Online learning in CMDPs with adversarial losses and stochastic hard constraints

Francesco Emanuele Stradi
Politecnico di Milano
francescoemanuele.stradi@polimi.it

Matteo Castiglioni
Politecnico di Milano
matteo.castiglioni@polimi.it

Alberto Marchesi
Politecnico di Milano
alberto.marchesi@polimi.it

Nicola Gatti
Politecnico di Milano
nicola.gatti@polimi.it

Abstract

We study online learning in *constrained Markov decision processes* (CMDPs) with *adversarial losses* and *stochastic hard constraints*, under *bandit* feedback. We consider two different scenarios. In the first one, we address general CMDPs, where we design an algorithm attaining sublinear regret and cumulative *positive constraints violation*. In the second scenario, under the mild assumption that a policy strictly satisfying the constraints exists and is known to the learner, we design an algorithm that achieves sublinear regret while ensuring that constraints are satisfied *at every episode* with high probability. To the best of our knowledge, our work is the first to study CMDPs involving both adversarial losses and hard constraints. Indeed, previous works either focus on much weaker soft constraints—allowing for positive violation to cancel out negative ones—or are restricted to stochastic losses. Thus, our algorithms can deal with general non-stationary environments subject to requirements much stricter than those manageable with state-of-the-art ones. This enables their adoption in a much wider range of real-world applications, ranging from autonomous driving to online advertising and recommender systems.

1 Introduction

Reinforcement learning [Sutton and Barto, 2018] studies problems where a learner sequentially takes actions in an environment modeled as a *Markov decision process* (MDP) [Puterman, 2014]. Most of the algorithms for such problems focus on learning policies that prescribe the learner how to take actions so as to minimize losses (equivalently, maximize rewards). However, in many real-world applications, the learner must fulfill additional requirements. For instance, autonomous vehicles must avoid crashing [Wen et al., 2020, Isele et al., 2018], bidding agents in ad auctions must not deplete their budget [Wu et al., 2018, He et al., 2021], recommender systems must not present offending items to their users [Singh et al., 2020] and dynamic pricing platforms must satisfy different sale constraints [Stradi et al., 2024a]. A commonly-used model that allows to capture such additional requirements is the *constrained* MDP (CMDP) [Altman, 1999], where the goal is to learn a loss-minimizing policy while at the same time satisfying some constraints.

We study online learning problems in *episodic* CMDPs with *adversarial losses* and *stochastic hard constraints*, under *bandit* feedback. In such settings, the goal of the learner is to minimize their *regret*—the difference between their cumulative loss and what they would have obtained by always selecting a best-in-hindsight policy—, while at the same time guaranteeing that the constraints are satisfied during the learning process. We consider two scenarios that differ in the way in which constraints are satisfied and are both usually referred to as *hard* constraints settings in the literature [Liu et al., 2021].

17th European Workshop on Reinforcement Learning (EWRL 2024).

In the first scenario, the learner aims at minimizing the *cumulative positive constraints violation*, while, in the second one, the learner's goal is to satisfy constraints at every episode.

To the best of our knowledge, our work is the first to study CMDPs that involve both adversarial losses and hard constraints. Indeed, all the works on adversarial CMDPs (see, *e.g.*, [Wei et al., 2018, Qiu et al., 2020]) consider settings with *soft* constraints. These are much weaker than hard constraints, as they are only concerned with the minimization of the cumulative (both positive and negative) constraints violation. As a result, they allow negative violations to cancel out positive ones across different episodes. Such cancellations are unreasonable in real-world applications. For instance, in autonomous driving, avoiding a collision clearly does *not* "repair" a crash occurred previously. Furthermore, the only few works addressing stochastic hard constraints in CMDPs [Liu et al., 2021, Shi et al., 2023] are restricted to *stochastic losses*. Thus, our CMDP settings capture many more applications than theirs, since being able to deal with adversarial losses allows to tackle general non-stationary environments, which are ubiquitous in the real world.

1.1 Original contributions

We start by addressing the first scenario, where we design an algorithm—called Bounded Violation Optimistic Policy Search (`BV-OPS`)—that guarantees both sublinear regret and sublinear cumulative positive constraints violation. `BV-OPS` builds on top of state-of-the-art learning algorithms in adversarial, unconstrained MDPs, by introducing the tools necessary to deal with constraints violation. Specifically, `BV-OPS` works by selecting policies that *optimistically* satisfy the constraints. `BV-OPS` updates the set of such policies in an online fashion, guaranteeing that it is always non-empty with high probability and that it collapses to the (true) set of constraints-satisfying policies as the number of episodes increases. This allows `BV-OPS` to attain sublinear violation. Crucially, even though such an "optimistic" set of policies changes during the execution of the algorithm, it always contains the (true) set of constraints-satisfying policies. This allows `BV-OPS` to attain sublinear regret. `BV-OPS` also addresses a problem left open by Qiu et al. [2020], *i.e.*, learning with *bandit* feedback in CMDPs with adversarial losses and stochastic constraints. Indeed, `BV-OPS` goes even further, as Qiu et al. [2020] were only concerned with soft constraints, while `BV-OPS` deals with *positive* violation.

Next, we switch the attention to the second scenario, where our goal is to design a *safe* algorithm, namely, one that satisfies the constraints at every episode. In order to achieve such a goal, we need to assume that the learner has knowledge about a policy strictly satisfying the constraints. Indeed, this is necessary even in simple stochastic multi-armed bandit settings, as shown in [Bernasconi et al., 2022]. This scenario begets considerable additional challenges compared to the first one, since assuring the safety property extremely limits the exploration capabilities of algorithms, rendering techniques for adversarial, unconstrained MDPs inapplicable. Nevertheless, we design an algorithm—called Safe Optimistic Policy Search (`S-OPS`)—that attains sublinear regret while being safe with high probability. `S-OPS` works by selecting, at each episode, a suitable randomization between the policy that `BV-OPS` would choose and the (known) policy strictly satisfying the constraints. As a result, `S-OPS` effectively plays *non-Markovian* policies. Crucially, the probability defining the randomization employed by the algorithm is carefully chosen in order to *pessimistically* account for constraints satisfaction. This guarantees that a sufficient amount of exploration is performed.

