernel estimate. In Framework 1, we needed a transition kernel estimate $\hat{p}_t$ that is accurate on the state–action pairs visited by agents initialized from the *true* initial distribution ρ_t under the computed policy π_t (see the analysis of R_T^{MDP} in App. 17). Since agents without resets are initialized from ρ_t, observing independent trajectories in each episode was sufficient in Sec. 3.

However, using $\hat{p}_t$ to estimate the initial distribution ρ_t does not yield a good approximation. Let $\tilde{\rho}_t$ denote its estimate at episode t. To reliably estimate the initial distribution for episode $t + 1$, we need a transition kernel estimate accurate over the state–action pairs visited by an agent initialized from the previously *estimated* initial distribution $\tilde{\rho}_t$ (see the proof of Prop. 5 in App. I). The difficulty is that, without resets, agents actually start from the true distribution ρ_t, not the estimated one $\tilde{\rho}_t$.

To address this challenge, we introduce two additional assumptions. First, we strengthen Ass. 2 so that it applies to all policies, not just periodic ones, which corresponds to the standard ergodicity assumption often used in infinite-horizon MDPs (Chen et al., 2022). Second, at the start of each episode t, we allow a randomly selected agent to be restarted from a state–action pair sampled according to the *estimated* initial distribution $\tilde{\rho}_t$. This agent can be the same each episode or vary across episodes.

Algorithm overview. Let $\tilde{\rho}_1 = \rho$. For all $t \geq 1$, in the start of episode t, we reset a randomly selected agent to a initial state-action pair sampled from $\tilde{\rho}_t$. We observe the trajectory followed by the restarted agent at episode t, and define $\tilde{p}_{t+1}$, a second empirical estimation of p, such as in Eq. (6), but using the trajectories of the restarted agent instead. We denote by $\widetilde{P}_{\pi_t}^t$ the stochastic

matrix induced by the transition kernel $(\tilde{p}_{t,n})_{n\in[N]}$ and the policy π_t, as defined in Eq. (3). We take $\tilde{\rho}_{t+1} = \tilde{\rho}_t \widetilde{P}^t_{\pi_t}$. Prop. 7 provides a key result on the accuracy of this estimate and is the main result for extending the methods of the previous section. The proof and additional details are in App. I. The final method for this setting is MDPP-U (Mirror Descent for Periodic Policies - Unknown initial distributions) in Alg. 3 in App. C. We state the main result of MDPP-U in Thm. 8.

Proposition 7 *For any $\delta \in (0,1)$, with probability at least $1 - 2\delta$,*

$$\sum_{t=1}^{T} \|\rho_t - \tilde{\rho}_t\|_1 \le (1-\alpha)^{-1}\big[3N|\mathcal{X}|\sqrt{2|\mathcal{A}|T\log\left(|\mathcal{X}||\mathcal{A}|NT/\delta\right)} + 2N|\mathcal{X}|\sqrt{2T\log\left(N/\delta\right)}\big].$$

Discussion. Deriving theoretical results in complex settings usually starts with simplifying assumptions. In the online infinite-horizon setting, non-asymptotic guarantees have so far been obtained only under known dynamics, since estimating them without resets can lead to errors accumulating linearly as agents drift from their initial distribution. In our analysis in Sec. 3, we show that if the initial distribution is known at each episode, with $M > 1$ homogeneous agents and a well-posed constrained problem, the dynamics can be reliably estimated from independent trajectories. When the initial distribution is unknown, additional exploration strategies are necessary. We demonstrate that allowing a single random agent to be reset each episode from a given distribution is sufficient to recover asymptotic guarantees even with unknown dynamics. This requirement is far weaker than resetting all agents simultaneously: for example, in the energy demand control scenario of Sec. 1, some consumers may have contracts permitting utilities to reset their water-heater temperature daily, whereas resetting all devices is impractical. Moreover, we believe that our analysis can provide insights for designing less constrained exploration strategies in future works.

Regarding the regret dependencies in Thm. 8, since the MDPP-U algorithm does not assume knowledge of the initial distribution at each episode, an additional quantity must be learned. This results in an increased dependence on the mixing time in the regret bound. Nevertheless, this dependence remains polynomial.

Theorem 8 *Running MDPP-U (Alg. 3) for T episodes in an MDP with unknown transition kernel p and initial state-action distribution ρ, against adversarial objectives $F_t := \sum_{n=1}^{N} f_{t,n}$, where each $f_{t,n}$ is convex and ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$, with parameter α, the optimal choice of η, $\|\nabla\psi(\mu_t)\|_{1,\infty} \le \Psi$, guarantees that, for any periodic policy π, with high probability,*

$$R_T(\pi) \le \tilde{O}\big(\ell N^3(1-\alpha)^{-1}|\mathcal{X}|^{5/4}|\mathcal{A}|^{1/4}\Psi^{1/2}T^{3/4} + \ell N^4(1-\alpha)^{-2}|\mathcal{X}|^2\sqrt{|\mathcal{A}|T}\big).$$

5. Experiments

We evaluate MDPP-K and MDPP-U on the *max-entropy* and *obstacle* tasks from Geist et al. (2022), without resets and with the added goal of returning agents to the initial state at the end of the episode. These tasks use a fixed objective. Adversarial MDPs are harder to implement due to the challenge of finding optimal stationary policies and limited benchmarks. The tasks under consideration could represent, for example, autonomous robots in a warehouse aiming to maximize task completion while ensuring they return to their charging stations at the end of each episode. For instance, in the *obstacle* experiment, the robot's objective is to reach a yellow target (representing a task) and then autonomously return to its initial state (the charging station) by the episode's end.[1]

---

1. All the code to reproduce the empirical results is available at: https://github.com/biancammoreno/OnlineMDPConstraintsLaw/.
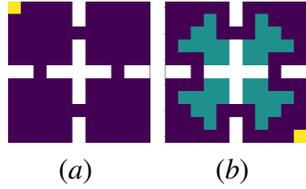
(a)   (b)

Figure 1: [left] Initial agent dist.; [right] *Obstacles* (reward in yellow, constraints in blue).

(a)   (b)
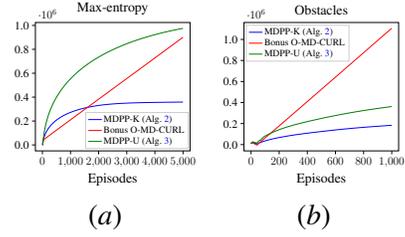
Figure 2: Periodic regret for each environment.

(a) $n = 1$ (b) $n = 14$ (c) $n = 24$ (d) $n = 30$ (e) $n = 40$

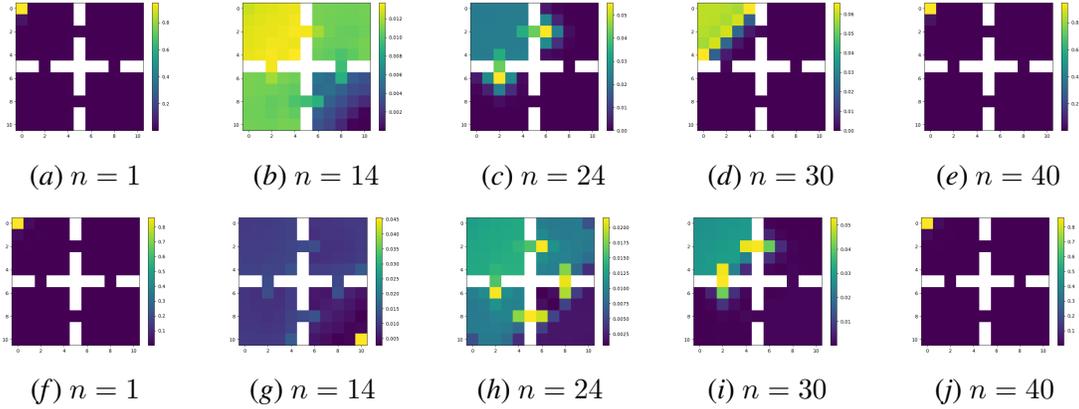(f) $n = 1$ (g) $n = 14$ (h) $n = 24$ (i) $n = 30$ (j) $n = 40$

Figure 3: *Max-entropy*: state dist. after 5000 eps. of MDPP-K [up] and Bonus MD-CURL [down].

We compare our periodic framework with the episodic framework, using the Bonus O-MD-CURL baseline from Moreno et al. (2025b), adding a regularization term $\gamma\|\rho_t - \rho\|_1$ to the objective of the episodic algorithm to allow fair comparison. Note that incorporating a L_1 regularizer with respect to the final distribution is only possible for an episodic algorithm in the convex RL case, as this regularizer is not a linear function of the state-action distribution.

The environment is an 11×11 four-room grid world, where rooms are connected by single doors. The agent moves in cardinal directions or stays in place: $x_{n+1} = x_n + a_n + \varepsilon_n$, where ε_n adds random displacement to a neighboring state. The initial state distribution is a Dirac located in the upper-left corner of the grid, as in Fig. 1 [left]. We set the learning rate $\eta = 0.01$, the dual learning rate $\eta_\lambda = 0.01$, the contraction parameter $\alpha = 0.1$, and $\gamma = 1000$. *Max-entropy:* The objective is to maximize entropy, aiming to reach a uniform distribution over the state space. At each step, $f_{t,n}(\mu_n^{\pi,\rho_t}) := \langle \mu_n^{\pi,\rho_t}, \log(\mu_n^{\pi,\rho_t}) \rangle$, and we take $N = 40$. *Obstacles:* The goal is to concentrate the state distribution on the yellow target in Fig. 1 [right] while avoiding the constraint states in blue. The objective function is defined as $f_{t,n}(\mu_n^{\pi,\rho_t}) := -\langle r, \mu_n^{\pi,\rho_t} \rangle + (\langle \mu_n^{\pi,\rho_t}, c \rangle)^2$, where $r, c \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{A}|}$. Here, r and c are zero everywhere except at the target and constraint states respectively, and $N = 80$.

We plot the state distributions at different time steps in Fig. 3, and Fig. 4, comparing MDPP-K [up] with Bonus O-MD-CURL [down]. For *max-entropy*, both methods lead agents to reach their goals midway through the episode and return to the initial state by the end. For the *obstacle* task, MDPP-K reaches the target and then returns to the initial state, whereas the episodic approach becomes trapped in the target square. To compute the optimal periodic policy in the periodic regret, we solve the constrained problem in Eq. (5) assuming known dynamics. In Fig. 2 we see that

13

(a) $n = 1$ (b) $n = 20$ (c) $n = 30$ (d) $n = 70$ (e) $n = 80$

(f) $n = 1$ (g) $n = 20$ (h) $n = 40$ (i) $n = 70$ (j) $n = 80$
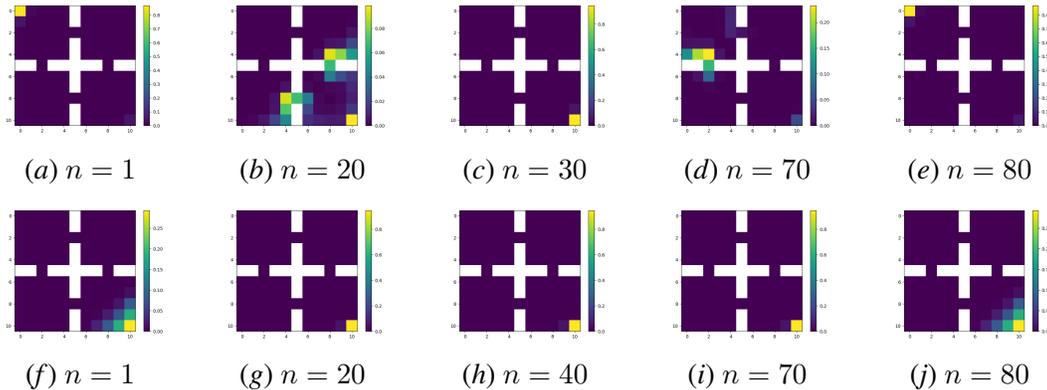
Figure 4: *Obstacles*: state dist. after 1000 eps. for MDPP-K [up] and Bonus MD-CURL [down].

MDPP-K and MDPP-U both achieve sub-linear regret, with MDPP-U suffering higher regret as it also has to estimate ρ_t. In contrast, Bonus O-MD-CURL incurs linear regret on both tasks.

Regularizers perform poorly in an episodic algorithm because such algorithms assume agents are reset at the end of each episode, an assumption that does not hold in our experiments. This leads to practical errors and makes tuning the regularization parameter difficult. For *max-entropy*, we observe that the regularizer $\gamma \|\rho_t - \rho\|_1$ is too strong, preventing correct optimization, while that in the *obstacle* task, it is to weak, failing to ensure a return to the initial state. Since agents actually start from a different distribution ρ_t rather than the assumed ρ, a fixed regularizer is ineffective; an adaptive γ_t would be needed, but computing it is infeasible without known dynamics. Our method, in contrast, effectively implements an implicit adaptive regularization, handling imperfect resets by automatically adjusting the Lagrange multipliers at each episode.

Acknowledgments

References

Francis Bach. *Learning Theory from First Principles*. The MIT Press, 2024. ISBN 9780262049443.

Peter L. Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, page 35–42, 2009.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, may 2003. ISSN 0167-6377. doi: 10.1016/S0167-6377(02)00231-6.

Bhavya, Lenart Treven, Florian Dorfler, Stelian Coros, and Andreas Krause. Neorl: Efficient exploration for nonepisodic rl. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 74966–74998, 2024.

Bruno Bouchard, Romuald Elie, and Cyril Imbert. Optimal control under stochastic target constraints. *SIAM J. Control Optim.*, 48(5):3501–3531, February 2010. ISSN 0363-0129.

Thibaut Bourdais, Nadia Oudjane, and Francesco Russo. An entropy penalized approach for stochastic optimization with marginal law constraints. complete version, 2025.

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, March 2003.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024. ISBN 9781713871088.

Gautam Chandrasekaran and Ambuj Tewari. Learning in online mdps: is there a price for handling the communicating case? In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision processes with constraints. In *International Conference on Machine Learning*, 2022.

Austin Coffman, Ana Bušić, and Prabir Barooah. A unified framework for coordination of thermostatically controlled loads. *Automatica*, 152:111002, 2023.

Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial markov decision processes with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 11437–11449, 2022.

Samuel Daudin. Optimal control of diffusion processes with terminal constraint in law. *Journal of Optimization Theory and Applications*, 195:1 – 41, 2020.

Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. In *AAAI Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence*, 2023.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps, 2020.

Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009. ISSN 0364765X, 15265471.

Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations*, 2018.

Hans Föllmer and Peter Leukert. Quantile hedging. *Finance and Stochastics*, 3(3):251–273, 1999.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, page 489–497, 2022.

Ivan Guo, Nicolas Langrené, Grégoire Loeper, and Wei Ning and. Portfolio optimization with a prescribed terminal wealth distribution. *Quantitative Finance*, 22(2):333–347, 2022. doi: 10.1080/14697688.2021.1967432.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, volume 97, pages 2681–2691, 2019.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2008.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, volume 119, pages 4860–4869, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. In *Advances in Neural Information Processing Systems*, volume 34, pages 22931–22942, 2021.

Bianca Moreno, Margaux Bregere, Pierre Gaillard, and Nadia Oudjane. Efficient model-based concave utility reinforcement learning through greedy mirror descent. In *International Conference on Artificial Intelligence and Statistics*, volume 238, pages 2206–2214, 02–04 May 2024.

Bianca Moreno, Margaux Brégère, Pierre Gaillard, and Nadia Oudjane. (Online) Convex Optimization for Demand-Side Management: Application to Thermostatically Controlled Loads. *J. Optim. Theory Appl.*, 205(3), 2025a. ISSN 0022-3239. doi: 10.1007/s10957-025-02658-9.

Bianca Moreno, Khaled Eldowa, Pierre Gaillard, Margaux Brégère, and Nadia Oudjane. Online episodic convex reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. In *International Conference on Machine Learning*, 2024.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvari. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 805–813, 2012.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. In *Conference on Neural Information Processing Systems*, 2017.

Laurent Pfeiffer. Optimality conditions in variational form for non-linear constrained stochastic control problems. *Mathematical Control and Related Fields*, 10(3):493–526, 2020. ISSN 2156-8472. doi: 10.3934/mcrf.2020008.

Laurent Pfeiffer, Xiaolu Tan, and Yu-Long Zhou. Duality and approximation of stochastic optimal control problems under expectation constraints. *SIAM Journal on Control and Optimization*, 59 (5):3231–3260, 2021. doi: 10.1137/20M1349886.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 15277–15287, 2020.

Aviv A. Rosenberg and Y. Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, 2019.

Shai Shalev-Shwartz. *Online Learning and Online Convex Optimization*. Now Publishers, 2012. doi: 10.1561/2200000018.

Archit Sharma, Rehaan Ahmad, and Chelsea Finn. A state-distribution matching approach to non-episodic reinforcement learning. In *International Conference on Machine Learning*, 2022a.

Archit Sharma, Kelvin Xu, Nikhil Sardana, Abhishek Gupta, Karol Hausman, Sergey Levine, and Chelsea Finn. Autonomous reinforcement learning: Formalism and benchmarking. In *International Conference on Learning Representations*, 2022b.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs: Handling stochastic and adversarial constraints. In *International Conference on Machine Learning*, volume 235, pages 46692–46721, 2024.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online Markov decision processes. In *International Conference on Machine Learning*, volume 162, pages 26865–26894, 2022.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Appendix

This appendix is organized as follows:
The code for reproducing the experiments in Sec. 5 can be found at: https://anonymous.4open.science/r/OnlineMDPConstraintsLaw-1115

Appendix A. List of Notations

In this appendix, we recall all notations that are used throughout the paper.

Below are generic notations for any finite sets $\mathcal{S}, \mathcal{D}$:

- $[s] := \{1, \ldots, s\}$

- $|\mathcal{S}| =$ cardinality of a set $\mathcal{S}$

- $\Delta_{\mathcal{S}} := \left\{ v \in \mathbb{R}^{|\mathcal{S}|} \mid \sum_{i=1}^{|\mathcal{S}|} v(i) = 1 \right\}$

- $(\Delta_{\mathcal{S}})^{\mathcal{D}} := \left\{ v = (v_d)_{d \in [D]} \mid v_d \in \mathbb{R}^{|\mathcal{S}|}, \ \sum_{i=1}^{|\mathcal{S}|} v_d(i) = 1, \forall d \in [D] \right\}$

- Given functions $f, g : \mathbb{N} \to [0, \infty)$ define $f(n) = O\big(g(n)\big) \Leftrightarrow \limsup_{n \to \infty} \frac{f(n)}{g(n)} < \infty$

- $f(n) = \tilde{O}(g(n)) \Leftrightarrow \limsup_{n \to \infty} \frac{f(n)}{g(n) \log^p(n)} < \infty$, for some power $p \in \mathbb{N}$

- $\| \cdot \|_1$ is the L_1 norm

- For all $v := (v_n)_{n \in [N]}$, such that $v_n \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$, we define $\|v\|_{\infty,1} := \sup_{n \in [N]} \|v_n\|_1$.

Below are notations related to the Markov decision process:

- $\mathcal{X}$ finite state space

- $\mathcal{A}$ finite action space

- $N :=$ episode length

- $n :=$ time step index within an episode, $n \in [N]$

- $\pi := (\pi_n)_{n \in [N]}$ with, for all $n \in [N], x \in \mathcal{X}$, $\pi_n(\cdot|x) \in \Delta_{\mathcal{A}}$ = policy at time step n conditioned on state x

- $\Pi := (\Delta_{\mathcal{A}})^{N \times \mathcal{X}}$ = space of policies (defined using the generic notations from above)

- $p := (p_n)_{n \in [N]}$ = true probability transition kernel with, for all $n \in [N], (x,a) \in \mathcal{X} \times \mathcal{A}$, $p_n(\cdot|x,a) \in \Delta_{\mathcal{X}}$

- $\rho \in \Delta_{\mathcal{X} \times \mathcal{A}}$ = target initial state-action distribution

- For any $\pi \in \Pi$, and $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, $\mu^{\pi,\nu} := (\mu_n^{\pi,\nu})_{n \in [N]}$, where for all (x,a), $\mu_0^{\pi,\nu}(x,a) = \nu(x,a)$, and for all n,

$$\mu_{n+1}^{\pi,\nu}(x,a) := \sum_{x',a'} \mu_n^{\pi,\nu}(x',a') p_{n+1}(x|x',a') \pi_{n+1}(a|x)$$

- For any $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, and for any probability transition p,

$$\mathcal{M}_\nu^p := \left\{ \mu \,\middle|\, \mu_0 = \nu \,,\, \sum_{a' \in \mathcal{A}} \mu_n(x',a') = \sum_{x \in \mathcal{X}, a \in \mathcal{A}} p_n(x'|x,a)\mu_{n-1}(x,a) \,,\, \forall x' \in \mathcal{X}, \forall n \in [N] \right\}$$

- For all $n \in [N]$, for $y_n := (x_n, a_n)$, and for all policy $\pi \in \Pi$, $P_n^\pi(y_{n-1}, y_n) := p_n(x_n|y_{n-1})\pi_n(a_n|x_n)$

- $P_\pi(y_0, y_N) := \sum_{y_1} P_1^\pi(y_0, y_1) \ldots \sum_{y_{N-1}} P_{N-1}^\pi(y_{N-2}, y_{N-1}) P_N^\pi(y_{N-1}, y_N)$

- For all $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, $\nu P_\pi(x,a) = \sum_{x',a'} \nu(x',a') P_\pi(x',a'; x,a)$

- M = number of agents

- α = contraction parameter defined in Ass. 2

Below are notations related to the online settings:

- $T :=$ number of episodes

- $t :=$ episode index, $t \in [T]$

- $\pi_t :=$ policy played at episode t associated to the state-action distribution sequence solution of Eq. (8) at episode t

- $\rho_t \in \Delta_{\mathcal{X} \times \mathcal{A}}$ = initial state-action distribution at episode t

- $\hat{p}_t$ empirical probability transition kernel estimated at episode t following Eq. 6

- For any $\pi \in \Pi$, and $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, $\hat{\mu}_t^{\pi,\nu} := (\hat{\mu}_{t,n}^{\pi,\nu})_{n \in [N]}$, where for all (x,a), $\hat{\mu}_{t,0}^{\pi,\nu}(x,a) = \nu(x,a)$, and for all n,

$$\hat{\mu}_{t,n+1}^{\pi,\nu}(x,a) := \sum_{x',a'} \hat{\mu}_{t,n}^{\pi,\nu}(x',a') \hat{p}_{t,n+1}(x|x',a') \pi_{n+1}(a|x)$$

- $\mathcal{M}_t := \mathcal{M}_{\rho_t}^{\hat{p}_t}$

- $\mu_t := \hat{\mu}_t^{\pi_t, \rho_t}$, the state-action distribution sequence solution of Eq. (8) at episode t

- $\ell_t := \nabla F_t(\mu_t)$

- For all $t \in [T], n \in [N], f_{t,n} : \mathbb{R}^{\mathcal{X} \times \mathcal{A}} \to \mathbb{R}$ any convex and ℓ-Lipschitz function with respect to $\| \cdot \|_1$

- $F_t := \sum_{n=1}^{N} f_{t,n}$

- $\gamma :=$ penalization parameter of the periodic regret defined in Eq. (4)

- $\eta =$ online mirror descent learning rate

- $\eta_\lambda =$ dual learning rate

- $\lambda =$ Lagrange multiplier

- $\tilde{p}_t =$ empirical probability transition kernel estimated using the trajectories of the restarted agent in framework 2

- $\tilde{\rho}_t =$ estimate of the initial state-action distribution at episode t used in framework 2

Appendix B. Related Work

Table 1 provides an overview of existing approaches to online MDPs with adversarial losses in both episodic and infinite-horizon settings, highlighting the distinctions with our periodic framework. Below, we discuss additional RL works that share certain similarities with our approach.

Autonomous RL: Many lines of work in RL address continual learning tasks from a practical point of view, without offering theoretical guarantees; see Sharma et al. (2022b) for an overview. Most of the deep RL approaches to continual learning concentrate on learning to reset. For instance, Eysenbach et al. (2018); Sharma et al. (2022a) propose ideas of state distribution matching similar to ours. These methods involve learning both a "forward policy" to optimize the task and a "backward policy" to drive the agent toward a predefined state distribution. However, they treat task optimization and distribution matching as distinct objectives, and do not assume a fixed episode length in which the agent must jointly minimize cumulative loss and return to the initial distribution by the episode's end, as we do. Furthermore, these works do not have the theoretical guarantees provided by our approach.

As for theoretical works addressing non-episodic settings, they typically focus on optimizing the infinite-horizon average loss Brafman and Tennenholtz (2003); Jaksch et al. (2008); Bartlett and Tewari (2009); Bhavya et al. (2024), which differs from minimizing periodic regret. Moreover, these works assume a fixed loss across episodes, in contrast to our setting, which considers an adversarially varying sequence of losses.

Stochastic control with constraints in law: Some works on the literature of stochastic control investigate constraints on the probability distribution of the terminal state Daudin (2020); Bouchard et al. (2010); Föllmer and Leukert (1999); Pfeiffer (2020); Pfeiffer et al. (2021). However, these studies primarily focus on deriving necessary optimality conditions under known dynamics. As such, they do not address learning problems, nor do they consider online settings with adversarially

changing losses as we do in this paper. Such control problems naturally arise in economics and finance, where an agent seeks to minimize a cost subject to constraints on the terminal distribution (e.g., achieving a specified distribution of terminal wealth Guo et al. (2022)), as well as in energy management applications, as detailed in the main paper Bourdais et al. (2025).

Appendix C. Algorithm schemes for both settings

We outline below the algorithms for computing a sequence of policies that achieve low periodic regret. The procedure is detailed for the two settings considered in the main paper: Framework 1, where ρ_t is observed at the beginning of each episode; and Framework 2, where ρ_t is unobserved and must be estimated by the agent. Steps specific to Framework 1 are highlighted in red, while those specific to Framework 2 are highlighted in blue. All remaining steps are similar for both settings.

Algorithm 2 MDPP-K: Framework 1: known initial distributions

Input: number of episodes T; length of each episode N; number of agents M; initial state-action distribution ρ; initial policy sequence π_1

Initialize: $\rho_1 = \rho$; $N_{1,n}(x,a) = M_{1,n}(x'|x,a) = 0$; $\hat{p}_{1,n}(x'|x,a) = 1/|\mathcal{X}|$ for all $(n,x,a,x') \in [N] \times \mathcal{X} \times \mathcal{A} \times \mathcal{X}$

Each agent $j \in [M]$ starts at $(x_{1,0}^j, a_{1,0}^j) \sim \rho(\cdot)$

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

  for $n = 1, \ldots, N$ **do**

    Environment samples new state $x_{t,n}^j \sim p_n(\cdot|x_{t,n-1}^j, a_{t,n-1}^j)$ for all agents $j \in [M]$

    Each agent $j \in [M]$ takes an action $a_{t,n}^j \sim \pi_{t,n}(\cdot|x_{t,n}^j)$

  end for

  Uniformly sample an agent $j \in [M]$ to ensure independence of the observed trajectories, and observe its state-action trajectory $(x_{t,n}^j, a_{t,n}^j)_{n \in [N]}$

  Update the counts using the entire trajectory

$$N_{t+1,n-1}(x_{t,n-1}^j, a_{t,n-1}^j) \leftarrow N_{t,n-1}(x_{t,n-1}^j, a_{t,n-1}) + 1$$
$$M_{t+1,n-1}(x_{t,n}^j|x_{t,n-1}^j, a_{t,n-1}^j) \leftarrow M_{t,n-1}(x_{t,n}^j|x_{t,n-1}, a_{t,n-1}) + 1$$

  Update the estimate $\hat{p}_{t+1}$ based on the updated counts following Eq. (6)

  Observe the objective function F_t (full-information)

  Recall that for each agent $j \in [M]$, $(x_{t+1,0}^j, a_{t+1,0}^j) = (x_{t,N}^j, a_{t,N}^j)$

  Observe $\rho_{t+1} = \rho_t P_{\pi_t}$ (the initial state-action distribution at episode $t+1$)

  Compute π_{t+1} solving Eq. (8) using ρ_{t+1}, $\hat{p}_{t+1}$ and F_t

end for

Algorithm 3 MDPP-U: Framework 2: unknown initial distributions

---

1: **Input:** number of episodes T; length of each episode N; number of agents M; initial state-action distribution ρ; initial policy sequence π_1; restarted agent, assumed here to be the agent M, but in practice it can be a different agent per episode

2: **Initialize:** $\rho_1 = \rho$; $\tilde{\rho}_1 = \rho$; $N_{1,n}(x,a) = M_{1,n}(x'|x,a) = 0$; $\widetilde{N}_{1,n}(x,a) = \widetilde{M}_{1,n}(x'|x,a) = 0$; $\hat{p}_{1,n}(x'|x,a) = 1/|\mathcal{X}|$; $\tilde{p}_{1,n}(x'|x,a) = 1/|\mathcal{X}|$ for all $(n,x,a,x') \in [N] \times \mathcal{X} \times \mathcal{A} \times \mathcal{X}$

3: Each agent $j \in [M]$ starts at $(x_{1,0}^j, a_{1,0}^j) \sim \rho(\cdot)$

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

5:     **for** $n = 1, \ldots, N$ **do**

6:         Environment samples new state $x_{t,n}^j \sim p_n(\cdot|x_{t,n-1}^j, a_{t,n-1}^j)$ for all agents $j \in [M]$

7:         Each agent $j \in [M]$ takes an action $a_{t,n}^j \sim \pi_{t,n}(\cdot|x_{t,n}^j)$

8:     **end for**

9:     Uniformly sample an agent $j \in [M]$ to ensure independence of the observed trajectories, and observe its state-action trajectory $(x_{t,n}^j, a_{t,n}^j)_{n \in [N]}$

10:     Update the counts using the entire trajectory

$$N_{t+1,n-1}(x_{t,n-1}^j, a_{t,n-1}^j) \leftarrow N_{t,n-1}(x_{t,n-1}^j, a_{t,n-1}^j) + 1$$
$$M_{t+1,n-1}(x_{t,n}^j|x_{t,n-1}^j, a_{t,n-1}^j) \leftarrow M_{t,n-1}(x_{t,n}^j|x_{t,n-1}^j, a_{t,n-1}^j) + 1$$

11:     Update the estimate $\hat{p}_{t+1}$ based on the updated counts following Eq. (6)

12:     Observe the objective function F_t (full-information)

13:     Recall that for each agent $j \in [M-1]$, $(x_{t+1,0}^j, a_{t+1,0}^j) = (x_{t,N}^j, a_{t,N}^j)$

14:     Update the second set of counts using the trajectory of agent M

$$\widetilde{N}_{t+1,n-1}(x_{t,n-1}^M, a_{t,n-1}^M) \leftarrow \widetilde{N}_{t,n-1}(x_{t,n-1}^M, a_{t,n-1}^M) + 1$$
$$\widetilde{M}_{t+1,n-1}(x_{t,n}^M|x_{t,n-1}^M, a_{t,n-1}^M) \leftarrow \widetilde{M}_{t,n-1}(x_{t,n}^M|x_{t,n-1}^M, a_{t,n-1}^M) + 1$$

15:     Update the estimate $\tilde{p}_{t+1}$ based on the second set of updated counts following Eq. (6)

16:     Compute the initial state-action distribution estimate $\tilde{\rho}_{t+1} = \tilde{\rho}_t \widetilde{P}_{\pi_t}$

17:     Reset agent M at $(x_{t+1,0}^M, a_{t+1,0}^M) \sim \tilde{\rho}_{t+1}(\cdot)$

18:     Compute π_{t+1} solving Eq. (8) but with $\tilde{\rho}_{t+1}$ instead of ρ_{t+1}, $\hat{p}_{t+1}$ and F_t

19: **end for**

Appendix D. Auxiliary Lemmas

Lemma 9 (Lem. 17 of Jaksch et al., 2008) *For any $0 < \delta < 1$, with a probability of at least $1 - \delta$,*

$$\|p_n(\cdot|x,a) - \hat{p}_n^t(\cdot|x,a)\|_1 \leq \sqrt{\frac{2|\mathcal{X}|\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)}{\max\left\{1, N_{n-1}^t(x,a)\right\}}}$$

holds simultaneously for all $(t, n, x, a) \in [T] \times [N] \times \mathcal{X} \times \mathcal{A}$.

Lemma 10 *For any strategy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X} \times N}$, for any two probability kernels $p = (p_n)_{n \in [N]}$ and $q = (q_n)_{n \in [N]}$ such that $p_n, q_n : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$, for any two initial state-action distributions $\nu, \nu' \in \Delta_{\mathcal{X} \times \mathcal{A}}$, and for all $n \in [N]$,*

$$\|\mu_n^{\pi,p,\nu} - \mu_n^{\pi,q,\nu'}\|_1 \leq$$
$$\sum_{i=0}^{n-1} \sum_{x,a} \mu_i^{\pi,p,\nu}(x,a)\|p_{i+1}(\cdot|x,a) - q_{i+1}(\cdot|x,a)\|_1 + \|\nu - \nu'\|_1,$$

where $\mu^{\pi,p,\nu}$ is the state-action distribution sequence induced by policy π, in the MDP with probability transition p, with initial state-action distribution ν.