1.2 Related works

Online learning [Cesa-Bianchi and Lugosi, 2006, Orabona, 2019] in MDPs has received considerable attention over the last decade (see, *e.g.*, [Auer et al., 2008, Even-Dar et al., 2009, Neu et al., 2010]). Two types of feedback are usually investigated: *full feedback*, with the entire loss function being observed by the learner, and *bandit feedback*, where the learner only observes the loss of chosen actions. Notably, Azar et al. [2017] study learning in episodic MDPs with unknown transitions and stochastic losses under bandit feedback, achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and matching the lower bound for these MDPs. Rosenberg and Mansour [2019b] study learning under full feedback in episodic MDPs with adversarial losses and unknown transitions, presenting an algorithm that attains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. The same setting is studied by Rosenberg and Mansour [2019a] under bandit feedback, obtaining a suboptimal $\widetilde{\mathcal{O}}(T^{3/4})$ regret. Jin et al. [2020] provide an algorithm with an optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, in the same setting. Bacchiocchi et al. [2023] study online learning in adversarial MDPs providing

regret bounds which depend on a behavioral policy. Finally, Maran et al. [2024] study online MDPs with stochastic losses when the agent is the configurator, under *bandit feedback*.

Online learning in CMDPs has generally been studied with stochastic losses and constraints. Zheng and Ratliff [2020] deal with fully-stochastic episodic CMDPs, assuming known transitions and bandit feedback. The regret of their algorithm is $\widetilde{\mathcal{O}}(T^{3/4})$, while its cumulative constraints violation is guaranteed to be below a threshold with a given probability. Bai et al. [2023] provide the first algorithm that achieves sublinear regret with unknown transitions, assuming that the rewards are deterministic and the constraints are stochastic with a particular structure. Efroni et al. [2020] propose two approaches to deal with the exploration-exploitation trade-off in episodic CMDPs. The first one resorts to a linear programming formulation of CMDPs and obtains sublinear regret and cumulative positive constraints violation. The second one relies on a primal-dual formulation of the problem and guarantees sublinear regret and cumulative (positive/negative) constraints violation, when transitions, losses, and constraints are unknown and stochastic, under bandit feedback. Liu et al. [2021] study stochastic *hard* constraints; however, the authors only focus on stochastic losses. Recently, Shi et al. [2023] study stochastic hard constraints on both states and actions. As concerns adversarial settings, [Wei et al., 2018, Qiu et al., 2020, Stradi et al., 2024b] address CMDPs with adversarial losses, but they only provide guarantees in terms of *soft* constraints. Moreover, [Wei et al., 2023, Ding and Lavaei, 2023, Stradi et al., 2024c] consider non-stationary losses/constraints with bounded variation. Thus, their results do *not* apply to general adversarial losses. Finally,Bacchiocchi et al. [2024] study CMDPs with partial observability on the constraints.

In conclusion, learning with hard constraints has been studied in online convex optimization [Guo et al., 2022], and also in stochastic settings with a simple tree-like sequential structure [Chen et al., 2018, Bernasconi et al., 2022]. Our results are much more general than those, since we jointly consider adversarial losses, bandit feedback, and an MDP sequential structure.

2 Preliminaries

2.1 Constrained Markov decision processes

We study online learning in *episodic constrained* MDPs [Altman, 1999] with *adversarial losses* and *stochastic constraints* (CMDPs for short). These are tuples $M := \left(X, A, P, \{\ell_t\}_{t=1}^T, \{G_t\}_{t=1}^T, \alpha\right)$:

- T is the number of episodes.[1]
- X and A are finite state and action spaces, respectively.
- $P : X \times A \times X \to [0, 1]$ is the transition function, where, for ease of notation, we denote by $P(x'|x, a)$ the probability of going from state $x \in X$ to $x' \in X$ by taking action $a \in A$.[2]
- $\{\ell_t\}_{t=1}^T$ is the sequence of vectors defining the losses at each episode $t \in [T]$, namely $\ell_t \in [0, 1]^{|X \times A|}$. We refer to the loss for a state-action pair $(x, a) \in X \times A$ as $\ell_t(x, a)$. Losses are adversarial, namely, no statistical assumption on how they are selected is made.
- $\{G_t\}_{t=1}^T$ is the sequence of matrices defining the *costs* that characterize the m constraints at each $t \in [T]$, namely $G_t \in [0, 1]^{|X \times A| \times m}$. For $i \in [m]$, the i-th constraint cost for a state-action pair $(x, a) \in X \times A$ is denoted by $g_{t,i}(x, a)$. Costs are stochastic, namely, the matrices G_t are i.i.d. random variables distributed according to a probability distribution $\mathcal{G}$.
- $\alpha = [\alpha_1, \ldots, \alpha_m] \in [0, L]^m$ is the vector of cost *thresholds* that characterize the m constraints, where α_i denotes the threshold for the i-th constraint.

At each episode of a CMDP, the learner chooses a *policy* $\pi : X \times A \to [0, 1]$, which defines a probability distribution over actions at each state. For ease of notation, we denote by $\pi(\cdot|x)$ the probability distribution of state $x \in X$, with $\pi(a|x)$ denoting the probability of action $a \in A$.

[1] We denote an episode by $t \in [T]$, where $[a \ldots b]$ is the set of all integers from a to b and $[b] := [1 \ldots b]$.

[2] In this paper, for ease of notation, we focus w.l.o.g. on *loop-free* CMDPs. This means that X is partitioned into $L + 1$ layers $X_0, \ldots, X_L$ with $X_0 = \{x_0\}$ and $X_L = \{x_L\}$. Moreover, the loop-free property requires that $P(x'|x, a) > 0$ only if $x' \in X_{k+1}$ and $x \in X_k$ for some $k \in [0 \ldots L - 1]$. Notice that any (episodic) CMDP with horizon H that is *not* loop-free can be cast into a loop-free one by suitably duplicating the state space H times, *i.e.*, a state x is mapped to a set of new states (x, k) with $k \in [H]$. In loop-free CMDPs, we let $k(x) \in [0 \ldots L]$ be the index of the layer which state $x \in X$ belongs to.

Algorithm 1 details the interaction between the learner and the environment in a CMDP. Notice that we assume that the learner has *bandit feedback*. In particular, the learner receives as feedback the trajectory of state-action pairs (x_k, a_k), for $k \in [0 \dots L-1]$, visited during the episode, as well as their losses $\ell_t(x_k, a_k)$ and costs $g_{t,i}(x_k, a_k)$ for $i \in [m]$. We assume that the learner knows X and A, but they do *not* know anything about the transition function P.