Proof Using that for $n \geq 1$,

$$\mu_n^{\pi,p,\nu}(x,a) = \sum_{x',a'} \mu_{n-1}^{\pi,p,\nu'}(x',a')p_n(x|x',a')\pi_n(a|x),$$

we can show that (see for example Lemma D.1 from Moreno et al. (2024))

$$\|\mu_n^{\pi,p,\nu} - \mu_n^{\pi,q,\nu'}\|_1 \leq \sum_{x,a} \mu_{n-1}^{\pi,p,\nu}(x,a)\|p_n(\cdot|x,a) - q_n(\cdot|x,a)\|_1 + \|\mu_{n-1}^{\pi,p,\nu'} - \mu_{n-1}^{\pi,q,\nu}\|_1.$$

Hence, by induction and using that $\mu_0^{\pi,p,\nu'} = \nu'$ and $\mu_0^{\pi,q,\nu} = \nu$, we conclude the proof. ∎

Lemma 11 *For any policy $\pi \in \Pi$, for any initial state-action distribution $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, for any $\delta \in (0, 1)$, for any episode $t \in [T]$, we have that with probability at least $1 - \delta$,*

$$\|\nu(P_\pi - \widehat{P}_\pi^t)\|_1 \leq \langle \nu \widehat{P}_\pi^t, b_t \rangle = \langle \hat{\mu}_{t,N}^{\pi,\nu}, b_t \rangle,$$

where b_t is the bonus vector defined for each $n \in \{0, 1, \ldots, N-1\}$, $(x, a) \in \mathcal{X} \times \mathcal{A}$, as

$$b_{t,n}(x,a) := \frac{C_\delta}{\sqrt{\max\{1, N_{t,n}(x,a)\}}},$$

with

$$C_\delta = \sqrt{2|\mathcal{X}|\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)}. \tag{9}$$

Proof Rewriting the initial L_1 norm in terms of the state-action distributions we have that

$$
\|\nu(P_\pi - \widehat{P}_\pi^t)\|_1 = \|\mu_N^{\pi,\nu} - \hat{\mu}_{t,N}^{\pi,\nu}\|_1
$$

$$
\underbrace{\leq}_{\text{Lemma 10}} \sum_{i=0}^{N-1} \sum_{x,a} \hat{\mu}_{t,i}^{\pi,\nu}(x,a)\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1
$$

$$
\underbrace{\leq}_{\text{Lemma 9}} \sum_{i=0}^{N-1} \sum_{x,a} \hat{\mu}_{t,i}^{\pi,\nu}(x,a) \frac{C_\delta}{\sqrt{\max\{1, N_{t,i}(x,a)\}}}
$$

$$
= \langle \hat{\mu}_t^{\pi,\nu}, b_t \rangle.
$$

$\blacksquare$

Proposition 12 *For $t \in [T]$, let $(x_{t,n}, a_{t,n})_{n \in [N]}$ represent the trajectory of an agent following policy π_t, with the initial state-action pair sampled independently from previous trajectories according to a distribution $\nu_t \in \Delta_{\mathcal{X} \times \mathcal{A}}$. Define the filtration $\mathcal{F}_t := \sigma\big((x_{s,n}, a_{s,n})_{n \in [N]}, s < t\big)$, capturing the sigma-algebra generated by all observations prior to the start of episode t. Assume that both π_t and ν_t are $\mathcal{F}_t$-measurable. Let $N_{t,n}(x,a) := \sum_{s=1}^{t-1} \mathbb{1}_{\{x_{s,n}=x, a_{s,n}=a\}}$ denote the number of times the state-action pair (x,a) has been visited at time step n up to the start of episode t, and define $\xi_{t,n}(x,a) \leq \frac{C}{\sqrt{\max\{1, N_{t,n}(x,a)\}}}$ for all (t,n,x,a), where C is a fixed constant, and $\xi_{t,n}(x,a)$ is $\mathcal{F}_t$-measurable. We also assume $\xi_{t,n}(x,a) \leq c$, for a fixed constant c that may differ from C. Hence, for all $n \in [N]$, for all $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$
\sum_{t=1}^T \langle \mu_n^{\pi_t, \nu_t}, \xi_{t,n} \rangle := \sum_{t=1}^T \sum_{x,a} \mu_n^{\pi_t, \nu_t}(x,a)\xi_{t,n}(x,a) \leq 3C\sqrt{|\mathcal{X}||\mathcal{A}|T} + c|\mathcal{X}|\sqrt{2T \log\left(\frac{N}{\delta}\right)}.
$$

Proof For a fixed $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have from Lemma 19 of Jaksch et al. (2008) that,

$$
\sum_{t=1}^T \mathbb{1}_{\{x_{t,n}=x, a_{t,n}=a\}}\xi_{t,n}(x,a) \leq C \sum_{t=1}^T \frac{\mathbb{1}_{\{x_{t,n}=x, a_{t,n}=a\}}}{\sqrt{\max\{N_{t,n}(x,a), 1\}}} \leq 3C\sqrt{N_{T,n}(x,a)}.
$$

Summing the inequality above over all possible (x,a), using Jensen's inequality and that $\sum_{x,a} N_{T,n}(x,a) = T$, we obtain that

$$
\sum_{t=1}^T \sum_{x,a} \mathbb{1}_{\{x_{t,n}=x, a_{t,n}=a\}}\xi_{t,n}(x,a) \leq 3C\sqrt{|\mathcal{X}||\mathcal{A}|T}. \tag{10}
$$

Therefore, to conclude the proof we need to analyze the term

$$
\sum_{t=1}^T \sum_{x,a} \big(\mu_n^{\pi_t, \nu_t}(x,a) - \mathbb{1}_{\{x_{t,n}=x, a_{t,n}=a\}}\big)\xi_{t,n}(x,a).
$$

Note that, for all (x,a), $\xi_{t,n}(x,a)$ is $\mathcal{F}_t$-measurable by construction, and, by assumption, so are π_t and ν_t. Consequently, we have $\mathbb{E}[\mu_n^{\pi_t, \nu_t}(x,a)\xi_{t,n}(x,a) \mid \mathcal{F}_t] = \mu_n^{\pi_t, \nu_t}(x,a)\xi_{t,n}(x,a)$. Furthermore,

since the trajectories in each episode are observed independently from agents initialized according to the distribution ν_t, it follows that

$$\mathbb{E}[\mathbb{1}_{\{x_{t,n}=x,a_{t,n}=a\}}\xi_{t,n}(x,a) \mid \mathcal{F}_t] = \xi_{t,n}(x,a)\mathbb{E}[\mathbb{1}_{\{x_{t,n}=x,a_{t,n}=a\}} \mid \mathcal{F}_t] = \xi_{t,n}(x,a)\mu_n^{\pi_t,\nu_t}(x,a).$$

Define the sequence $M_{0,n} = 0$, and for all $t \in [T]$

$$M_{t,n} := \sum_{s=1}^{t}\sum_{x,a}\left(\mu_n^{\pi_s,\nu_s}(x,a) - \mathbb{1}_{\{x_{s,n}=x,a_{s,n}=a\}}\right)\xi_{s,n}(x,a).$$

From the arguments above, $(M_{t,n})_{t\in[T]}$ defines a martingale difference sequence. Recall that $|\xi_{t,n}(x,a)| \leq c$. Hence, from Azuma-Hoeffding inequality we obtain that, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$M_{T,n} \leq c|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}. \tag{11}$$

Joining the bounds on Eq. (10) and (11), we obtain that, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T}\langle\mu_n^{\pi_t,\nu_t},\xi_{t,n}\rangle \leq 3C\sqrt{|\mathcal{X}||\mathcal{A}|T} + c|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)},$$

concluding the proof. ■

Corollary 13 *Let $(b_t)_{t\in[T]}$ and $(\bar{b}_t)_{t\in[T]}$ be the two sequences of bonus vectors defined in Eq. (7). Hence, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T}\langle b_t,\mu^{\pi_t,\rho_t}\rangle = \tilde{O}(N|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}), \quad and \quad \sum_{t=1}^{T}\langle\bar{b}_t,\mu^{\pi_t,\rho_t}\rangle = \tilde{O}(\ell N^2|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T})$$

and with probability at least $1 - 2\delta$,

$$\sum_{t=1}^{T}\langle b_t,\hat{\mu}_t^{\pi_t,\rho_t}\rangle = \tilde{O}(N^2|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}), \quad and \quad \sum_{t=1}^{T}\langle\bar{b}_t,\hat{\mu}_t^{\pi_t,\rho_t}\rangle = \tilde{O}(\ell N^3|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}).$$

Proof From Prop. 12 with $\nu_t = \rho_t$, the policy π_t solution of Eq. (8), the trajectories observed in MDPP-K (Alg. 2) that are independent across episodes, and $\xi_t = b_t$, such that $C = c = C_\delta$ we obtain that, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T}\langle b_t,\mu^{\pi_t,\rho_t}\rangle \leq 3NC_\delta\sqrt{|\mathcal{X}||\mathcal{A}|T} + NC_\delta|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}$$

$$\leq N(3|\mathcal{X}| + 2|\mathcal{X}|^{3/2})\sqrt{2|\mathcal{A}|T\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\log\left(\frac{N}{\delta}\right)}.$$

For the bonus sequence $(\bar{b}_t)_{t\in[T]}$, we apply Prop. 12 in a similar manner, this time with $C = c = \ell NC_\delta$, hence we obtain that

$$\sum_{t=1}^{T}\langle\bar{b}_t,\mu^{\pi_t,\rho_t}\rangle \leq \ell N^2(3|\mathcal{X}| + 2|\mathcal{X}|^{3/2})\sqrt{2|\mathcal{A}|T\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\log\left(\frac{N}{\delta}\right)}.$$

26

As for the terms involving the sum of the product between a bonus vector and $\hat{\mu}_t^{\pi_t,\rho_t}$, we have that

$$\sum_{t=1}^{T}\langle b_t, \hat{\mu}_t^{\pi_t,\rho_t}\rangle = \underbrace{\sum_{t=1}^{T}\langle b_t, \hat{\mu}_t^{\pi_t,\rho_t} - \mu^{\pi_t,\rho_t}\rangle}_{(i)} + \underbrace{\sum_{t=1}^{T}\langle b_t, \mu^{\pi_t,\rho_t}\rangle}_{(ii)},$$

where we have just shown a high-probability upper bound for term (ii).

We now analyze term (i): using Holder's inequality, and that from the definition of the bonus vector, $\|b_{t,n}\|_\infty \leq C_\delta$, we have that

$$(i) \leq \sum_{t=1}^{T}\sum_{n=1}^{N}\|b_{t,n}\|_\infty\|\hat{\mu}_{t,n}^{\pi_t,\rho_t} - \mu_n^{\pi_t,\rho_t}\|_1$$

$$\underbrace{\leq}_{\text{Lemma } 10} C_\delta \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{i=0}^{n-1}\sum_{x,a}\mu_i^{\pi_t,\rho_t}(x,a)\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1.$$

We then apply Prop. 12 again with the trajectories observed in MDPP-K (Alg. 2), policy π_t, $\nu_t = \rho_t$, but now with $\xi_{t,n}(x,a) := \|p_{n+1}(\cdot|x,a) - \hat{p}_{t,n+1}(\cdot|x,a)\|_1 \leq \frac{C_\delta}{\max\{1, \sqrt{N_{t,n}(x,a)}\}}$, where $\xi_{t,n}(x,a) \leq 2$ by definition. Hence, we can take $C = C_\delta$ and $c = 2$ to obtain that, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$(i) \leq C_\delta N^2\left[3C_\delta\sqrt{|\mathcal{X}||\mathcal{A}|T} + 2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right]$$

$$\leq 10N^2|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}.$$

Joining both terms, we get that for any $\delta \in (0,1)$, with probability at least $1 - 2\delta$,

$$\sum_{t=1}^{T}\langle b_t, \hat{\mu}_t^{\pi_t,\rho_t}\rangle \leq 15N^2|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}$$

$$= O\left(N^2|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\right),$$

where we use the definition of C_δ in Eq. (9).

Similarly, for the bonus sequence $\bar{b}_t$, we have $\|\bar{b}_{t,n}\|_\infty \leq \ell NC_\delta$. Therefore, by applying a similar reasoning with adjusted constants, we obtain that

$$\sum_{t=1}^{T}\langle \bar{b}_t, \hat{\mu}_t^{\pi_t,\rho_t}\rangle \leq \ell C_\delta N^3\left[3C_\delta\sqrt{|\mathcal{X}||\mathcal{A}|T} + 2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right] + \sum_{t=1}^{T}\langle \bar{b}_t, \mu^{\pi_t,\rho_t}\rangle$$

$$\leq 10\ell N^3|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)} + \sum_{t=1}^{T}\langle \bar{b}_t, \mu^{\pi_t,\rho_t}\rangle$$

$$\leq 15\ell N^3|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}$$

$$= O\left(\ell N^3|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\right).$$

■

Lemma 14 *[Lemma A.3 from [Moreno et al. (2025b)]) For all $n \in [N]$, $(x, a, x') \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$, and $t \in [T]$, let $\hat{p}_{t,n+1}(x'|x, a)$ be defined as in Eq. (6). Hence,*

$$\|\hat{p}_{t+1,n+1}(\cdot|x, a) - \hat{p}_{t,n+1}(\cdot|x, a)\|_1 \leq \frac{\mathbb{1}_{\{x_{t,n}=x, a_{t,n}=a\}}}{\max\{1, N_{t+1,n}(x, a)\}}.$$

Lemma 15 *[Lemma A.4 from [Moreno et al. (2025b)]) For $(n, x, a) \in [N] \times \mathcal{X} \times \mathcal{A}$, let $(q_t)_{t \in [T]}$ be a sequence of probability transition kernels with $q_t := (q_{t,n})_{n \in [N]}$ such that*

$$\|q_{t+1,n}(\cdot|x, a) - q_{t,n}(\cdot|x, a)\|_1 \leq \frac{c\mathbb{1}_{\{x_{t,n-1}=x, a_{t,n-1}=a\}}}{\max\{1, N_{t+1,n-1}(x, a)\}}$$

for some constant $c > 0$. Then,

$$\sum_{t=1}^{T} \|q_{t+1,n}(\cdot|x, a) - q_{t,n}(\cdot|x, a)\|_1 \leq ec \log(T).$$

Appendix E. Feasibility

E.1. Proof of Lemma 2

Proof We show that for every episode t, there is a state-action distribution sequence $\mu \in \mathcal{M}_t$ that satisfies the constraint

$$\|\mu_N - \rho\|_1 \leq \langle \mu, b_t \rangle + \alpha \|\rho_t - \rho\|_1. \tag{12}$$

Recall that we denote by ρ_t the initial state-action distribution at episode t. From the equivalence between solving the problem over the set of state-action distributions sequences satisfying the Bellman flow and the set of policies, we have that $\mu_N = \rho_t \widehat{P}_\pi^t$ for some policy π.

From Ass. 1, there exists a periodic policy π in the true MDP with stationary distribution ρ, meaning that $\rho P_\pi = \rho$. Thus, this periodic policy π satisfies that

$$
\begin{aligned}
\|\rho_t \widehat{P}_\pi^t - \rho\|_1 &= \|\rho_t \widehat{P}_\pi^t - \rho_t P_\pi + \rho_t P_\pi - \rho\|_1 \\
&\leq \|\rho_t \widehat{P}_\pi^t - \rho_t P_\pi\|_1 + \|\rho_t P_\pi - \rho\|_1 \\
&\leq \langle \hat{\mu}_t^{\pi,\rho_t}, b_t \rangle + \alpha \|\rho_t - \rho\|_1
\end{aligned}
$$

where the last inequality follows from Lemma 11 stating that for any $\delta \in (0, 1)$, with probability $1 - \delta$, $\|\rho_t(\widehat{P}_\pi^t - P_\pi)\|_1 \leq \langle \hat{\mu}_t^{\pi,\rho_t}, b_t \rangle$, and from Ass. 2. Hence, for $\bar{\alpha} = \alpha$, we showed the existence of a state-action distribution sequence in $\mathcal{M}_t$ satisfying the constraints in Eq. (12). ■

E.2. Proof of Lemma 3

Proof Recall that we define $\mu_t := \hat{\mu}_t^{\pi_t, \rho_t}$ as the solution to Eq. (8) at episode t.

Since μ_t is the solution to the optimization problem in Eq. (8), it satisfies the following constraint:

$$\|\mu_{t,N} - \rho\|_1 \leq \langle \mu_t, b_t \rangle + \alpha \|\rho_t - \rho\|_1.$$

Using that $\mu_{t,N} = \rho_t \widehat{P}_{\pi_t}^t$, and that the final state-action distribution we reach when playing policy π_t starting from the initial distribution ρ_t in the true MDP is $\rho_{t+1} = \mu_N^{\pi_t, \rho_t} = \rho_t P_{\pi_t}$, we have that

$$
\begin{aligned}
\|\mu_N^{\pi_t, \rho_t} - \rho\|_1 &= \|\rho_t P_{\pi_t} - \rho\|_1 \\
&\leq \|\rho_t P_{\pi_t} - \rho_t \widehat{P}_{\pi_t}^t\|_1 + \|\rho_t \widehat{P}_{\pi_t}^t - \rho\|_1 \\
&\leq 2\langle \hat{\mu}_t^{\pi_t, \rho_t}, b_t \rangle + \alpha \|\rho_t - \rho\|_1,
\end{aligned}
$$

where the last inequality follows from Lemma 11 stating that for any $\delta \in (0,1)$, with probability $1 - \delta$, $\|\rho_t P_{\pi_t} - \rho_t \widehat{P}_{\pi_t}^t\|_1 \leq \langle \hat{\mu}_t^{\pi, \rho_t}, b_t \rangle$, and that the policy π_t induces a state-action distribution sequence in $\mathcal{M}_t$ satisfying the inequality constraint.