Algorithm 1 CMDP Interaction at episode $t \in [T]$

1: ℓ_t, G_t chosen *adversarially* and *stochastically*, resp.
2: Learner chooses a policy $\pi_t : X \times A \to [0,1]$
3: Environment is initialized to state x_0
4: **for** $k = 0, \dots, L-1$ **do**
5: Learner takes action $a_k \sim \pi_t(\cdot|x_k)$
6: Learner sees $\ell_t(x_k, a_k)$, $g_{t,i}(x_k, a_k) \forall i \in [m]$
7: Environment evolves to $x_{k+1} \sim P(\cdot|x_k, a_k)$
8: Learner observes the next state x_{k+1}

2.2 Occupancy measures

Next, we introduce the notion of *occupancy measure* [Rosenberg and Mansour, 2019a]. Given a transition function P and a policy π, the occupancy measure $q^{P,\pi} \in [0,1]^{|X \times A \times X|}$ induced by P and π is such that, for every $x \in X_k$, $a \in A$, and $x' \in X_{k+1}$ with $k \in [0 \dots L-1]$, it holds $q^{P,\pi}(x, a, x') = \mathbb{P}[x_k = x, a_k = a, x_{k+1} = x' | P, \pi]$. Moreover, we also define:

$$q^{P,\pi}(x,a) = \sum_{x' \in X_{k+1}} q^{P,\pi}(x,a,x') \quad \text{and} \quad q^{P,\pi}(x) = \sum_{a \in A} q^{P,\pi}(x,a). \tag{1}$$

The next lemma characterizes *valid* occupancy measures.

Lemma 2.1 (Rosenberg and Mansour [2019b])**.** *A vector $q \in [0,1]^{|X \times A \times X|}$ is a valid occupancy measure of an episodic loop-free MDP if and only if the following holds:*

$$\begin{cases} \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x,a,x') = 1 & \forall k \in [0 \dots L-1] \\ \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x,a,x') = \sum_{x' \in X_{k-1}} \sum_{a \in A} q(x',a,x) & \forall k \in [1 \dots L-1], \forall x \in X_k \\ P^q = P, \end{cases}$$

where P is the transition function of the MDP and P^q is the one induced by q (see Equation (2)*).*

Notice that any valid occupancy measure q induces a transition function P^q and a policy π^q, with:

$$P^q(x'|x,a) = \frac{q(x,a,x')}{q(x,a)} \quad \text{and} \quad \pi^q(a|x) = \frac{q(x,a)}{q(x)}. \tag{2}$$

2.3 Baseline

Our *baseline* for evaluating the performances of the learner is defined through a linear programming formulation of the (offline) learning problem in constrained MDPs. Specifically, given a constrained MDP $M := (X, A, P, \ell, G, \alpha)$ characterized by a loss vector $\ell \in [0,1]^{|X \times A|}$, a cost matrix $G \in [0,1]^{|X \times A| \times m}$, and a threshold vector $\alpha \in [0, L]^m$, such a problem consists in finding a policy minimizing the loss while ensuring that all the constraints are satisfied. Thus, our baseline $\text{OPT}_{\ell,G,\alpha}$ is defined as the optimal value of a parametric linear program, which reads as follows:

$$\text{OPT}_{\ell,G,\alpha} := \begin{cases} \min_{q \in \Delta(M)} & \ell^\top q \quad \text{s.t.} \\ & G^\top q \leq \alpha, \end{cases} \tag{3}$$

where $q \in [0,1]^{|X \times A|}$ is a vector encoding an occupancy measure whose entries are defined as in Equation (1), while $\Delta(M)$ is the set of valid occupancy measures. Notice that, given the equivalence between policy and occupancy, the (offline) learning problem can be formulated as a linear program working in the space of the occupancy measures q, since expected losses and costs are linear in q.

2.4 Online learning with hard constraints

As customary in settings with adversarial losses, we measure the performance of a learning algorithm by comparing it with the *best-in-hindsight constraint-satisfying policy*. The performance of the

learner is evaluated in terms of the *(cumulative) regret* $R_T := \sum_{t=1}^{T} \ell_t^\top q^{P,\pi_t} - T \cdot \text{OPT}_{\overline{\ell},\overline{G},\alpha}$, where $\overline{\ell} := \frac{1}{T}\sum_{t=1}^{T} \ell_t$ is the average of the adversarial losses over the T episodes and $\overline{G} := \mathbb{E}_{G\sim\mathcal{G}}[G]$ is the expected value of the stochastic cost matrices. For ease of presentation, we let q^* be a best-in-hindsight constraint-satisfying occupancy measure, *i.e.*, one achieving value $\text{OPT}_{\overline{\ell},\overline{G},\alpha}$, while we let π^* be its corresponding policy. Thus, the regret reduces to $R_T := \sum_{t=1}^{T} \ell_t^\top (q^{P,\pi_t} - q^*)$. For ease of notation, we refer to q^{P,π_t} by simply using q_t, thus omitting the dependency on P and π_t.

Our goal is to design learning algorithms with regret growing sublinearly in T, namely $R_T = o(T)$, while at the same time ensuring that the m constraints are satisfied. In this work, we consider two different settings, both usually falling under the umbrella of *hard constraints* settings in the literature [Guo et al., 2022]. In the first one (Section 2.4.1), constraints satisfaction is measured by the cumulative *positive* constraints violation incurred by the algorithm. In the second one (Section 2.4.2), the goal is to design algorithms ensuring that the constraints are satisfied at every episode.

2.4.1 Guaranteeing bounded violation

In this setting, our objective is expressed in terms of *cumulative (positive) constraints violation* $V_T := \max_{i\in[m]} \sum_{t=1}^{T} \left[\overline{G}^\top q_t - \alpha\right]_i^+$, where where we let $[x]^+ := \max\{0, x\}$. Our goal is to design algorithms with sublinear V_T, namely $V_T = o(T)$. To achieve such a goal, we only need to assume that the problem is well posed, namely, there exists a policy satisfying the constraints in expectation.

Assumption 2.2. There is an occupancy measure $q^\diamond$, called *feasible solution*, such that $\overline{G}^\top q^\diamond \leq \alpha$.

2.4.2 Guaranteeing safety

In this setting, our goal is to design algorithms ensuring that the following *safety property* is met:

Definition 2.3 (Safe algorithm)**.** An algorithm is *safe* if and only if $\overline{G}^\top q_t \leq \alpha$ for all $t \in [T]$.

As shown by Bernasconi et al. [2022], without further assumptions, it is *not* possible to achieve $R_T = o(T)$ while at the same time guaranteeing that the safety property holds with high probability, even in simple stochastic multi-armed bandit instances. To design safe learning algorithms, we need the following two assumptions. The first one is about the possibility of *strictly* satisfying constraints.

Assumption 2.4 (Slater's condition)**.** There exists an occupancy measure $q^\diamond$ such that $\overline{G}^\top q^\diamond < \alpha$. We call $q^\diamond$ *strictly feasible solution*, while a policy $\pi^\diamond$ induced by $q^\diamond$ is called *strictly feasible policy*.

The second assumption is related to learner's knowledge about a strictly feasible policy.

Assumption 2.5. The policy $\pi^\diamond$ and its costs $\beta = [\beta_1, \ldots, \beta_m] := \overline{G}^\top q^\diamond$ are known to the learner.

Intuitively, Assumption 2.5 is needed to guarantee that safety holds during the first episodes, namely, when learner's uncertainty about costs is high. Notice that Assumptions 2.4 and 2.5 are often employed in CMDPs (see, *e.g.*, [Liu et al., 2021]), as they are usually met in real-world applications of interest, where it is common to have access to a "do-nothing" policy resulting in *no* constraints costs.

3 Concentration bounds

In the following Sections 4 and 5, we design two algorithms that work by estimating expected values of the stochastic parameters in a CMDP, namely costs and transitions. In this section, as a preliminary step towards the analysis of our algorithms, we provide concentration bounds for such estimates. Notice that losses need a completely different treatment, since they are selected adversarially.

Concentration bounds for costs Let $N_t(x, a)$ be the total number of episodes up to $t \in [T]$ in which $(x, a) \in X \times A$ is *visited*. Then, $\widehat{g}_{t,i}(x, a) := \frac{\sum_{\tau\in[t]} g_{\tau,i}(x,a)\mathbb{1}_\tau\{x,a\}}{\max\{1, N_t(x,a)\}}$, with $\mathbb{1}_\tau\{x, a\} = 1$ if and only if (x, a) is visited in episode τ, is an unbiased estimator of the expected cost of constraint $i \in [m]$ for (x, a), namely $\overline{g}_i(x, a) := \mathbb{E}_{G\sim\mathcal{G}}[g_{t,i}(x, a)]$. Thus, by applying Hoeffding's inequality:

Lemma 3.1. *Given a confidence parameter $\delta \in (0,1)$, with probability at least $1-\delta$, for every $i \in [m]$, episode $t \in [T]$, and pair $(x,a) \in X \times A$, it holds $|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)| \leq \xi_t(x,a)$, where we let the confidence bound $\xi_t(x,a) := \min\{1, \sqrt{4\ln(T|X||A|m/\delta)/\max\{1, N_t(x,a)\}}\}$.*

For ease of notation, we let $\widehat{G}_t \in [0,1]^{|X\times A|\times m}$ be the matrix of the estimated costs $\widehat{g}_{t,i}(x,a)$. Moreover, we denote by $\xi_t \in [0,1]^{|X\times A|}$ the vector whose entries are the bounds $\xi_t(x,a)$, and we let $\Xi_t \in [0,1]^{|X\times A|\times m}$ be a matrix built by in such a way that the statement of Lemma 3.1 becomes: $|\widehat{G}_t - \overline{G}| \preceq \Xi_t$ holds with probability at least $1-\delta$, where $|\cdot|$ and $\preceq$ are applied component wise. In the following, given any $\delta \in (0,1)$, we refer to the event defined in Lemma 3.1 as $\mathcal{E}^G(\delta)$.

Concentration bounds for transitions Next, we introduce *confidence sets* for the transition function of a CMDP, by exploiting suitable concentration bounds for estimated transition probabilities. By letting $M_t(x,a,x')$ be the total number of episodes up to $t \in [T]$ in which $(x,a) \in X \times A$ is visited and the environment transitions to state $x' \in X$, the estimated transition probability at t for (x,a,x') is $\widehat{P}_t(x' \mid x,a) = \frac{M_t(x,a,x')}{\max\{1,N_t(x,a)\}}$. Then, the confidence set for P at episode $t \in [T]$ is $\mathcal{P}_t := \bigcap_{(x,a,x')\in X\times A\times X} \mathcal{P}_t^{x,a,x'}$, where: $\mathcal{P}_t^{x,a,x'} := \{\overline{P} : |\overline{P}(x'|x,a) - \widehat{P}_t(x'|x,a)| \leq \epsilon_t(x,a,x')\}$, with $\epsilon_t(x,a,x') := 2\sqrt{\frac{\widehat{P}_t(x'|x,a)\ln(T|X||A|/\delta)}{\max\{1,N_t(x,a)-1\}}} + \frac{14\ln(T|X||A|/\delta)}{3\max\{1,N_t(x,a)-1\}}$ for some confidence $\delta \in (0,1)$. The next lemma establishes that $\mathcal{P}_t$ is a proper confidence set.

Lemma 3.2 (Jin et al. [2020])**.** *Given a confidence parameter $\delta \in (0,1)$, with probability at least $1-4\delta$, it holds that the transition function P belongs to $\mathcal{P}_t$ for all $t \in [T]$.*

At each $t \in [T]$, given a confidence set $\mathcal{P}_t$, it is possible to efficiently build a set $\Delta(\mathcal{P}_t)$ that comprises all the occupancy measures that are valid with respect to every transition function $\overline{P} \in \mathcal{P}_t$. For reasons of space, we defer the formal definition of $\Delta(\mathcal{P}_t)$ to Appendix D. Lemma 3.2 implies that, with high probability, the set $\Delta(M)$ of valid occupancy measure is included in all the "estimated" sets $\Delta(\mathcal{P}_t)$, for $t \in [T]$. In the following, given a confidence parameter $\delta \in (0,1)$, we refer to the event $\Delta(M) \subseteq \bigcap_{t\in[T]} \Delta(\mathcal{P}_t)$ as $\mathcal{E}^\Delta(\delta)$, which holds with probability at least $1-4\delta$ thanks to Lemma 3.2. Finally, for ease of presentation, given $\delta \in (0,1)$ we define a *clean event* $\mathcal{E}^{G,\Delta}(\delta)$ in which all the concentration bounds for costs and transitions correctly hold. Formally, $\mathcal{E}^{G,\Delta}(\delta) := \mathcal{E}^G(\delta) \cap \mathcal{E}^\Delta(\delta)$, which holds with probability at least $1-5\delta$ by a union bound (and Lemmas 3.1 and 3.2).