Recall that $\rho_t := \rho_{t-1} P_{\pi_{t-1}} = \mu_N^{\pi_{t-1}, \rho_{t-1}}$. Substituting the inequalities iteratively for all t, until the first episode where $\rho_1 = \rho$, we derive that:

$$
\begin{aligned}
\|\mu_N^{\pi_t, \rho_t} - \rho\|_1 &\leq 2\langle \mu_t, b_t \rangle + \alpha \|\rho_t - \rho\|_1 \\
&= 2\langle \mu_t, b_t \rangle + \alpha \|\mu_N^{\pi_{t-1}, \rho_{t-1}} - \rho\|_1 \\
&\leq 2\langle \mu_t, b_t \rangle + \alpha [2\langle \mu_{t-1}, b_{t-1} \rangle + \alpha \|\rho_{t-1} - \rho\|] \\
&\leq 2\sum_{s=0}^{t-1} \alpha^s \langle \mu_{t-s}, b_{t-s} \rangle + \alpha^t \underbrace{\|\rho_1 - \rho\|_1}_{=0}.
\end{aligned}
$$

Therefore, summing over $t \in [T]$, with probability $1 - T\delta$,

$$\sum_{t=1}^T \|\mu_N^{\pi_t, \rho_t} - \rho\|_1 \leq 2 \sum_{t=1}^T \sum_{s=0}^{t-1} \alpha^s \langle \mu_{t-s}, b_{t-s} \rangle.$$

By rearranging the sum and using Corollary 13 we have that with probability $1 - (2+T)\delta$,

$$
\begin{aligned}
\sum_{t=0}^T \alpha^t \sum_{s=1}^{T-t} \langle \mu_s, b_s \rangle &\leq \sum_{t=0}^T \alpha^t 15 N^2 |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right) \sqrt{\log\left(\frac{N}{\delta}\right)} \\
&\leq \frac{15 N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right) \sqrt{\log\left(\frac{N}{\delta}\right)},
\end{aligned}
$$

where we use that $\sum_{t=0}^T \alpha^t \leq \frac{1}{1-\alpha}$ as $0 \leq \alpha < 1$ by definition.

Therefore, we obtain that for any $\delta \in (0,1)$, with probability at least $1 - (2+T)\delta$,

$$
\begin{aligned}
\sum_{t=1}^T \|\rho_{t+1} - \rho\|_1 &\leq \frac{30 N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right) \sqrt{\log\left(\frac{N}{\delta}\right)}, \\
&= \tilde{O}\left(\frac{N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T}\right).
\end{aligned}
$$

∎

Appendix F. Upper bound on Mirror Descent term

Recall that for any initial state-action distribution $\rho \in \Delta_{\mathcal{X} \times \mathcal{A}}$, and for any probability transition kernel p, we define in Eq. (2) the set of state-action distribution sequences satisfying the MDP dynamics:

$$\mathcal{M}_\rho^p := \Big\{ \mu := (\mu_n)_{n \in [N]} \ \Big| \ \mu_0 = \rho \text{ and } \sum_a \mu_{n+1}(x,a) := \sum_{x',a'} \mu_n(x',a') p_{n+1}(x|x',a') \Big\}.$$

Let $\mathcal{M}_\rho^{p,*}$ be the subset of $\mathcal{M}_\rho^p$ where the corresponding policies (in the sense of Eq. (1)) satisfy $\pi_n(a|x) \neq 0$. We denote by $\psi : \mathbb{R}^{N \times |\mathcal{X}| \times |\mathcal{A}|} \to \mathbb{R}$ the function inducing the Bregman divergence D_ψ in the sense that for all $\mu, \mu' \in \mathbb{R}^{N \times |\mathcal{X}| \times |\mathcal{A}|}$,

$$D_\psi(\mu, \mu') := \psi(\mu) - \psi(\mu') - \langle \nabla\psi(\mu'), \mu - \mu' \rangle. \tag{13}$$

Assumption 3 *For any μ and μ',*

$$D_\psi(\mu, \mu') \geq \frac{1}{2} \|\mu - \mu'\|_{\infty,1}^2. \tag{14}$$

From now on, we assume D_ψ is one of the two following commonly used Bregman divergences satisfying Ass. 3:

- Using the convention that $0 \log(0) = 0$ and assuming $\mu'_n(x,a) > 0$ for all (n,x,a), let D_ψ be the Kullback-Leibler divergence:

$$D_\psi(\mu, \mu') = \mathrm{KL}(\mu, \mu') := \sum_{n=0}^N \sum_{x,a} \mu_n(x,a) \log\left(\frac{\mu_n(x,a)}{\mu'_n(x,a)}\right). \tag{15}$$

  By Pinsker's inequality, the result in Eq. (14) holds. Note that, in this case, Pinsker's inequality remains valid even if μ and μ' are induced by different probability transitions.

- For any two probability transition kernels p, q, for any two initial state-action distributions ρ, ρ', for all $\mu \in \mathcal{M}_\rho^p$, and $\mu' \in \mathcal{M}_{\rho'}^{q,*}$ with respective policies π, π', let

$$D_\psi(\mu, \mu') = \Gamma(\mu, \mu') := \sum_{n=1}^N \sum_{x,a} \mu_n(x,a) \log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right) + \rho(x,a) \log\left(\frac{\rho(x,a)}{\rho'(x,a)}\right). \tag{16}$$

  If $p = q$, Lemma B.1 from Moreno et al. (2024) shows that D_ψ satisfies the inequality from Eq. (14).

Lemma 16 presents an auxiliary result for Online Mirror Descent, where, in each episode t, the iteration is performed over a space of measures $\mathcal{M}_{\rho_t}^{q_t}$ determined by a probability transition kernel q_t and an initial state-action distribution ρ_t, both of which may vary across episodes. Additionally, each iteration is performed over a set of convex constraints $\mathcal{W}_t$ that may also vary across episodes. We compare the sequence of state-action distributions $(\mu_t)_{t \in [T]}$, generated by solving each OMD iteration, with any sequence $\nu_t := \nu^{q_t, \pi, \rho_t}$ that satisfies the constraints in $\mathcal{W}_t$ and is induced by a fixed policy π.

We then apply Lemma 16 for the case when $q_t := \hat{p}_t$ as defined in Eq. (6), $\rho_t := \rho_{t-1} P_{\pi_{t-1}} = \mu_N^{\pi_{t-1}, p, \rho_{t-1}}$, where the initial distribution in the first episode is set as $\rho_1 := \rho$, and $\mathcal{W}_t := \{\mu := (\mu_n)_{n \in [N]} \mid \|\mu_n - \rho\|_1 \leq \langle \mu, b_t \rangle + \bar{\alpha}\|\rho_t - \rho\|_1\}$. Joining this result with the feasibility results from App. E, and some auxiliary Lemmas in App. D, we show in Prop. 6 a final upper bound on the term R_T^{MD} from the periodic regret decomposition presented in Sec. 3.3.

Lemma 16 *Consider a sequence of vectors $(z_t)_{t \in [T]}$ where $z_t \in \mathbb{R}^{N \times |\mathcal{X}| \times |\mathcal{A}|}$ such that $\max_{t \in [T]} \|z_t\|_{1,\infty} \leq \zeta$. Let $(\mathcal{W}_t)_{t \in [T]}$ define a sequence of convex constraint sets for the state-action distributions sequences. Let $(\rho_t)_{t \in [T]}$ define a sequence of state-action distributions, $(q_t)_{t \in [T]}$ be a sequence of probability transitions, and $\mathcal{M}_t := \mathcal{M}_{\rho_t}^{q_t}$. Assume that $\mathcal{M}_t \cap \mathcal{W}_t$ is non-empty for all $t \in [T]$. Let*

$$\mu_{t+1} \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}_{t+1} \cap \mathcal{W}_{t+1}} \left\{ \eta \langle z_t, \mu \rangle + D_\psi(\mu, \mu_t) \right\}$$

for D_ψ as given in either Eq. (15) or Eq. (16), μ_1 initialized such that $\nabla \psi(\mu_1) = 0$, and $\eta > 0$. For any sequence of distributions $(\nu_t)_{t \in [T]}$ with $\nu_t := \nu^{q_t, \pi, \rho_t}$ for a common policy π,

$$\eta \sum_{t=1}^T \langle z_t, \mu_t - \nu_t \rangle \leq \eta^2 \zeta^2 T + N V_T \mathbb{1}_{\{D_\psi = \Gamma\}} - \psi(\mu_1)$$

$$+ \sum_{t=1}^T \langle \nabla \psi(\mu_t), \nu_t - \nu_{t+1} \rangle + \eta \sum_{t=1}^T \langle z_t, \nu_{t+1} - \nu_t \rangle,$$

where $V_T \geq 1 + \max_{(n,x,a)} \sum_{t=1}^{T-1} \|q_{t,n}(\cdot|x,a) - q_{t+1,n}(\cdot|x,a)\|_1$.

Proof Since this is a convex problem and the constraint set is non-empty, using the optimality conditions and the three points Bregman inequality, for all $\nu_{t+1} \in \mathcal{M}_{t+1} \cap \mathcal{W}_{t+1}$,

$$\eta \langle z_t, \mu_{t+1} - \nu_{t+1} \rangle \leq D_\psi(\nu_{t+1}, \mu_t) - D_\psi(\nu_{t+1}, \mu_{t+1}) - D_\psi(\mu_{t+1}, \mu_t).$$

Re-arranging the terms,

$$\eta \sum_{t=1}^T \langle z_t, \mu_t - \nu_t \rangle \leq \underbrace{\eta \sum_{t=1}^T \langle z_t, \mu_t - \mu_{t+1} \rangle - D_\psi(\mu_{t+1}, \mu_t)}_{(1)} + \underbrace{\sum_{t=1}^T D_\psi(\nu_{t+1}, \mu_t) - D_\psi(\nu_{t+1}, \mu_{t+1}))}_{(2)}$$

$$+ \eta \sum_{t=1}^T \langle z_t, \nu_{t+1} - \nu_t \rangle.$$

$$(17)$$

We analyse each term separately.

Term 1 analysis: Applying Young's inequality for some $\sigma > 0$ we have that for each $t \in [T]$,

$$\eta \langle z_t, \mu_t - \mu_{t+1} \rangle - D_\psi(\mu_{t+1}, \mu_t) \leq \frac{\eta^2 \|z_t\|_{1,\infty}^2}{2\sigma} + \frac{\sigma}{2} \|\mu_t - \mu_{t+1}\|_{\infty,1}^2 - D_\psi(\mu_{t+1}, \mu_t).$$

- $D_\psi = \text{KL}$: If D_ψ is the Kullback-Leibler divergence as in Eq. (15), then for $\sigma = 1$, using the inequality from Eq. (14),

$$\frac{\sigma}{2}\|\mu_t - \mu_{t+1}\|_{\infty,1}^2 - \text{KL}(\mu_{t+1}, \mu_t) \leq 0.$$

Hence, $(1) \leq \frac{\eta^2\zeta^2 T}{2}$.

- $D_\psi = \Gamma$: If D_ψ is the divergence defined in Eq. (16), then it does not satisfy Inequality (14), as the probability transition kernel q_t, which induces μ_t, may differ from q_{t+1} inducing μ_{t+1}. However, by the definition of Γ, the probability transition kernel inducing the term in the second entry is irrelevant; only the policy and initial distribution matter. Therefore, we have

$$\Gamma(\mu_{t+1}, \mu_t) = \Gamma(\mu_{t+1}, \mu_{t+1}^{\pi_t,\rho_t}) \geq \frac{1}{2}\|\mu_{t+1} - \mu_{t+1}^{\pi_t,\rho_t}\|_{\infty,1}^2.$$

Thus, for $\sigma = 1/2$ we have that

$$\sum_{t=1}^{T}\frac{1}{4}\|\mu_t - \mu_{t+1}\|_{\infty,1}^2 - \Gamma(\mu_{t+1}, \mu_t) \leq \sum_{t=1}^{T}\frac{1}{4}\|\mu_t - \mu_{t+1}\|_{\infty,1}^2 - \frac{1}{2}\|\mu_{t+1} - \mu_{t+1}^{\pi_t,\rho_t}\|_{\infty,1}^2$$

$$\leq \sum_{t=1}^{T}\frac{1}{2}\|\mu_t - \mu_{t+1}^{\pi_t,\rho_t}\|_{\infty,1}^2$$

$$\leq \sum_{t=1}^{T}\sup_{n\in\{0,\ldots,N\}}\sum_{i=0}^{n-1}\sum_{x,a}\mu_{t,i}(x,a)\|q_{t,i+1}(\cdot|x,a) - q_{t+1,i+1}(\cdot|x,a)\|_1$$

$$\leq NV_T,$$

where the second to last inequality comes from Lemma 10. Hence,

$$(1) \leq \eta^2\zeta^2 T + NV_T.$$

Term 2 **analysis:** In the analysis of classic online mirror descent Shalev-Shwartz (2012), the sum in term (2) is telescopic. However, because in our case the constraint set changes at each episode due to varying probability transitions and changing initial state-action distributions, this term requires further analysis. To address this, we sum and subtract $D_\psi(\nu_t, \mu_t)$:

$$(2) = \sum_{t=1}^{T}D_\psi(\nu_{t+1}, \mu_t) - D_\psi(\nu_{t+1}, \mu_{t+1})$$

$$= \underbrace{\sum_{t=1}^{T}D_\psi(\nu_{t+1}, \mu_t) - D_\psi(\nu_t, \mu_t)}_{(i)} + \underbrace{\sum_{t=1}^{T}D_\psi(\nu_t, \mu_t) - D_\psi(\nu_{t+1}, \mu_{t+1})}_{(ii)}.$$

For (i), using that ψ is the function inducing the Bregman divergence D_ψ in the sense of Eq. (13), we have that

$$(i) = \sum_{t=1}^{T}\psi(\nu_{t+1}) - \psi(\nu_t) - \langle\nabla\psi(\mu_t), \nu_{t+1} - \nu_t\rangle$$

$$= \psi(\nu_{T+1}) - \psi(\nu_1) + \sum_{t=1}^{T}\langle\nabla\psi(\mu_t), \nu_t - \nu_{t+1}\rangle.$$

Term (ii) is a telescopic sum, therefore, as a Bregman divergence is always positive,

$$(ii) = D_\psi(\nu_1, \mu_1) - D_\psi(\nu_{T+1}, \mu_{t+1}) \leq D_\psi(\nu_1, \mu_1).$$

Joining terms (i) and (ii), using that μ_1 is initialized such that $\nabla \psi(\mu_1) = 0$, and that $\psi(\mu) \leq 0$ for all state-action distribution sequence μ for both Bregman divergences considered, we get that

$$(2) \leq \psi(\nu_{T+1}) - \psi(\nu_1) + \sum_{t=1}^{T} \langle \nabla \psi(\mu_t), \nu_t - \nu_{t+1} \rangle + D_\psi(\nu_1, \mu_1) \leq -\psi(\mu_1) + \sum_{t=1}^{T} \langle \nabla \psi(\mu_t), \nu_t - \nu_{t+1} \rangle.$$

Joining all terms: replacing the upper bounds for each term in the inequality of Eq. (17), we obtain that,

$$\eta \sum_{t=1}^{T} \langle z_t, \mu_t - \nu_t \rangle \leq \eta^2 \zeta^2 T + N V_T \mathbb{1}_{\{D_\psi = \Gamma\}} - \psi(\mu_1)$$

$$+ \sum_{t=1}^{T} \langle \nabla \psi(\mu_t), \nu_t - \nu_{t+1} \rangle + \eta \sum_{t=1}^{T} \langle z_t, \nu_{t+1} - \nu_t \rangle.$$

∎

F.1. Proof of Prop. 6

We prove Prop. 6 which bounds the R_T^{MD} term of the periodic regret decomposition.

Proof Let $\nu_t := \hat{\mu}_t^{\pi, \rho_t}$ for all $t \in [T]$, where π is any periodic policy in the true MDP for the initial distribution ρ, i.e., $\rho P_\pi = \rho$. From the feasibility result on Lemma 2, for all $t \in [T]$, we have that with high probability,

$$\nu_t \in \mathcal{W}_t := \{ \mu := (\mu_n)_{n \in [N]} \mid \|\mu_n - \rho\|_1 \leq \langle \mu, b_t \rangle + \bar{\alpha} \|\rho_t - \rho\|_1 \}, \tag{18}$$

and $\mathcal{W}_t$ is a convex set on the sequences of state-action distributions. Hence, the set $\mathcal{M}_t \cap \mathcal{W}_t$ is convex and non-empty.