4 Guaranteeing bounded violation

We start by designing an algorithm, called `BV-OPS`, which guarantees that both the regret R_T and the cumulative positive constraints violation V_T grow sublinearly in T. We recall that, in order to get to this result, we only need to assume the existence of a feasible solution (Assumption 2.2).

Dealing with adversarial losses while limiting constraints violation begets considerable challenges, which go beyond classical exploration-exploitation trade-offs faced in unconstrained settings. On the one hand, using state-of-the-art algorithms for online learning in adversarial, unconstrained MDPs would lead to sublinear regret, but constraints violation would grow linearly. On the other hand, a naïve approach that randomly explores to compute a set of policies satisfying the constraints with high probability can lead to sublinear constraints violation, at the cost of suffering linear regret. Thus, a clever adaptation of the techniques employed for unconstrained settings is needed. Our approach builds on top of an algorithm developed by Jin et al. [2020] for adversarial, unconstrained MDPs, by equipping it with the tools necessary to deal with adversarial losses and constraints violation.

4.1 The `BV-OPS` algorithm

Our algorithm—called Bounded Violation Optimistic Policy Search (`BV-OPS`)—works by selecting policies derived from a set of occupancy measures that *optimistically* satisfy cost constraints. Such an "optimistic" set is built in an online fashion by using lower confidence bounds on the costs characterizing the constraints. This ensures that the set is always non-empty with high probability and that it collapses to the (true) set of constraint-satisfying occupancy measures as the number of episodes increases, enabling `BV-OPS` to attain sublinear constraints violation. The fundamental property preserved by `BV-OPS` is that, even though the "optimistic" set changes during the execution

of the algorithm, it always subsumes the (true) set of constraint-satisfying occupancy measures. This crucially allows BV-OPS to employ classical policy-selection methods for unconstrained MDPs.

Algorithm 2 provides the pseudocode of BV-OPS. At the beginning, the algorithm initializes all the counters (Line 2), it sets the occupancy measure $\widehat{q}_1$ for the first episode to be equal to a uniform vector (Line 3), and it selects the policy π_1 for the first episode as the one induced by $\widehat{q}_1$ (Line 4; see the definition of $\pi^{\widehat{q}_1}$ in Equation (2)). At each episode $t \in [T]$, BV-OPS plays policy π_t and receives feedback as described in Algorithm 1 (Line 6). Then, BV-OPS computes an *upper occupancy bound* $u_t(x_k, a_k)$ for every state-action pair (x_k, a_k) visited during Algorithm 1, by using the confidence set for the transition function $\mathcal{P}_{t-1}$ computed in the previous episode, namely, it sets $u_t(x_k, a_k) := \max_{\overline{P} \in \mathcal{P}_{t-1}} q^{\overline{P}, \pi_t}(x, a)$ for every $k \in [0 \dots L-1]$ (Line 7). Intuitively, $u_t(x_k, a_k)$ represents the maximum probability with which (x_k, a_k) is visited when using policy π_t, given the confidence set for the transition function built so far. The upper occupancy bounds are combined with the exploration factor γ to compute an *optimistic loss estimator* $\widehat{\ell}_t(x, a)$ for every state-action pair $(x, a) \in X \times A$ (see Line 8 for its definition). After that, BV-OPS updates all the counters given the path traversed in Algorithm 1 (Lines 10–11), it builds the new confidence set $\mathcal{P}_t$, and it computes the matrices $\widehat{G}_t$ and Ξ_t containing the estimated costs and their corresponding bounds, respectively, by using the received feedback (Line 12).

Algorithm 2 BV-OPS

Require: X, A, α, T, δ, η, γ

1: **for** $k \in [0 \dots L-1]$, $(x, a, x') \in X_k \times A \times X_{k+1}$ **do**
2: $\quad N_0(x, a) \leftarrow 0$; $M_0(x, a, x') \leftarrow 0$
3: $\quad \widehat{q}_1(x, a, x') \leftarrow 1/|X_k||A||X_{k+1}|$
4: $\pi_1 \leftarrow \pi^{\widehat{q}_1}$
5: **for** $t \in [T]$ **do**
6: $\quad$ Choose π_t in Algorithm 1 and receive feedback
7: $\quad$ Build *upper occupancy bounds* for $k \in [0 \dots L-1]$:

$$u_t(x_k, a_k) \leftarrow \max_{\overline{P} \in \mathcal{P}_{t-1}} q^{\overline{P}, \pi_t}(x_k, a_k)$$

8: $\quad$ Build *optimistic loss estimator* for $(x, a) \in X \times A$:

$$\widehat{\ell}_t(x, a) \leftarrow \begin{cases} \frac{\ell_t(x,a)}{u_t(x,a)+\gamma} & \text{if } \mathbb{1}_t\{x, a\} = 1 \\ 0 & \text{otherwise} \end{cases}$$

9: $\quad$ **for** $k \in [0 \dots L-1]$ **do**
10: $\quad\quad N_t(x_k, a_k) \leftarrow N_{t-1}(x_k, a_k) + 1$
11: $\quad\quad M_t(x_k, a_k, x_{k+1}) \leftarrow M_{t-1}(x_k, a_k, x_{k+1}) + 1$
12: $\quad$ Build $\mathcal{P}_t$, $\widehat{G}_t$, and Ξ_t as in Section 3
13: $\quad$ Build *unconstrained occupancy* for all (x, a, x'):

$$\tilde{q}_{t+1}(x, a, x') \leftarrow \widehat{q}_t(x, a, x') e^{-\eta \widehat{\ell}_t(x,a)}$$

14: $\quad$ **if** $\text{PROJ}\big(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\big)$ is *feasible* **then**
15: $\quad\quad \widehat{q}_{t+1} \leftarrow \text{PROJ}\big(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\big)$
16: $\quad$ **else**
17: $\quad\quad \widehat{q}_{t+1} \leftarrow$ any $q \in \Delta(\mathcal{P}_t)$
18: $\quad \pi_{t+1} \leftarrow \pi^{\widehat{q}_{t+1}}$

In order to choose a policy π_{t+1} for the next episode, BV-OPS first computes an *unconstrained occupancy measure* $\tilde{q}_{t+1}$ according to a classical unconstrained OMD update [Orabona, 2019] (see Line 13 for its definition). Then, $\tilde{q}_{t+1}$ is projected on a suitably-defined set of occupancy measures that *optimistically* satisfy the constraints. This latter step is crucial to jointly manage adversarial losses and constraints violation. Next, we formally define the projection step (Line 14).

$$\text{PROJ}\big(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\big) := \begin{cases} \arg \min_{q \in \Delta(\mathcal{P}_t)} & D(q||\tilde{q}_{t+1}) \quad \text{s.t.} \\ & \Big(\widehat{G}_t - \Xi_t\Big)^\top q \leq \alpha, \end{cases} \tag{4}$$

where $D(q||\tilde{q}_{t+1})$ is the unnormalized KL-divergence between q and $\tilde{q}_{t+1}$, which is defined as

$$D(q||\tilde{q}_{t+1}) := \sum_{x,a,x'} q(x, a, x') \ln \frac{q(x, a, x')}{\tilde{q}_{t+1}(x, a, x')} - \sum_{x,a,x'} \Big(q(x, a, x') - \tilde{q}_{t+1}(x, a, x')\Big).$$

Notice that Problem (4) is a linearly-constrained convex mathematical program, and, thus, it can be solved efficiently for an arbitrarily-good approximate solution.[3] Intuitively, Problem (4) performs a projection onto the set of occupancy measures $q \in \Delta(\mathcal{P}_t)$ that additionally satisfy the constraint $\Big(\widehat{G}_t - \Xi_t\Big)^\top q \leq \alpha$, where lower confidence bounds $\widehat{G}_t - \Xi_t$ for the costs are used in order to take an optimistic approach with respect to constraints satisfaction. Finally, if Problem (4) is feasible, then at the next episode BV-OPS selects the policy $\pi^{\widehat{q}_{t+1}}$ induced by a solution $\widehat{q}_{t+1}$ to Problem (4) (Line 15), otherwise it chooses a policy induced by any occupancy measure in $\Delta(\mathcal{P}_t)$ (Line 17).

[3] As customary in adversarial MDPs, we assume that an optimal solution to Problem (4) can be computed efficiently. Indeed, by dropping this assumption, we can still derive all of our results up to small approximations.

The optimistic approach adopted in Problem (4) crucially allows to prove the following lemma.

Lemma 4.1. *Given confidence $\delta \in (0,1)$, Algorithm 2 ensures that* $\text{PROJ}(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t)$ *is feasible at every episode $t \in [T]$ with probability at least $1 - 5\delta$.*

Intuitively, Lemma 4.1 follows from the fact that, under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, the set on which the projection is performed subsumes the (true) set of constraints-satisfying occupancy measures. Lemma 4.1 is fundamental, as it allows to prove that `BV-OPS` attains sublinear V_T and R_T.

4.2 Cumulative constraints violation

In order to prove that the cumulative constraints violation achieved by `BV-OPS` is sublinear, we exploit the fact that both the concentration bounds for costs and those associated with transition probabilities shrink at a rate of $\mathcal{O}(1/\sqrt{T})$. This allows us to show the following result.

Theorem 4.2. *Given $\delta \in (0,1)$, Algorithm 2 attains cumulative positive constraints violation $V_T \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left({}^{T|X||A|m}/_{\delta}\right)}\right)$ with probability at least $1 - 8\delta$.*

4.3 Cumulative regret

The crucial observation that allows us to prove that the regret attained by `BV-OPS` grows sublinearly in T is that the set on which the algorithm perform its projection step (Problem (4)) always contains the (true) set of occupancy measures that satisfy the cost constraints, and, thus, it also always contains the best-in-hindsight constraint-satisfying occupancy measure q^*. As a result, even though cost estimates may be arbitrarily bad during the first episodes, `BV-OPS` is still guaranteed to select policies resulting in losses that are smaller than or equal to those incurred by q^*. This allows us to show the following:

Theorem 4.3. *Given $\delta \in (0,1)$, by setting $\eta = \gamma = \sqrt{{}^{L\ln(L|X||A|/\delta)}/_{T|X||A|}}$ in Algorithm 2, the algorithm attains regret $R_T \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left({}^{T|X||A|}/_{\delta}\right)}\right)$ with probability at least $1 - 10\delta$.*

5 Guaranteeing safety

In this section, we design another algorithm, called `S-OPS`, attaining sublinear regret and enjoying the safety property with high probability. In order to do this, we work under Assumptions 2.4 and 2.5. Designing safe algorithms raises many additional challenges compared to the case studied in Section 4, where one seeks for the weaker goal of sublinear cumulative positive constraints violation. Indeed, adapting techniques for adversarial, unconstrained MDPs does *not* work anymore, and, thus, *ad hoc* approaches are needed. This is because adhering to the safety property extremely limits exploration.

5.1 The `S-OPS` algorithm

Our algorithm—Safe Optimistic Policy Search (`S-OPS`)—builds on top of the `BV-OPS` algorithm developed in Section 4. Selecting policies derived from the "optimistic" set of occupancy measures, as done by `BV-OPS`, is *not* sufficient anymore, as it would clearly result in the safety property being unsatisfied during the first episodes. Our new algorithm circumvents such an issue by employing, at each episode, a suitable randomization between the policy derived from the "optimistic" set (the one `BV-OPS` would select) and the strictly feasible policy $\pi^\diamond$. Crucially, as we show next, such a randomization accounts for constraints satisfaction by taking a *pessimistic* approach, namely, by considering upper confidence bounds on the costs characterizing the constraints. This is needed in order to guarantee the safety property. Moreover, having access to the strictly feasible policy $\pi^\diamond$ and its expected costs β (Assumption 2.5) allows `S-OPS` to always place a sufficiently large probability on the policy derived from the "optimistic" set, so that a sufficient amount of exploration is guaranteed, and, in its turn, sublinear regret is attained. Notice that `S-OPS` effectively selects *non-Markovian* policies, as it employs a randomization between two Markovian policies at each episode.