Since $f_{t,n}$ is ℓ-Lipschitz with respect to the $\|\cdot\|_1$ norm, we have $\|\ell_t\|_{1,\infty} \leq N\ell$. From the definition of the bonus vector $\bar{b}_t$ in Eq. (7), it follows that $\|\bar{b}_t\|_{1,\infty} \leq \ell C_\delta N^2$, where C_δ is given in Eq. (9). By applying Lemma 16 with $z_t := \ell_t - \bar{b}_t$ such that $\zeta \leq 2\ell N^2 C_\delta$, $q_t := \hat{p}_t$ as defined in Eq. (6), $\rho_t := \rho_{t-1} P_{\pi_{t-1}} = \mu_N^{\pi_{t-1}, \rho_{t-1}}$ for all $1 < t \leq T$, $\rho_1 = \rho$, and $\mathcal{W}_t$ as defined in Eq. (18), and noting that μ_t is the solution to the optimization problem in Eq. (8), we obtain:

$$R_T^{\mathrm{MD}} := \sum_{t=1}^{T} \langle \ell_t - \bar{b}_t, \mu_t - \hat{\mu}_t^{\pi, \rho_t} \rangle \leq \eta (2\ell N^2 C_\delta)^2 T + \frac{2}{\eta} N e \log(T) \mathbb{1}_{\{D_\psi = \Gamma\}} - \frac{\psi(\mu_1)}{\eta}$$

$$+ \underbrace{\frac{1}{\eta} \sum_{t=1}^{T} \langle \nabla \psi(\mu_t), \nu_t - \nu_{t+1} \rangle}_{(i)} + \underbrace{\sum_{t=1}^{T} \langle \ell^t - \bar{b}_t, \nu_{t+1} - \nu_t \rangle}_{(ii)}, \tag{19}$$

where we also use that for all (n, x, a), $V_T := 1 + \sum_{t=1}^{T} \|\hat{p}_{t,n}(\cdot|x, a) - \hat{p}_{t+1,n}(\cdot|x, a)\|_1 \leq e \log(T)$ from Lemmas 14 and 15. We now analyze terms (i) and (ii) considering that $\nu_t = \hat{\mu}_t^{\pi, \rho_t}$.

Term (i): We assume that $\|\nabla\psi(\mu_t)\|_{1,\infty} \leq \Psi$. In practice, to ensure this upper bound, one can define a smoothed version of the policy by mixing it with a uniform policy, as proposed in Moreno et al. (2024). This modification results in only minor changes to the regret, affecting the final regret bound by an additional logarithmic term. Therefore, for clarity of presentation, we assume directly that the gradient of ψ is bounded.

From Holder's inequality, we have that, for any $\delta \in (0,1)$, with probability at least $1 - (2+T)\delta$,

$$
\sum_{t=1}^{T}\langle\nabla\psi(\mu_t), \nu_t - \nu_{t+1}\rangle \leq \sum_{t=1}^{T}\|\nabla\psi(\mu_t)\|_{1,\infty}\|\nu_t - \nu_{t+1}\|_{\infty,1}
$$

$$
\leq \Psi\sum_{t=1}^{T}\sup_{n\in\{0,...,N\}}\|\nu_{t,n} - \nu_{t+1,n}\|_{\infty,1}
$$

$$
= \Psi\sum_{t=1}^{T}\sup_{n\in\{0,...,N\}}\|\hat{\mu}_{t,n}^{\pi,\rho_t} - \hat{\mu}_{t+1,n}^{\pi,\rho_{t+1}}\|_1
$$

$$
\underbrace{\leq}_{\text{Lemma } 10} \Psi\sum_{t=1}^{T}\Bigg[\sup_{\{0,...,N\}}\sum_{i=1}^{n}\sum_{x,a}\hat{\mu}_{t,i-1}^{\pi,\rho_t}(x,a)\|\hat{p}_{t,i}(\cdot|x,a) - \hat{p}_{t+1,i}(\cdot|x,a)\|_1 + \|\rho_t - \rho_{t+1}\|_1\Bigg]
$$

$$
\underbrace{\leq}_{\text{Lemmas } 14 \text{ and } 15} \Psi N e\log(T) + \Psi\sum_{t=1}^{T}\|\rho_t - \rho_{t+1}\|_1
$$

$$
\underbrace{\leq}_{\text{Lemma } 3} \Psi\Bigg[Ne\log(T) + \frac{60N^2}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}\Bigg].
$$

$$(20)$$

Term (ii): The analysis of the second term is similar to the first, but it depends on $|\ell_t - \bar{b}_t|_{1,\infty}$ instead of Ψ. Thus,

$$
\sum_{t=1}^{T}\langle\ell_t - \bar{b}_t, \nu_{t+1} - \nu_t\rangle \leq 2\ell N^2 C_\delta\Bigg[Ne\log(T) + \frac{60N^2}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}\Bigg].
$$

Final step on $R_T^{\mathbf{MD}}$ analysis: let

$$
\beta = Ne\log(T)\mathbb{1}_{\{D_\psi=\Gamma\}} - \psi(\mu_1) + \Psi\Bigg[Ne\log(T) + \frac{60N^2}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}\Bigg].
$$

Joining the bounds of terms (i) and (ii) in Eq. (19), we obtain that with probability at least $1 - 2(2 + T)\delta$,

$$
R_T^{\mathbf{MD}} \leq \eta(2\ell N^2 C_\delta)^2 T + \frac{1}{\eta}\beta + 2\ell N^2 C_\delta\Bigg[Ne\log(T) + \frac{60N^2}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}\Bigg].
$$

For $\eta = \sqrt{\frac{\beta}{(2\ell N^2 C_\delta)^2 T}}$, we then obtain that with high probability

$$
R_T^{\mathbf{MD}} \leq \tilde{O}\left(\ell N^3|\mathcal{X}|^{5/4}|\mathcal{A}|^{1/4}\sqrt{\frac{\Psi}{1-\alpha}}T^{3/4}\right).
$$

34

Appendix G. Main result known ρ_t

Before presenting the main periodic regret bound for our algorithm with known ρ_t, we analyze the terms R_T^{MDP} and R_T^{bonus} from the regret decomposition in Sec. 3. These results build upon and generalize existing results in the literature (see, for example, Moreno et al. (2025b)). We also prove below Prop. 5 upper bounding the regret term $R_T^{\text{diff. } \rho_t}$.

G.1. Upper bound on R_T^{MDP}

Proposition 17 *For any $\delta \in (0,1)$, with probability at least $1 - 2\delta$,*

$$R_T^{MDP} \le \ell N^2 \left(3\sqrt{2}|\mathcal{X}|\sqrt{|\mathcal{A}|T \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2|\mathcal{X}|\sqrt{2T \log\left(\frac{N}{\delta}\right)} \right).$$

Proof Using that $f_{t,n}$ is convex and ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $t \in [T]$ and $n \in [N]$, recalling that $(x_{t,n}, a_{t,n})_{n \in [N]}$ denotes the agent trajectory observed at episode t, and that the trajectory observed in episode t is independent of the previous trajectories by the design of MDPP-K (Alg. 2), we obtain that

$$
\begin{aligned}
R_T^{\text{MDP}} &= \sum_{t=1}^T F_t(\mu^{\pi_t, \rho_t}) - F_t(\hat{\mu}_t^{\pi_t, \rho_t}) \\
&\le \sum_{t=1}^T \langle \nabla F_t(\mu^{\pi_t, \rho_t}), \mu^{\pi_t, \rho_t} - \hat{\mu}_t^{\pi_t, \rho_t} \rangle \\
&\le \ell \sum_{t=1}^T \sum_{n=1}^N \|\mu_n^{\pi_t, \rho_t} - \hat{\mu}_{t,n}^{\pi_t, \rho_t}\|_1 \\
&\underbrace{\le}_{\text{Lemma 10}} \ell \sum_{t=1}^T \sum_{n=1}^N \sum_{i=0}^{n-1} \sum_{x,a} \mu_i^{\pi_t, \rho_t}(x,a) \|p_{i+1}(\cdot|x,a) - \hat{p}_{t,i+1}(\cdot|x,a)\|_1.
\end{aligned}
$$

Applying Prop. 12 with $\nu_t = \rho_t$, with π_t the policy used to observe the trajectories in the true MDP, and with $\xi_{t,n}(x,a) = \|p_{n+1}(\cdot|x,a) - \hat{p}_{t,n+1}(\cdot|x,a)\|_1 \le \frac{C_\delta}{\sqrt{N_{t,n}(x,a)}}$, such that $C = C_\delta$, and $c = 2$, and where the inequality happens with high probability and follows from Lemma 9, we have that

$$
\begin{aligned}
R_T^{\text{MDP}} &\le \ell \sum_{n=1}^N \sum_{i=0}^{n-1} \left(3C_\delta \sqrt{|\mathcal{X}||\mathcal{A}|T} + 2|\mathcal{X}|\sqrt{2T \log\left(\frac{N}{\delta}\right)} \right) \\
&\le \ell N^2 \left(3\sqrt{2}|\mathcal{X}|\sqrt{|\mathcal{A}|T \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2|\mathcal{X}|\sqrt{2T \log\left(\frac{N}{\delta}\right)} \right).
\end{aligned}
$$

∎

G.2. Upper bound on R_T^{bonus}

Proposition 18 *For any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$R_T^{bonus} \leq \frac{1}{1-\alpha} \tilde{O}\big(\ell N^3 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\big).$$

Proof Using Holder's inequality and that $\|\ell_{t,n}\|_\infty \leq \ell$ for all $t \in [T]$ and $n \in [N]$, as ℓ is the Lipschitz constant of $f_{t,n}$ with respect to the norm $\|\cdot\|_1$, it follows that

$$\sum_{t=1}^{T} \langle \ell_t, \hat{\mu}_t^{\pi,\rho} - \mu^{\pi,\rho} \rangle \leq \ell \sum_{t=1}^{T} \sum_{n=1}^{N} \|\hat{\mu}_{t,n}^{\pi,\rho} - \mu_n^{\pi,\rho}\|_1$$

$$\underset{\text{Lemma 10}}{\leq} \ell \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{i=0}^{n-1} \sum_{x,a} \hat{\mu}_{t,i}^{\pi,\rho}(x,a) \|\hat{p}_{t,i+1}(\cdot|x,a) - p_{i+1}(\cdot|x,a)\|_1$$

$$\underset{\text{Lemma 9}}{\leq} \ell \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{i=0}^{n-1} \sum_{x,a} \hat{\mu}_{t,i}^{\pi,\rho}(x,a) \frac{C_\delta}{\sqrt{N_{t,i}(x,a)}}$$

$$= \ell \sum_{t=1}^{T} \sum_{n=0}^{N} (N-n) \sum_{x,a} \hat{\mu}_{t,n}^{\pi,\rho}(x,a) \frac{C_\delta}{\sqrt{N_{t,n}(x,a)}}$$

$$= \sum_{t=1}^{T} \langle \bar{b}_t, \hat{\mu}_t^{\pi,\rho} \rangle + \sum_{t=1}^{T} \langle \bar{b}_{t,0}, \rho \rangle,$$

where $\bar{b}_t := (\bar{b}_{t,n})_{n \in [N]}$ is the sequence of bonus vectors defined for all $n \in \{0, \ldots, N\}$ and (x,a) as in Eq. (7) by

$$\bar{b}_{t,n}(x,a) := \ell(N-n) \frac{C_\delta}{\sqrt{N_{t,n}(x,a)}},$$

with $C_\delta := \sqrt{2|\mathcal{X}| \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)}$.

Hence, replacing it in the bonus regret term we have that

$$R_T^{\text{bonus}} = \sum_{t=1}^{T} \langle \bar{b}_t, \mu_t - \hat{\mu}_t^{\pi,\rho} \rangle + \langle \ell_t, \hat{\mu}_t^{\pi,\rho} - \mu^{\pi,\rho} \rangle$$

$$\leq \sum_{t=1}^{T} \langle \bar{b}_t, \mu_t \rangle + \sum_{t=1}^{T} \langle \bar{b}_{t,0}, \rho \rangle$$

$$= \underbrace{\sum_{t=1}^{T} \langle \bar{b}_t, \mu_t \rangle + \langle \bar{b}_{t,0}, \rho_t \rangle}_{(i)} + \underbrace{\sum_{t=1}^{T} \langle \bar{b}_{t,0}, \rho - \rho_t \rangle}_{(ii)}.$$

Note that term (ii) arises because we cannot guarantee that the initial distribution at each episode is equal to ρ. This issue does not occur when using bonuses in episodic MDPs, as done in Moreno et al. (2025b). We analyze this additional term in more detail below.

36

As $\mu_0^t = \rho_t$, we have from Corollary 13 that the first term satisfies for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$(i) = \sum_{t=1}^{T} \langle \bar{b}_t, \mu_t \rangle + \langle \bar{b}_{t,0}, \rho_t \rangle \leq 15\ell N^3 |\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T} \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)}.$$

We now proceed with the second term. Using Holder's inequality, that $\|\bar{b}_{t,0}\|_\infty \leq \ell N C_\delta$, and Lemma 3, we have that with high probability

$$\begin{aligned}
(ii) = \sum_{t=1}^{T} \langle \bar{b}_{t,0}, \rho - \rho_t \rangle &\leq \sum_{t=1}^{T} \|\bar{b}_{t,0}\|_\infty \|\rho - \rho_t\|_1 \\
&\leq \ell N C_\delta \sum_{t=1}^{T} \|\rho - \rho_t\|_1 \\
&\leq \ell N C_\delta \frac{30N^2}{1-\alpha} |\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T} \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)} \\
&= \frac{\ell 30\sqrt{2}N^3}{1-\alpha} |\mathcal{X}|^2\sqrt{|\mathcal{A}|T}\log^{3/2}\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)},
\end{aligned}$$

where for the last inequality we use the definition of C_δ.

Joinining both terms, we obtain that with high probability,

$$\begin{aligned}
R_T^{\text{bonus}} &\leq 15\ell N^3 |\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T}\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)} \\
&\quad + \frac{\ell 30\sqrt{2}N^3}{1-\alpha}|\mathcal{X}|^2\sqrt{|\mathcal{A}|T}\log^{3/2}\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)\sqrt{\log\left(\frac{N}{\delta}\right)} \\
&= \tilde{O}\left(\frac{\ell N^3}{1-\alpha}|\mathcal{X}|^2\sqrt{|\mathcal{A}|T}\right).
\end{aligned}$$

$\blacksquare$

G.3. Upper bound on $R_T^{\text{diff. } \rho_t}$: proof of Prop. 5

Proof Recall that $\|\ell_{t,n}\|_\infty \leq \ell$ as ℓ is the Lipschitz constant of $f_{t,n}$ with respect to the norm $\|\cdot\|_1$ for all $n \in [N]$ and $t \in [T]$. Given the definition of $\bar{b}_t$ in Eq. (7), we have that for all $t \in [T]$ and

$n \in [N]$, $\|\bar{b}_{t,n}\|_\infty \leq NC_\delta \ell$. Hence, from Holder's inequality we have that with high probability

$$
\begin{aligned}
R_T^{\mathrm{diff}, \rho_t} &= \sum_{t=1}^{T} \langle \ell_t - \bar{b}_t, \hat{\mu}_t^{\pi, \rho_t} - \hat{\mu}_t^{\pi, \rho} \rangle \\
&\leq 2\ell N C_\delta \sum_{t=1}^{T} \sum_{n=1}^{N} \|\hat{\mu}_{t,n}^{\pi, \rho_t} - \hat{\mu}_{t,n}^{\pi, \rho}\|_1 \\
&\underset{\text{Lemma } 10}{\leq} 2\ell N C_\delta \sum_{t=1}^{T} N \|\rho_t - \rho\|_1 \\
&\underset{\text{Lemma } 3}{\leq} 2\ell N^2 C_\delta \frac{30 N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right) \sqrt{\log\left(\frac{N}{\delta}\right)} \\
&= \tilde{O}\left(\frac{1}{1-\alpha} \ell N^4 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\right).
\end{aligned}
$$

∎

G.4. Final result: proof of Thm. 4

Proof Following the decomposition of the periodic regret in Sec. 3, and using the convexity of F_t with $\ell_t := \nabla F_t(\mu_t)$, we have that for all periodic policy π,

$$
R_T(\pi) \leq \underbrace{\sum_{t=1}^{T} F_t(\mu^{\pi_t, \rho_t}) - F_t(\hat{\mu}_t^{\pi_t, \rho_t})}_{R_T^{\mathrm{MDP}}} + \underbrace{\sum_{t=1}^{T} \langle \ell_t - \bar{b}_t, \mu_t - \hat{\mu}_t^{\pi, \rho_t} \rangle}_{R_T^{\mathrm{MD}}} + \underbrace{\sum_{t=1}^{T} \langle \ell_t - \bar{b}_t, \hat{\mu}_t^{\pi, \rho_t} - \hat{\mu}_t^{\pi, \rho} \rangle}_{R_T^{\mathrm{diff}.\ \rho_t}}
$$
$$
+ \underbrace{\sum_{t=1}^{T} \langle \bar{b}_t, \mu_t - \hat{\mu}_t^{\pi, \rho} \rangle + \langle \ell_t, \hat{\mu}_t^{\pi, \rho} - \mu^{\pi, \rho} \rangle}_{R_T^{\mathrm{bonus}}} + \underbrace{\gamma \sum_{t=1}^{T} \|\rho_t - \rho\|_1}_{R_T^{\mathrm{reg}}}.
$$

We analyze each term separately:

- From Prop. 17, we have that with high probability

$$
\begin{aligned}
R_T^{\mathrm{MDP}} &\leq \ell N^2 \left(3\sqrt{2}|\mathcal{X}|\sqrt{|\mathcal{A}|T \log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2|\mathcal{X}|\sqrt{2T \log\left(\frac{N}{\delta}\right)}\right) \\
&= \tilde{O}\left(\ell N^2 |\mathcal{X}|\sqrt{|\mathcal{A}|T}\right).
\end{aligned}
$$

- From Prop. 6, we have that with high probability

$$
R_T^{\mathrm{MD}} \leq \tilde{O}\left(\ell N^3 |\mathcal{X}|^{5/4} |\mathcal{A}|^{1/4} \sqrt{\frac{\Psi}{1-\alpha}} T^{3/4}\right).
$$

- From Prop. 5, we obtain that with high probability

$$R_T^{\text{diff. } \rho_t} \leq \tilde{O}\left(\frac{1}{1-\alpha}\ell N^4 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\right).$$

- From Prop. 18 we have that with high probability,

$$R_T^{\text{bonus}} \leq \tilde{O}\left(\frac{1}{1-\alpha}\ell N^3 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\right).$$

- Finally, from Lemma 3,

$$R_T^{\text{reg}} \leq \tilde{O}\left(\frac{\gamma}{1-\alpha} N^2 |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T}\right).$$

By replacing all upper bounds in the decomposition of the periodic regret, we conclude that, with high probability, for all periodic policies π,

$$R_T(\pi) \leq \tilde{O}\left(\ell N^2 |\mathcal{X}|^{5/4} |\mathcal{A}|^{1/4} \sqrt{\frac{\Psi}{1-\alpha}} T^{3/4} + \frac{1}{1-\alpha}\ell N^4 |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T}\right).$$

∎

Appendix H. Practical Solution Using Dynamic Programming

We show in this section how to solve the optimization problem in Eq. (8) of the main paper that we use to build the main algorithm in practice:

$$\mu_{t+1} \in \underset{\mu \in \mathcal{M}_{t+1}}{\arg\min} \quad \left\{\eta\langle \ell_t - \bar{b}_t, \mu\rangle + D_\psi(\mu, \mu_t)\right\}$$
$$\text{s.t.} \quad \|\mu_N - \rho\|_1 \leq \langle \mu, b_t\rangle + \bar{\alpha}\|\rho_t - \rho\|_1.$$

Let $\lambda \geq 0$ be the Lagrange multiplier corresponding to the inequality constraint. The Lagrangian is defined as

$$\mathcal{L}(\mu, \lambda) := \langle \ell_t - \bar{b}_t, \mu\rangle + \frac{1}{\eta}D_\psi(\mu, \mu_t) + \lambda[\|\mu_N - \rho\|_1 - \langle \mu, b_t\rangle - \bar{\alpha}\|\rho_t - \rho\|_1].$$

We also represent the constraint function as

$$G(\mu) := \|\mu_N - \rho\|_1 - \langle \mu, b_t\rangle - \bar{\alpha}\|\rho_t - \rho\|_1.$$

Assume there exists $\nu \in \mathcal{M}_{t+1}$ that strictly satisfies the constraint, *i.e.*, $G(\nu) < 0$ (to ensure that, we can slightly increase the value of α in practice). Since the problem is convex, strong duality holds, and thus

$$\mu_{t+1} \in \underset{\mu \in \mathcal{M}_{t+1}}{\arg\min} \quad \left\{\eta\langle \ell_t - \bar{b}_t, \mu\rangle + D_\psi(\mu, \mu_t)\right\} \quad := \quad \max_{\lambda \geq 0} \min_{\mu \in \mathcal{M}_{t+1}} \mathcal{L}(\mu, \lambda).$$
$$\text{s.t.} \quad G(\mu) \leq 0 \tag{21}$$

Minimizing the Lagrangian for a given λ: For a fixed $\lambda \geq 0$, we provide a solution for $\min_{\mu \in \mathcal{M}_{t+1}} \mathcal{L}(\mu, \lambda)$ when the Bregman divergence is defined as in Eq. (16). We derive a closed-form solution for the policies using dynamic programming.

Since the following procedure applies to all $t \in [T]$ and any probability transitions, we omit the episode index and the fact that the optimization is performed on the set of distribution sequences induced by the estimated MDP. Thus, instead of using $\ell_t, \bar{b}_t, b_t, \mathcal{M}_{t+1}, \mu_t, \pi_t, \hat{\mu}_{t+1}^{\pi, \rho_{t+1}}, \hat{p}_{t+1}, \rho_{t+1}$, we simplify the notation to $\ell, \bar{b}, b, \mathcal{M}, \bar{\mu}, \bar{\pi}, \mu^\pi, \bar{p}, \bar{\rho}$.

Let $\ell^\lambda := \ell - \bar{b} - \lambda b$. Hence,

$$
\begin{aligned}
\min_{\mu \in \mathcal{M}} \mathcal{L}(\mu, \lambda) &= \min_{\mu \in \mathcal{M}} \left\{ \langle \ell^\lambda, \mu \rangle + \frac{1}{\eta} D_\psi(\mu, \bar{\mu}) + \lambda \|\mu_N - \rho\|_1 \right\} \\
&= \min_{\mu \in \mathcal{M}} \left\{ \sum_{n=1}^N \sum_{x,a} \ell_n^\lambda(x,a) \mu_n(x,a) + \frac{1}{\eta} \sum_{n=1}^N \sum_{x,a} \mu_n(x,a) \log \left(\frac{\pi_n(a|x)}{\bar{\pi}_n(x,a)} \right) + \lambda \sum_{x,a} |\mu_N(x,a) - \rho(x,a)| \right\} \\
&= \min_{\mu \in \mathcal{M}} \left\{ \sum_{n=1}^N \sum_{x,a} \mu_n(x,a) \left[\ell_n^\lambda(x,a) + \frac{1}{\eta} \log \left(\frac{\pi_n(a|x)}{\bar{\pi}_n(x,a)} \right) \right] + \lambda \sum_{x,a} |\mu_N(x,a) - \rho(x,a)| \right\} \\
&= \min_{\pi \in \Pi} \left\{ \sum_{n=1}^N \sum_{x,a} \mu_n^\pi(x,a) \left[\ell_n^\lambda(x,a) + \frac{1}{\eta} \log \left(\frac{\pi_n(a|x)}{\bar{\pi}_n(x,a)} \right) \right] + \lambda \sum_{x,a} |\mu_N^\pi(x,a) - \rho(x,a)| \right\} \\
&= \min_{\pi \in \Pi} \left\{ \mathbb{E} \left[\sum_{n=1}^N \ell_n^\lambda(x_n, a_n) + \log \left(\frac{\pi_n(a_n|x_n)}{\bar{\pi}_n(a_n|x_n)} \right) \Big| \bar{\rho}, \bar{p}, \pi \right] + \lambda \sum_{x,a} \left| \mathbb{E} \left[\mathbb{1}_{\{(x_N, a_N)=(x,a)\}} \big| \bar{\rho}, \bar{p}, \pi \right] - \rho(x,a) \right| \right\}.
\end{aligned}
$$

$$(22)$$

We define the state-action value function sequence $(Q_i)_{i \in [N]}$ where, for $i = N$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$
\begin{aligned}
Q_N^\lambda(x,a) &= \ell_N^\lambda(x,a) + \lambda \sum_{\bar{x}, \bar{a}} \left| \mathbb{E} \left[\mathbb{1}_{\{(x_N, a_N)=(\bar{x}, \bar{a})\}} \big| (x_N, a_N) = (x,a) \right] - \rho(\bar{x}, \bar{a}) \right| \\
&= \ell_N^\lambda(x,a) + \lambda \sum_{\bar{x}, \bar{a}} \left| \mathbb{1}_{\{(x,a)=(\bar{x}, \bar{a})\}} - \rho(\bar{x}, \bar{a}) \right| \\
&= \ell_N^\lambda(x,a) + \lambda 2(1 - \rho(x,a)),
\end{aligned}
$$

and for each $i \in \{1, \ldots, N-1\}$, where we denote $\pi_{i+1:N} := (\pi_j)_{j \in \{i+1, \ldots, N\}}$,

$$
\begin{aligned}
Q_i^\lambda(x,a) := \min_{\pi_{i+1:N}} \Bigg\{ &\ell_i^\lambda(x,a) + \mathbb{E} \left[\sum_{n=i+1}^N \ell_n^\lambda(x_n, a_n) + \log \left(\frac{\pi_n(a_n|x_n)}{\bar{\pi}_n(a_n|x_n)} \right) \Big| (x_i, a_i) = (x,a), \bar{p}, \pi \right] \\
&+ \lambda \sum_{\bar{x}, \bar{a}} \left| \mathbb{E} \left[\mathbb{1}_{\{(x_N, a_N)=(\bar{x}, \bar{a})\}} \big| (x_i, a_i) = (\bar{x}, \bar{a}), \bar{p}, \pi \right] - \rho(\bar{x}, \bar{a}) \right| \Bigg\}.
\end{aligned}
$$

Note that this state-action value function satisfies

$$
\begin{aligned}
Q_i^\lambda(x,a) = \ell_i^\lambda(x,a) + \min_{\pi_{i+1:N}} \Bigg\{ &\sum_{x'} \bar{p}_{i+1}(x'|x,a) \sum_{a'} \pi_{i+1}(a'|x') \Bigg[\ell_{i+1}^\lambda(x',a') + \log\left(\frac{\pi_{i+1}(a'|x')}{\bar{\pi}_{i+1}(a'|x')} \right) \\
&+ \mathbb{E}\Bigg[\sum_{n=i+2}^{N} \ell_n^\lambda(x_n,a_n) + \log\left(\frac{\pi_n(a_n|x_n)}{\bar{\pi}_n(a_n|x_n)} \right) \Big| (x_{i+1},a_{i+1}) = (x',a'), \bar{p}, \pi \Bigg] \\
&+ \lambda \sum_{\bar{x},\bar{a}} \Bigg| \mathbb{E}\big[\mathbb{1}_{\{(x_N,a_N)=(\bar{x},\bar{a})\}} \big| (x_{i+1},a_{i+1}) = (x',a'), \bar{p}, \pi \big] - \rho(\bar{x},\bar{a}) \Bigg| \Bigg] \Bigg\} \\
= \ell_i^\lambda(x,a) + \min_{\pi_{i+1}} \Bigg\{ &\sum_{x'} \bar{p}_{i+1}(x'|x,a) \sum_{a'} \pi_{i+1}(a'|x') \Bigg[\log\left(\frac{\pi_{i+1}(a'|x')}{\bar{\pi}_{i+1}(a'|x')} \right) + Q_{i+1}^\lambda(x',a') \Bigg] \Bigg\}.
\end{aligned}
\tag{23}
$$

Observe that computing $\mathbb{E}_{(x,a)\sim\bar{\rho}}[Q_0^\lambda(x,a)]$ is equivalent to solving the Lagrangian minimization for a fixed λ given the decomposition in Eq. (22). Hence, our goal is to find the sequence of policies that is used to compute $\mathbb{E}_{(x,a)\sim\bar{\rho}}[Q_0(x,a)]$. Using the relation satisfied by the state-action value function shown in Eq. (23), we can compute each element in the sequence of optimal policies, that we denote by $(\pi_n^\lambda)_{n\in[N]}$, separately backwards in time by taking for all $x \in \mathcal{X}$

$$
\pi_{i+1}^\lambda(\cdot|x) = \operatorname*{arg\,min}_{\pi(\cdot|x)\in\Delta_\mathcal{A}} \big\{ \langle \pi(\cdot|x), Q_{i+1}^\lambda(x,\cdot) \rangle + \frac{1}{\eta} \mathrm{KL}\big(\pi(\cdot|x), \bar{\pi}_{i+1}(\cdot|x) \big) \big\},
$$

which is well known to have a closed-form solution (see for example Moreno et al. (2024)) with for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$
\pi_{i+1}^\lambda(a|x) = \frac{\bar{\pi}_{i+1}(a|x) \exp(-\eta Q_{i+1}^\lambda(x,a))}{\sum_{a'} \bar{\pi}_{i+1}(a'|x) \exp(-\eta Q_{i+1}^\lambda(x,a'))}.
\tag{24}
$$

Finally, using the recursive relation from Eq. (1), we can recover the state-action distribution sequence μ^λ corresponding to π^λ, which is the solution to the Lagrangian minimization $\min_{\mu\in\mathcal{M}} \mathcal{L}(\mu,\lambda)$ for a fixed Lagrange multiplier λ.

Maximizing the Lagrangian: For a given Lagrange multiplier $\lambda > 0$, we previously described how to compute $\mu^\lambda \in \arg\min_{\mu\in\mathcal{M}} \mathcal{L}(\mu,\lambda)$. To find the pair $(\mu^{\lambda_*}, \lambda_*) \in \arg\min_{\lambda\geq 0} \min_{\mu\in\mathcal{M}} \mathcal{L}(\mu^\lambda, \lambda)$, we adopt the min-max iterative procedure outlined in Algorithm 4. We begin by initializing the multiplier $\lambda_1 \geq 0$ and proceed iteratively. At each iteration s, we compute the exact state-action distribution sequence μ^{λ_s} associated with the current multiplier λ_s, following the method described above. We then check whether the inequality constraint is satisfied up to a small constant ε, *i.e.*, $G(\mu^{\lambda_s}) \leq \varepsilon$. If it is, we return μ^{λ_s} as the solution to the original optimization problem. Otherwise, we update the multiplier by performing a gradient ascent step with a tunable learning rate η_λ. We omit the analysis for Algorithm 4 as it is not the main focus of this work and it follows from the classic gradient ascent analysis (see Bach (2024)).

Computing $\bar{\alpha}$: In practice, the algorithm only requires a value $\bar{\alpha}$ that ensures feasibility. From Lemma 3, we know that Eq. (8) is feasible for any $\bar{\alpha} \geq \alpha$. Therefore, we can construct a grid over the interval $(0,1)$ and search for the smallest value of $\bar{\alpha}$ in this grid for which the problem remains feasible. To test feasibility, the L_1 constraint in Eq. (8) can be reformulated as a set of linear inequalities, allowing the use of a solver to efficiently verify feasibility.

Algorithm 4 Lagrangian min-max approach

Input: initial multiplier λ_1, learning rate η_λ, $\varepsilon > 0$
while $G(\mu^{\lambda_s}) > \varepsilon$ **do**
    Update Lagrangian multiplier: $\lambda_{s+1} = \lambda_s + \eta_\lambda G(\mu^{\lambda_s})$
    Compute the associated state-action distribution $\mu^{\lambda_{s+1}}$
end while

Appendix I. Main results unknown ρ_t

In this section, we present the main results for the second framework, where ρ_t is not observed by the learner. The algorithmic scheme is detailed in Algorithm 3 in App. C, with the key differences with respect to the method for the first framework highlighted in blue.

To establish the regret bounds for this framework, we also introduce Ass. 4 below.

Assumption 4 *There exists some $0 \leq \alpha < 1$ such that for all policies $\pi \in \Pi$ and distributions $\nu, \nu' \in \Delta_{\mathcal{X} \times \mathcal{A}}$,*

$$\|\nu P_\pi - \nu' P_\pi\|_1 \leq \alpha \|\nu - \nu'\|_1.$$

I.1. Estimating ρ_t

We build an estimate of the initial state-action distribution at episode t by designating one of the M agents as an agent that can be restarted at the beginning of each episode. We refer to this agent as the *special* agent. Importantly, only this single agent is restarted, unlike in standard episodic RL, where all agents would be reset, making this a weaker assumption. For further discussion, see Sec. 4 in the main paper. Let $\tilde{\rho}_t$ denote the estimate of ρ_t for all $t \in [T]$.

Let $\tilde{\ell}_{t-1} := \nabla F_{t-1}(\hat{\mu}_{t-1}^{\pi_{t-1}, \tilde{\rho}_{t-1}})$, and let $\widetilde{\mathcal{M}}_t := \mathcal{M}_{\tilde{\rho}_t}^{\hat{p}_t}$, the set of state-action distributions sequences that start at the initial distribution $\tilde{\rho}_t$, and that satisfies the dynamics of the MDP induced by the estimated probability kernel $\hat{p}_t$. Algorithm 3 computes the policy π_t by running each online mirror descent iteration over the distributions initialized in $\tilde{\rho}_t$. Hence, at episode t, for some $\eta > 0$ to be tuned later, the learner computes

$$
\begin{aligned}
\mu_t \in \underset{\mu \in \widetilde{\mathcal{M}}_t}{\arg\min} \left\{ \eta \langle \tilde{\ell}_{t-1} - \bar{b}_{t-1}, \mu \rangle + D_\psi(\mu, \mu_{t-1}) \right\} \\
\text{s.t.} \quad \|\mu_N - \rho\| \leq \langle \mu, b_{t-1} \rangle + \bar{\alpha} \|\tilde{\rho}_{t-1} - \rho\|,
\end{aligned}
\tag{25}
$$

and takes π_t as the policy associated to the state-action distribution μ_t as discussed in Sec. 2.