Algorithm 3 provides the pseudocode of `S-OPS`. Differently from `BV-OPS`, the policy selected at the first episode is *not* the one derived from a uniform occupancy measure, but it is obtained by randomizing the latter with the strictly feasible policy $\pi^\diamond$ (Line 4). The probability λ_0 of selecting $\pi^\diamond$ is chosen pessimistically. Intuitively, in the first episode, being pessimistic means that λ_0 must

guarantee that the constraints are satisfied for any possible choice of costs and transitions, and, thus, $\lambda_0 := \max_{i\in[m]} \{^{L-\alpha_i}/_{L-\beta_i}\}$. Thanks to Assumptions 2.4 and 2.5, it is always the case that $\lambda_0 < 1$. Thus, $\pi_1 \neq \pi^\diamond$ with positive probability and some exploration is performed even in the first episode.

Analogously to `BV-OPS`, at each $t \in [T]$, `S-OPS` selects a policy π_t and receives feedback as described in Algorithm 1, it computes optimistic loss estimators, it updates the confidence set for the transitions, and it computes the matrices of estimated costs and their bounds. Then, as in `BV-OPS`, an update step of unconstrained OMD is performed. Although identical to the update done in `BV-OPS`, the one in `S-OPS` uses loss estimators computed when using a randomization between the policy obtained by solving Problem (4) and the strictly feasible policy $\pi^\diamond$. Thus, there is a mismatch between the occupancy measure used to estimate losses and the one computed by the projection step.

The projection step performed by `S-OPS` (Line 14) is the same as the one in `BV-OPS`. Specifically, the algorithm projects the unconstrained occupancy measure $\tilde{q}_{t+1}$ onto an "optimistic" set by solving Problem (4), which, if the problem is feasible, results in occupancy measure $\widehat{q}_{t+1}$. However, differently from `BV-OPS`, when the problem is feasible, `S-OPS` does *not* select the policy $\pi^{\widehat{q}_{t+1}}$ derived from $\widehat{q}_{t+1}$, but it rather uses a randomization between such a policy and the strictly feasible policy $\pi^\diamond$ (Line 22). The probability λ_t of selecting $\pi^\diamond$ is chosen pessimistically with respect to constraints satisfaction, by using upper confidence bounds for the costs and upper occupancy bounds given the policy $\pi^{\widehat{q}_{t+1}}$ (Lines 17 and 19). Such a pessimistic approach ensures that the constraints are satisfied with high probability, thus making the algorithm safe with high probability. Notice that, if Problem (4) is *not* feasible, then any occupancy measure in $\Delta(\mathcal{P}_t)$ can be selected (Line 21).

Algorithm 3 Safe Optimistic Policy Search

Require: $X, A, \alpha, T, \delta, \eta, \gamma, \pi^\diamond, \beta$

1: **for** $k \in [0 \dots L-1]$, $(x, a, x') \in X_k \times A \times X_{k+1}$ **do**
2: $\quad N_0(x,a) \leftarrow 0$; $M_0(x,a,x') \leftarrow 0$
3: $\quad \widehat{q}_1(x,a,x') \leftarrow \frac{1}{|X_k||A||X_{k+1}|}$
4: $\pi_1 \leftarrow \begin{cases} \pi^\diamond & \text{w. probability } \lambda_0 := \max_{i\in[m]} \left\{\frac{L-\alpha_i}{L-\beta_i}\right\} \\ \pi^{\widehat{q}_1} & \text{w. probability } 1-\lambda_0 \end{cases}$
5: **for** $t \in [T]$ **do**
6: $\quad$ Select π_t in Algorithm 1 and receive feedback
7: $\quad$ Build *upper occupancy bounds* for $k \in [0 \dots L-1]$:

$$u_t(x_k, a_k) \leftarrow \max_{\overline{P}\in\mathcal{P}_{t-1}} q^{\overline{P},\pi_t}(x_k, a_k)$$

8: $\quad$ Build *optimistic loss estimator* for $(x,a) \in X \times A$:

$$\widehat{\ell}_t(x,a) \leftarrow \begin{cases} \frac{\ell_t(x,a)}{u_t(x,a)+\gamma} & \text{if } \mathbb{1}_t\{x,a\} = 1 \\ 0 & \text{otherwise} \end{cases}$$

9: $\quad$ **for** $k \in [0 \dots L-1]$ **do**
10: $\quad\quad N_t(x_k, a_k) \leftarrow N_{t-1}(x_k, a_k) + 1$
11: $\quad\quad M_t(x_k, a_k, x_{k+1}) \leftarrow M_{t-1}(x_k, a_k, x_{k+1}) + 1$
12: $\quad$ Build $\mathcal{P}_t$, $\widehat{G}_t$, and Ξ_t as in Section 3
13: $\quad$ Build *unconstrained occupancy* for all (x,a,x'):

$$\tilde{q}_{t+1}(x,a,x') \leftarrow \widehat{q}_t(x,a,x') e^{-\eta \widehat{\ell}_t(x,a)}$$

14: $\quad$ **if** PROJ$\big(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\big)$ is *feasible* **then**
15: $\quad\quad \widehat{q}_{t+1} \leftarrow \text{PROJ}\big(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\big)$
16: $\quad\quad \widehat{\pi}_{t+1} \leftarrow \pi^{\widehat{q}_{t+1}}$
17: $\quad\quad$ Build $\widehat{u}_{t+1} \in [0,1]^{|X\times A|}$ so that for all (x,a):

$$\widehat{u}_{t+1}(x,a) \leftarrow \max_{\overline{P}\in\mathcal{P}_t} q^{\overline{P},\widehat{\pi}_{t+1}}(x,a)$$

18: $\quad\quad \sigma \leftarrow \max_{i\in[m]} \left\{ \frac{\min\left\{(\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1}, L\right\} - \alpha_i}{\min\left\{(\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1}, L\right\} - \beta_i} \right\}$
19: $\quad\quad \lambda_t \leftarrow \begin{cases} \sigma & \text{if } \exists i \in [m] : (\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1} > \alpha_i \\ 0 & \text{if } \forall i \in [m] : (\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1} \leq \alpha_i \end{cases}$
20: $\quad$ **else**
21: $\quad\quad \widehat{q}_{t+1} \leftarrow$ take any $q \in \Delta(\mathcal{P}_t)$; $\lambda_t \leftarrow 1$
22: $\quad \pi_{t+1} \leftarrow \begin{cases} \pi^\diamond & \text{with probability } \lambda_t \\ \pi^{\widehat{q}_{t+1}} & \text{with probability } 1-\lambda_t \end{cases}$