We denote the trajectory of the restarted agent under policy π_t by $(\tilde{x}_{t,n}, \tilde{a}_{t,n})_{n \in [N]}$. At the beginning of episode t, the restarted agent is restarted at an initial state-action pair sampled from $\tilde{\rho}_t$, *i.e.*, $(\tilde{x}_{t,0}, \tilde{a}_{t,0}) \sim \tilde{\rho}_t$. For each tuple (n, x, a, x'), we define the corresponding counts as follows

$$\widetilde{N}_{t,n}(x, a) = \sum_{s=1}^{t-1} \mathbb{1}_{\{\tilde{x}_{s,n} = x, \tilde{a}_{s,n} = a\}}, \qquad \widetilde{M}_{t,n}(x'|x, a) = \sum_{s=1}^{t-1} \mathbb{1}_{\{\tilde{x}_{s,n+1} = x', \tilde{x}_{s,n} = x, \tilde{a}_{s,n} = a\}}.$$

We also define another estimate of the transition probability kernel at the end of episode $t - 1$ as

$$\tilde{p}_{t,n+1}(x'|x, a) := \frac{\widetilde{M}_{t,n}(x'|x, a)}{\max\{1, \widetilde{N}_{t,n}(x, a)\}}.$$

We define the stochastic matrix corresponding to $\tilde{p}_t := (\tilde{p}_{t,n})_{n \in [N]}$ under a policy π, as in Eq. (3), and denote it by $\widetilde{P}_\pi^t$. We then set $\tilde{\rho}_1 := \rho$, and for each episode $1 < t \leq T$, we recursively define $\tilde{\rho}_t := \tilde{\rho}_{t-1} \widetilde{P}_{\pi_t}^t$.

We state and prove below Prop. 7, giving a result on the accuracy of the estimated initial state-action distribution $\tilde{\rho}_t$ computed and used in Algorithm 3.

Proposition *Assuming Ass. 4 holds for $0 \leq \alpha < 1$, we obtain that for any $\delta \in (0,1)$, with probability at least $1 - 2\delta$,*

$$\sum_{t=1}^T \|\rho_t - \tilde{\rho}_t\|_1 \leq \frac{1}{1-\alpha}\left[3N|\mathcal{X}|\sqrt{2|\mathcal{A}|T\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2N|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right].$$

Proof

For $i < j$, for all $\nu \in \Delta_{\mathcal{X} \times \mathcal{A}}$, let $\nu P_{\pi_{i:j}} := \nu P_{\pi_i} P_{\pi_{i+1}} \dots P_{\pi_j}$, and $\nu \widetilde{P}_{\pi_{i:j}}^{i:j} := \nu \widetilde{P}_{\pi_i}^i \widetilde{P}_{\pi_{i+1}}^{i+1} \dots \widetilde{P}_{\pi_j}^j$. Hence, note that $\tilde{\rho}_t := \rho \widetilde{P}_{\pi_{1:t-1}}^{1:t-1}$, and $\rho_t = \rho P_{\pi_{1:t-1}}^{1:t-1}$. Thus, we have that for each $t \in [T]$,

$$
\begin{aligned}
\rho_t - \tilde{\rho}_t &= \rho P_{\pi_{1:t-1}} - \rho \widetilde{P}_{\pi_{1:t-1}}^{1:t-1} \\
&= \rho(P_{\pi_1} - \widetilde{P}_{\pi_1}^1)P_{\pi_{2:t-1}} + \rho \widetilde{P}_{\pi_1}^1(P_{\pi_2} - \widetilde{P}_{\pi_2}^2)P_{\pi_{3:t-1}} + \dots + \rho \widetilde{P}_{\pi_{1:t-2}}^{1:t-2}(P_{\pi_{t-1}} - \widetilde{P}_{\pi_{t-1}}^{t-1}) \\
&= \sum_{i=1}^{t-1} \rho \widetilde{P}_{\pi_{1:i-1}}^{1:i-1}(P_{\pi_i} - \widetilde{P}_{\pi_i}^i)P_{\pi_{i+1:t-1}},
\end{aligned}
$$

where $\widetilde{P}_{\pi_{1:0}}^{1:0} = 1$ and $P_{\pi_{t:t-1}} = 1$. Thus, by taking the L_1 norm and summing over $t \in [T]$ we get that

$$
\begin{aligned}
\sum_{t=1}^T \|\rho_t - \tilde{\rho}_t\|_1 &\leq \sum_{t=1}^T \sum_{i=1}^{t-1} \|\rho \widetilde{P}_{\pi_{1:i-1}}^{1:i-1}(P_{\pi_i} - \widetilde{P}_{\pi_i}^i)P_{\pi_{i+1:t-1}}\|_1 \\
&\leq \sum_{t=1}^T \sum_{i=1}^{t-1} \alpha^{t-i-1}\|\rho \widetilde{P}_{\pi_{1:i-1}}^{1:i-1}(P_{\pi_i} - \widetilde{P}_{\pi_i}^i)\|_1 \qquad (26) \\
&= \sum_{t=1}^T \sum_{i=1}^{t-1} \alpha^{t-i-1}\|\tilde{\rho}_i(P_{\pi_i} - \widetilde{P}_{\pi_i}^i)\|_1
\end{aligned}
$$

where for the before to last inequality we use Ass. 4, and for the last equality we use that $\rho \widetilde{P}_{\pi_{1:i-1}}^{1:i-1} = \tilde{\rho}_i$.

We now express the final L_1 norm using the state-action distribution notation. Let $\mu^{\pi,\nu}$ denote the sequence of state-action distributions induced by executing policy π in an episode starting from initial distribution ν under the true MDP, as defined in Eq. (1). Similarly, let $\tilde{\mu}_t^{\pi,\nu}$ represent the corresponding distribution in the MDP defined by the estimated transition model $\tilde{p}_t$. In particular, we

have that $\tilde{\rho}_i P_{\pi_i} = \mu_N^{\pi_i,\tilde{\rho}_i}$ and $\tilde{\rho}_i \widetilde{P}_{\pi_i}^i = \tilde{\mu}_{i,N}^{\pi_i,\tilde{\rho}_i}$. Therefore, from the inequality above we have that

$$
\sum_{t=1}^{T} \|\rho_t - \tilde{\rho}_t\|_1 \leq \sum_{t=1}^{T} \sum_{i=1}^{t-1} \alpha^{t-i-1} \|\mu_N^{\pi_i,\tilde{\rho}_i} - \tilde{\mu}_{i,N}^{\pi_i,\tilde{\rho}_i}\|_1
$$

$$
\underbrace{\leq}_{\text{Lemma 10}} \sum_{t=1}^{T} \sum_{i=1}^{t-1} \alpha^{t-i-1} \sum_{j=0}^{N-1} \sum_{x,a} \mu_j^{\pi_i,\tilde{\rho}_i}(x,a)\|p_{j+1}(\cdot|x,a) - \tilde{p}_{i,j+1}(\cdot|x,a)\|_1 \quad (27)
$$

$$
= \sum_{t=1}^{T-1} \alpha^{t-1} \sum_{i=1}^{T-t} \sum_{j=0}^{N-1} \sum_{x,a} \mu_j^{\pi_i,\tilde{\rho}_i}(x,a)\|p_{j+1}(\cdot|x,a) - \tilde{p}_{i,j+1}(\cdot|x,a)\|_1.
$$

We apply Prop. 12 to the trajectory of the restarted agent $(\tilde{x}_{t,n}, \tilde{a}_{t,n})_{n\in[N]}$ when following the policy π_t (solution to Eq. (25)), with initial distribution $\nu_t = \tilde{\rho}_t$, and where $\xi_{t,n}(x,a) = \|p_{n+1}(\cdot \mid x,a) - \tilde{p}_{t,n}(\cdot \mid x,a)\|_1 \leq C_\delta/\sqrt{\max\{1, \widetilde{N}_{t,n}(x,a)\}}$. This inequality follows from Lemma 9, allowing us to set $C = C_\delta$ and $c = 2$. Since the restarted agent is restarted at the beginning of each episode from $\tilde{\rho}_t$, the resulting trajectories are independent across episodes. Therefore, Prop. 12 yields that

$$
\sum_{i=1}^{T-t} \sum_{j=0}^{N-1} \sum_{x,a} \mu_j^{\pi_i,\tilde{\rho}_i}(x,a)\|p_{j+1}(\cdot|x,a) - \tilde{p}_{i,j+1}(\cdot|x,a)\|_1
$$

$$
\leq 3N|\mathcal{X}|\sqrt{2|\mathcal{A}|(T-t)\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2N|\mathcal{X}|\sqrt{2(T-t)\log\left(\frac{N}{\delta}\right)}.
$$

Replacing it in the decomposition on Eq. (27), we have that

$$
\sum_{t=1}^{T} \|\rho_t - \tilde{\rho}_t\|_1 \leq \sum_{t=1}^{T-1} \alpha^{t-1}\left[3N|\mathcal{X}|\sqrt{2|\mathcal{A}|(T-t)\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2N|\mathcal{X}|\sqrt{2(T-t)\log\left(\frac{N}{\delta}\right)}\right]
$$

$$
\leq \frac{1}{1-\alpha}\left[3N|\mathcal{X}|\sqrt{2|\mathcal{A}|T\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2N|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right],
$$

where for the last inequality we use that $0 \leq \alpha < 1$. ∎

I.2. Auxiliary results

We now present some auxiliary results that are used in establishing the main periodic regret bound for Algorithm 3 in Thm. 8. These are mainly adaptations of the results from the previous sections to take into account the use of the estimated initial state-action distribution $\tilde{\rho}_t$ rather than the actual distribution ρ_t when calculating the policy at episode t.

New feasibility result: We begin by presenting a new feasibility result. In Lemma 2, we showed that the problem in Eq. (8) is feasible when the sequence of distributions starts from the true state-action distribution ρ_t. However, since the learner now uses the estimated distribution $\tilde{\rho}_t$ to compute the policy, a corresponding feasibility result is required.

Lemma 19 *There exists $0 \le \bar{\alpha} < 1$ such that for every episode $t \in [T]$, the optimization problem in Eq. (25) is feasible with high-probability.*

Proof Let π be any periodic policy. Hence,

$$\|\tilde{\rho}_t \hat{P}_\pi^t - \rho\|_1 \le \|\tilde{\rho}_t(\hat{P}_\pi^t - P_\pi)\|_1 + \|(\tilde{\rho}_t - \rho)P_\pi\|_1$$
$$\le \langle \hat{\mu}_t^{\pi,\tilde{\rho}_t}, b_t \rangle + \alpha\|\tilde{\rho}_t - \rho\|_1,$$

where we use the high-probability result from Lemma 11 and Ass. 2. ∎

Corollary 20 *Let $(b_t)_{t\in[T]}$ and $(\bar{b}_t)_{t\in[T]}$ be the two sequences of bonus vectors defined in Eq. (7). Recall that $\hat{\mu}_t^{\pi,\nu}$ is the state-action distribution sequence induced in the estimated MDP $\hat{p}_t$, with policy π, with initial distribution ν. Thus, for any $\delta \in (0,1)$, with probability at least $1 - 4\delta$,*

$$\sum_{t=1}^{T} \langle b_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} \rangle \le \tilde{O}\left(\frac{N^2}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T} \right) \quad and \quad \sum_{t=1}^{T} \langle \bar{b}_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} \rangle \le \tilde{O}\left(\frac{\ell N^3}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T} \right).$$

Proof We start by decomposing the product as

$$\sum_{t=1}^{T} \langle b_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} \rangle = \underbrace{\sum_{t=1}^{T} \langle b_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} - \hat{\mu}_t^{\pi_t,\rho_t} \rangle}_{(i)} + \underbrace{\sum_{t=1}^{T} \langle b_t, \hat{\mu}_t^{\pi_t,\rho_t} \rangle}_{(ii)}.$$

From Corollary 13, we have that for any $\delta \in (0,1)$, with probability $1 - 2\delta$, $(ii) = \tilde{O}(N^2|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T})$. As for the first term, using Holder's inequality and that for each $n \in [N]$, $\|b_{t,n}\|_\infty \le C_\delta$, we get that

$$(i) \le C_\delta \sum_{t=1}^{T}\sum_{n=1}^{N} \|\hat{\mu}_{t,n}^{\pi_t,\tilde{\rho}_t} - \hat{\mu}_{t,n}^{\pi_t,\rho_t}\|_1$$

$$\underbrace{\le}_{\text{Lemma } 10} C_\delta N \sum_{t=1}^{T} \|\tilde{\rho}_t - \rho_t\|_1$$

$$\underbrace{\le}_{\text{Prop. } 7} \tilde{O}\left(\frac{N^2}{1-\alpha}|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}|T} \right).$$

By summing both terms, we conclude the proof for the bonus vector sequence $(b_t)_{t\in[T]}$.

The result for the bonus vector sequence $(\bar{b}_t)_{t\in[T]}$ follows from a similar decomposition, the application of Corollary 13, and the observation that for each $n \in [N]$, $\|\bar{b}_{t,n}\|_\infty \le \ell N C_\delta$. ∎

Distance between $\tilde{\rho}_t$ and ρ: We now present a result analogous to Lemma 3, which shows that the estimated initial state-action distribution remains, on average, close to the target distribution ρ. This result is crucial for establishing a low periodic regret bound.

Lemma 21 *With high probability we have that*

$$\sum_{t=1}^{T} \|\tilde{\rho}_{t+1} - \rho\|_1 \leq \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right).$$

Proof We denote by $\tilde{\mu}_t^{\pi,\nu}$ the sequence of state-action distributions induced by policy π in the estimated MDP starting from ν, according to the recursive relation in Eq. (1), with initial state-action distribution ν.

For short, we let $\tilde{\mu}_t := \tilde{\mu}_t^{\pi_t, \tilde{\rho}_t}$, and $\mu_t := \hat{\mu}_t^{\pi_t, \tilde{\rho}_t}$. Additionally, we define a bonus vector based on the trajectory of the restarted agent; that is, for all (n, x, a), we set:

$$\tilde{b}_{t,n}(x, a) := \frac{C_\delta}{\sqrt{\max\{\tilde{N}_{t,n}(x, a), 1\}}}.$$

Note that by definition, $\tilde{\rho}_{t+1} = \tilde{\rho}_t \tilde{P}_{\pi_t}^t$, hence

$$\begin{aligned}
\|\tilde{\rho}_{t+1} - \rho\|_1 &= \|\tilde{\rho}_t \tilde{P}_{\pi_t}^t - \rho\|_1 \\
&\leq \|\tilde{\rho}_t (\tilde{P}_{\pi_t}^t - P_{\pi_t})\|_1 + \|\tilde{\rho}_t (\hat{P}_{\pi_t}^t - P_{\pi_t})\|_1 + \|\tilde{\rho}_t \hat{P}_{\pi_t}^t - \rho\|_1 \\
&\leq \underbrace{\langle \tilde{\mu}_t, \tilde{b}_t \rangle + 2\langle \mu_t, b_t \rangle}_{:=\beta_t} + \alpha \|\tilde{\rho}_t - \rho\|_1 \\
&\leq \beta_t + \alpha \big[\beta_{t-1} + \alpha \|\tilde{\rho}_{t-1} - \rho\|_1 \big] \\
&\leq \sum_{s=1}^{t} \alpha^{t-s} \beta_s,
\end{aligned}$$

where for the second inequality we use that π_t is a solution of the OMD iteration at episode t in Eq. (25), therefore satisfying the constraint $\|\tilde{\rho}_t \hat{P}_{\pi_t}^t - \rho\|_1 \leq \langle \mu_t, b_t \rangle + \alpha \|\tilde{\rho}_t - \rho\|_1$, and we also use Lemma 11 for both the stochastic matrices $\hat{P}_{\pi_t}^t$ and $\tilde{P}_{\pi_t}^t$.

For any $s \in [T]$,

$$\sum_{t=1}^{s} \beta_s = \underbrace{\sum_{t=1}^{s} \langle \tilde{\mu}_t, \tilde{b}_t \rangle}_{(i)} + \underbrace{\sum_{t=1}^{s} 2\langle \mu_t, b_t \rangle}_{(ii)}.$$

We analyze each term individually.

Term (ii) can be upper bounded via a direct application of Corollary 20, which shows that

$$(ii) = \sum_{t=1}^{T} \langle \mu_t, b_t \rangle \leq \tilde{O}\left(\frac{N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|s} \right).$$

For term (i), we apply Prop. 12 to the trajectory of the restarted agent under policy π_t, *i.e.*, $(\tilde{x}_{t,n}, \tilde{a}_{t,n})_{n \in [N]}$, using the sequence of initial state-action distributions $(\tilde{\rho}_t)_{t \in [T]}$ from which the restarted agent is reset at the beginning of each episode. We set $\xi_{t,n}(x, a) = \tilde{b}_{t,n}(x, a)$ for all

(n, x, a), and therefore take $C = c = C_\delta$. As a result, for any $\delta \in (0, 1)$, we obtain that with probability at least $1 - \delta$,

$$(i) = \sum_{t=1}^{s} \langle \tilde{\mu}_t, \tilde{b}_t \rangle \leq 3NC_\delta \sqrt{|\mathcal{X}||\mathcal{A}|s} + NC_\delta |\mathcal{X}| \sqrt{2s \log\left(\frac{N}{\delta}\right)}$$

$$= 3N|\mathcal{X}| \sqrt{|\mathcal{A}|s \log\left(\frac{|\mathcal{X}||\mathcal{A}|Ns}{\delta}\right)} + N|\mathcal{X}|^{3/2} \sqrt{2s \log\left(\frac{|\mathcal{X}||\mathcal{A}|Ns}{\delta} \log\left(\frac{N}{\delta}\right)\right)}$$

$$= \tilde{O}\big(N|\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|s}\big),$$

where we use the definition of C_δ in Eq. (9).

Joining the upper bounds on terms (i) and (ii) we obtain that

$$\sum_{t=1}^{s} \beta_s \leq \tilde{O}\left(\frac{N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|s}\right).$$

Summing over $t \in [T]$ we then get that

$$\sum_{t=1}^{T} \|\tilde{\rho}_{t+1} - \rho\|_1 \leq \sum_{t=1}^{T} \sum_{s=1}^{t} \alpha^{t-s} \beta_s$$

$$= \sum_{t=1}^{T-1} \alpha^{t-1} \sum_{s=1}^{T-t} \beta_s$$

$$\leq \sum_{t=1}^{T-1} \alpha^{t-1} \tilde{O}\left(\frac{N^2}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|(T-t)}\right)$$

$$\leq \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T}\right),$$

concluding the proof. ∎

As a consequence of Lemma 21, Corollary 22 establishes that the true state-action distribution ρ_t remains, on average, close to the target distribution ρ, despite the fact that in Framework 2, we do not observe ρ_t directly and instead rely on the estimate $\tilde{\rho}_t$ to compute the policies.

Corollary 22 *With high probability, we have that*

$$\sum_{t=1}^{T} \|\rho_t - \rho\|_1 \leq \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T}\right).$$

Proof The proof follows from the triangle inequality, then from Prop. 7 and Lemma 21:

$$\sum_{t=1}^{T} \|\rho_t - \rho\|_1 \leq \sum_{t=1}^{T} \|\rho_t - \tilde{\rho}_t\|_1 + \sum_{t=1}^{T} \|\tilde{\rho}_t - \rho\|_1$$

$$\leq \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T}\right).$$

∎

New online mirror descent proof: We present here a result analogous to Prop. 6, concerning the policy computed via online mirror descent (OMD), but now accounting for the fact that OMD is performed as in Eq. (25), over the set of state-action distributions initialized at $\tilde{\rho}_t$ at the start of each episode.

Lemma 23 *For each episode t, we let $\hat{\mu}_t^{\pi_t, \tilde{\rho}_t}$ be the solution of Eq. (25). Suppose further that $\|\nabla \psi(\hat{\mu}_t^{\pi_t, \tilde{\rho}_t})\|_{1,\infty} \leq \psi$. Then, with high probability,*

$$\sum_{t=1}^{T} \langle \tilde{\ell}_t - \bar{b}_t, \hat{\mu}_t^{\pi_t, \tilde{\rho}_t} - \hat{\mu}_t^{\pi, \tilde{\rho}_t} \rangle = \tilde{O}\left(\frac{\ell N^2}{1-\alpha} |\mathcal{X}|^{5/4} |\mathcal{A}|^{1/4} \Psi^{1/2} T^{3/4} \right).$$

For π a periodic policy, we apply Lemma 16 with $\nu_t = \hat{\mu}_t^{\pi, \tilde{\rho}_t}$. We take $q_t = \hat{p}_t$, and $\rho_t = \tilde{\rho}_t$, with $\rho_1 = \rho$, such that $\widetilde{\mathcal{M}}_t = \mathcal{M}_{\tilde{\rho}_t}^{\hat{p}_t}$ following the definition in Eq. (2). Note that for

$$\widetilde{\mathcal{W}}_t := \{ \mu := (\mu_n)_{n \in [N]} \mid \|\mu_N - \rho\|_1 \leq \langle \mu, b_t \rangle + \alpha \|\tilde{\rho}^t - \rho\|_1 \},$$

we have $\nu_t \in \widetilde{\mathcal{W}}_t \cap \widetilde{\mathcal{M}}_t$ from the new feasibility result in Lemma 19. Recall that $f_{t,n}$ is ℓ-Lipschitz with respect to the norm $\| \cdot \|_1$, hence $\|\tilde{\ell}_{t,n}\|_\infty \leq \ell$ and $\|\bar{b}_{t,n}\|_\infty \leq b := \ell N C_\delta$ for all (n, t). As $\hat{\mu}_t^{\pi_t, \tilde{\rho}_t}$ is solution to Eq. (25), we have that

$$\sum_{t=1}^{T} \langle \tilde{\ell}_t - \bar{b}_t, \hat{\mu}_t^{\pi_t, \tilde{\rho}_t} - \hat{\mu}_t^{\pi, \tilde{\rho}_t} \rangle \leq \eta (2\ell N C_\delta)^2 T + \frac{2}{\eta} N e \log(T) \mathbb{1}_{\{D_\psi = \Gamma\}} - \frac{\psi(\mu^1)}{\eta}$$

$$+ \frac{1}{\eta} \underbrace{\sum_{t=1}^{T} \langle \nabla \psi(\mu_t), \nu_t - \nu_{t+1} \rangle}_{(i)} + \underbrace{\sum_{t=1}^{T} \langle \tilde{\ell}^t - \bar{b}_t, \nu_{t+1} - \nu_t \rangle}_{(ii)},$$

where we also use that for all (n, x, a), $V_T := 1 + \sum_{t=1}^{T} \|\hat{p}_n^t(\cdot|x, a) - \hat{p}_n^{t+1}(\cdot|x, a)\|_1 \leq e \log(T)$ from Lemmas 14 and 15. The analysis of terms (i) and (ii) is similar as in the proof of Prop. 6, except that we now need to account for the difference between two consecutive estimates of the initial state-action distribution:

$$\sum_{t=1}^{T} \|\tilde{\rho}_t - \tilde{\rho}_{t+1}\|_1 \leq \sum_{t=1}^{T} \|\tilde{\rho}_t - \rho\|_1 + \sum_{t=1}^{T} \|\rho - \tilde{\rho}_{t+1}\|_1 \qquad (28)$$

$$\underbrace{\leq}_{\text{Lemma 21}} \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right).$$

Thus, by repeating the analysis of terms (i) and (ii) from the proof of Prop. 6, but replacing occurrences of $\|\rho_t - \rho_{t+1}\|_1$ with $\|\tilde{\rho}_t - \tilde{\rho}_{t+1}\|_1$, and applying the upper bound from Eq. (28), we obtain that

$$(i) \leq \Psi \left[N e \log(T) + \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right) \right],$$

and

$$(ii) \leq 2\ell N C_\delta \left[N e \log(T) + \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right) \right].$$

Therefore, joining the bounds of terms (i) and (ii), and optimizing over η, we obtain that

$$\sum_{t=1}^{T}\langle \tilde{\ell}_t - \bar{b}_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} - \hat{\mu}_t^{\pi,\tilde{\rho}_t}\rangle = \tilde{O}\left(\frac{\ell N^3}{1-\alpha}|\mathcal{X}|^{5/4}|\mathcal{A}|^{1/4}\Psi^{1/2}T^{3/4}\right).$$

I.3. Final periodic regret bound: proof of Theorem 8

Proof We begin by decomposing the periodic regret as follows: for any periodic policy π,

$$
\begin{aligned}
R_T(\pi) &= \sum_{t=1}^{T}\left[F_t(\mu^{\pi_t,\rho_t}) - F_t(\mu^{\pi,\rho})\right] + \sum_{t=1}^{T}\|\rho_t - \rho\|_1 \\
&= \underbrace{\sum_{t=1}^{T}\left[F_t(\mu^{\pi_t,\rho_t}) - F_t(\hat{\mu}_t^{\pi_t,\rho_t})\right]}_{R_T^{\text{MDP}}} + \underbrace{\sum_{t=1}^{T}\left[F_t(\hat{\mu}_t^{\pi_t,\rho_t}) - F_t(\hat{\mu}_t^{\pi_t,\tilde{\rho}_t})\right]}_{R_T^{\text{diff. }\tilde{\rho}_t - \rho_t}} \\
&\quad + \underbrace{\sum_{t=1}^{T}\left[F_t(\hat{\mu}_t^{\pi_t,\tilde{\rho}_t}) - F_t(\mu^{\pi,\rho})\right]}_{R_T^{\text{policy}}} + \gamma\sum_{t=1}^{T}\|\rho_t - \rho\|_1.
\end{aligned}
$$

We analyze each term individually

R_T^{MDP}: The results of Prop. 17 still apply to this term, hence for any $\delta \in (0,1)$, with probability at least $1 - 2\delta$,

$$R_T^{\text{MDP}} \le \ell N^2\left(3\sqrt{2}|\mathcal{X}|\sqrt{|\mathcal{A}|T\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)} + 2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right).$$

$R_T^{\text{diff. } \tilde{\rho}_t - \rho_t}$: Using the convexity of F_t, Holder's inequality, and that $f_{t,n}$ is ℓ-Lipschitz with respect to the norm $\| \cdot \|_1$, we have that

$$
\begin{aligned}
R_T^{\text{diff. } \tilde{\rho}_t - \rho_t} &= \sum_{t=1}^{T} F_t(\hat{\mu}_t^{\pi_t, \rho_t}) - F_t(\hat{\mu}_t^{\pi_t, \tilde{\rho}_t}) \\
&\leq \sum_{t=1}^{T} \sum_{n=1}^{N} \langle \nabla f_{t,n}(\hat{\mu}_{t,n}^{\pi_t, \rho_t}), \hat{\mu}_{t,n}^{\pi_t, \rho_t} - \hat{\mu}_{t,n}^{\pi_t, \tilde{\rho}_t} \rangle \\
&\leq \ell \sum_{t=1}^{T} \sum_{n=1}^{N} \| \hat{\mu}_{t,n}^{\pi_t, \tilde{\rho}_t} - \hat{\mu}_{t,n}^{\pi_t, \rho_t} \|_1 \\
&\underbrace{\leq}_{\text{Lemma } 10} \ell N \sum_{t=1}^{T} \| \rho_t - \tilde{\rho}_t \|_1 \\
&\underbrace{\leq}_{\text{Prop. } 7} \frac{\ell N}{1 - \alpha} \left[3N|\mathcal{X}| \sqrt{2|\mathcal{A}|T \log \left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta} \right)} + 2N|\mathcal{X}| \sqrt{2T \log \left(\frac{N}{\delta} \right)} \right] \\
&= \tilde{O} \left(\frac{\ell N^2}{1 - \alpha} |\mathcal{X}| \sqrt{|\mathcal{A}|T} \right).
\end{aligned}
$$

Regularization term : By applying Corollary 22

$$
\sum_{t=1}^{T} \| \rho_t - \rho \|_1 \leq \tilde{O} \left(\frac{\gamma N^2}{(1 - \alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right).
$$

R_T^{policy}: Using the convexity of F_t, and defining $\tilde{\ell}_t := \nabla F_t(\hat{\mu}_t^{\pi_t, \tilde{\rho}_t})$, we decompose this term in three parts:

$$
\begin{aligned}
R_T^{\text{policy}} &\leq \sum_{t=1}^{T} \langle \tilde{\ell}_t, \hat{\mu}_t^{\pi_t, \tilde{\rho}_t} - \mu^{\pi, \rho} \rangle \\
&\leq \underbrace{\sum_{t=1}^{T} \langle \tilde{\ell}_t - \bar{b}_t, \hat{\mu}_t^{\pi_t, \tilde{\rho}_t} - \hat{\mu}_t^{\pi, \tilde{\rho}_t} \rangle}_{R_T^{\text{MD}}} + \underbrace{\sum_{t=1}^{T} \langle \tilde{\ell}_t - \bar{b}_t, \hat{\mu}_t^{\pi, \tilde{\rho}_t} - \hat{\mu}_t^{\pi, \rho} \rangle}_{R_T^{\text{diff. } \tilde{\rho}_t - \rho}} \\
&\quad + \underbrace{\sum_{t=1}^{T} \langle \bar{b}_t, \hat{\mu}_t^{\pi_t, \tilde{\rho}_t} - \hat{\mu}_t^{\pi, \rho} \rangle + \sum_{t=1}^{T} \langle \tilde{\ell}_t, \hat{\mu}_t^{\pi, \rho} - \mu^{\pi, \rho} \rangle}_{R_T^{\text{bonus}}}.
\end{aligned}
$$

We again break the analysis of each term:

- R_T^{MD}: The bound for this term follows directly from Lemma 23, which gives us that with high probability
$$
R_T^{\text{MD}} \leq \tilde{O} \left(\frac{\ell N^3}{1 - \alpha} |\mathcal{X}|^{5/4} |\mathcal{A}|^{1/4} \Psi^{1/2} T^{3/4} \right).
$$

- $R_T^{\text{diff. } \tilde{\rho}_t - \rho}$: Using Holder's inequality, and that $\|\tilde{\ell}_t - \bar{b}_t\|_\infty \leq 2\ell N C_\delta$, we have that

$$
\begin{aligned}
R_T^{\text{diff. } \tilde{\rho}_t - \rho} &= \sum_{t=1}^T \langle \tilde{\ell}_t - \bar{b}_t, \hat{\mu}_t^{\pi,\tilde{\rho}_t} - \hat{\mu}_t^{\pi,\rho} \rangle \\
&\leq 2\ell N C_\delta \sum_{t=1}^T \sum_{n=1}^N \|\hat{\mu}_t^{\pi,\tilde{\rho}_t} - \hat{\mu}_t^{\pi,\rho}\|_1 \\
&\underbrace{\leq}_{\text{Lemma } 10} 2\ell N^2 C_\delta \sum_{t=1}^T \|\tilde{\rho}_t - \rho\|_1 \\
&\underbrace{\leq}_{\text{Lemma. } 21} \tilde{O}\left(\frac{\ell N^4}{(1-\alpha)^2} |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T} \right).
\end{aligned}
$$

- R_T^{bonus} : Following the proof of Prop. 18, we obtain the following decomposition:

$$
\begin{aligned}
R_T^{\text{bonus}} &= \sum_{t=1}^T \langle \bar{b}_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} - \hat{\mu}_t^{\pi,\rho} \rangle + \sum_{t=1}^T \langle \tilde{\ell}_t, \hat{\mu}_t^{\pi,\rho} - \mu^{\pi,\rho} \rangle \\
&\leq \sum_{t=1}^T \langle \bar{b}_t, \hat{\mu}_t^{\pi_t,\tilde{\rho}_t} - \hat{\mu}_t^{\pi,\rho} \rangle + \sum_{t=1}^T \langle \bar{b}_t, \hat{\mu}_t^{\pi,\rho} \rangle + \sum_{t=1}^T \langle \bar{b}_{t,0}, \rho \rangle \\
&= \underbrace{\sum_{t=1}^T \langle \bar{b}_t, \hat{\mu}_t^{\pi_t \tilde{\rho}_t} \rangle + \langle \bar{b}_{t,0}, \tilde{\rho}_t \rangle}_{(i)} + \underbrace{\sum_{t=1}^T \langle \bar{b}_{t,0}, \rho - \tilde{\rho}_t \rangle}_{(ii)} .
\end{aligned}
$$

We begin by upper bounding term (i). Since, by definition, $\hat{\mu}_{t,0}^{\pi_t,\tilde{\rho}_t} = \tilde{\rho}_t$, Corollary 20 implies that

$$
(i) \leq \tilde{O}\left(\frac{\ell N^3}{1-\alpha} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right).
$$

As for term (ii), using Holder's inequality, that $\|\bar{b}_{t,0}\|_1 \leq \ell N C_\delta$, and Lemma 21, we have that

$$
\begin{aligned}
(ii) &= \sum_{t=1}^T \langle \bar{b}_{t,0}, \rho - \tilde{\rho}_t \rangle \\
&\leq \sum_{t=1}^T \|\bar{b}_{t,0}\|_\infty \|\rho - \tilde{\rho}_t\|_1 \\
&\leq \ell N C_\delta \sum_{t=1}^T \|\rho - \tilde{\rho}_t\|_1 \\
&\leq \ell N C_\delta \tilde{O}\left(\frac{N^2}{(1-\alpha)^2} |\mathcal{X}|^{3/2} \sqrt{|\mathcal{A}|T} \right) \\
&= \tilde{O}\left(\frac{\ell N^3}{(1-\alpha)^2} |\mathcal{X}|^2 \sqrt{|\mathcal{A}|T} \right).
\end{aligned}
$$

Joining the upper bounds on terms (i) and (ii) we obtain that

$$R_T^{\text{bonus}} = \tilde{O}\left(\frac{\ell N^3}{(1-\alpha)^2}|\mathcal{X}|^2\sqrt{|\mathcal{A}|T}\right).$$

Conclusion: By summing the upper bounds from the terms of the periodic regret decomposition above, we obtain the final upper bound on the periodic regret of Algorithm 3.

■