5.2 Safety property

In the following, we show that `S-OPS` enjoys the safety property with high probability. Formally:

Theorem 5.1. *Given a confidence $\delta \in (0,1)$, Algorithm 3 is safe with probability at least $1 - 5\delta$.*

Intuitively, Theorem 5.1 follows from the way in which the randomization probability λ_t is defined. Indeed, λ_t relies on two crucial components: (i) a pessimistic estimate of the costs for state-action pairs, namely, the upper confidence bounds $\widehat{g}_{t,i} + \xi_t$, and (ii) a pessimistic choice of transition probabilities, encoded by the upper occupancy bounds defined by the vector $\widehat{u}_t$. Notice that the $\max_{i\in[m]}$ operator allows to be conservative with respect to all the constraints.

5.3 Cumulative regret

Proving that S-OPS attains sublinear regret begets challenges that, to the best of our knowledge, have never been addressed in the online learning literature. In particular, analyzing the estimates of the adversarial losses requires non-standard techniques in our setting, since the policy π_t that is used by the algorithm and determines the received feedback is *not* the one resulting from an OMD-like update, as it is obtained via a non-standard randomization procedure. Nevertheless, the particular shape of the randomization probability λ_t can be exploited to overcome such a challenge. Indeed, we show that each λ_t can be upper bounded by the initial value λ_0, and, thus, a loss estimator from feedback received by using a policy computed by an OMD-like update is available with probability at least $1 - \lambda_0$. This observation is crucial in order ot prove the following result:

Theorem 5.2. *Given $\delta \in (0, 1)$, by setting $\eta = \gamma = \sqrt{L \ln(L|X||A|/\delta) / T|X||A|}$ in Algorithm 3, the algorithm attains regret $R_T \leq \mathcal{O}\left(\Psi L^3 |X| \sqrt{|A| T \ln\left(T|X||A|m/\delta\right)}\right)$ with probability at least $1 - 11\delta$, where $\Psi := \max_{i \in [m]} \{1/\min\{(\alpha_i - \beta_i), (\alpha_i - \beta_i)^2\}\}$.*

The regret bound in Theorem 5.2 is in line with the one achieved by BV-OPS in the bounded violation setting, with an additional ΨL^2 factor. Such a factor comes from the mismatch between loss estimators and the occupancy measure chosen by the OMD-like update. Notice that Ψ depends on the violation gap $\min_{i \in [m]} \{\alpha_i - \beta_i\}$, which represents how much the strictly feasible solution satisfies the constraints. Such a dependence is expected, since the better the strictly feasible solution (in terms of constraints satisfaction), the larger the exploration performed during the first episodes.

References

E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/e4a6222cdb5b34375400904f03d8e6a5-Paper.pdf.

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

Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Papini, Alberto Maria Metelli, and Nicola Gatti. Online adversarial mdps with off-policy feedback and known transitions. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Markov persuasion processes: Learning to persuade from scratch. *arXiv preprint arXiv:2402.03077*, 2024.

Qinbo Bai, Vaneet Aggarwal, and Ather Gattami. Provably sample-efficient model-free algorithm for mdps with peak constraints. *Journal of Machine Learning Research*, 24(60):1–25, 2023.

Martino Bernasconi, Federico Cacciamani, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, and Francesco Trovò. Safe learning in tree-form sequential decision making: Handling hard and soft constraints. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1854–1873. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bernasconi22a.html.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Kun Chen, Kechao Cai, Longbo Huang, and John CS Lui. Beyond the click-through rate: web link selection with multi-level feedback. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3308–3314, 2018.

Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7396–7404, 2023.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps, 2020. URL https://arxiv.org/abs/2003.02189.

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

Hengquan Guo, Xin Liu, Honghao Wei, and Lei Ying. Online convex optimization with hard constraints: Towards the best of two worlds and beyond. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36426–36439. Curran Associates, Inc., 2022.

Yue He, Xiujun Chen, Di Wu, Junwei Pan, Qing Tan, Chuan Yu, Jian Xu, and Xiaoqiang Zhu. A unified solution to constrained bidding in online display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2993–3001, 2021.

David Isele, Alireza Nakhaei, and Kikuo Fujimura. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20c.html.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

Davide Maran, Pierriccardo Olivieri, Francesco Emanuele Stradi, Giuseppe Urso, Nicola Gatti, and Marcello Restelli. Online markov decision processes configuration with continuous decision space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14315–14322, 2024.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

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

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019. URL http://arxiv.org/abs/1912.13213.

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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15277–15287. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ae95296e27d7f695f891cd26b4f37078-Paper.pdf.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/a0872cc5b5ca4cc25076f3d868e1bdf8-Paper.pdf.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/rosenberg19a.html.

Ming Shi, Yingbin Liang, and Ness Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. *arXiv preprint arXiv:2302.04375*, 2023.

Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, E Chi, Jilin Chen, and Alex Beutel. Building healthy recommendation sequences for everyone: A safe reinforcement learning approach. In *Proceedings of the FAccTRec Workshop, Online*, pages 26–27, 2020.

Francesco Emanuele Stradi, Filippo Cipriani, Lorenzo Ciampiconi, Marco Leonardi, Alessandro Rozza, and Nicola Gatti. A primal-dual online learning approach for dynamic pricing of sequentially displayed complementary items under sale constraints. *arXiv preprint arXiv:2407.05793*, 2024a.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs: Handling stochastic and adversarial constraints. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46692–46721. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/stradi24a.html.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning constrained markov decision processes with non-stationary rewards and constraints. *arXiv preprint arXiv:2405.14372*, 2024c.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and Xingyu Zhou. Provably efficient model-free algorithms for non-stationary cmdps. In *International Conference on Artificial Intelligence and Statistics*, pages 6527–6570. PMLR, 2023.

Xiaohan Wei, Hao Yu, and Michael J. Neely. Online learning in weakly coupled markov decision processes: A convergence time study. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(1), apr 2018. doi: 10.1145/3179415. URL https://doi.org/10.1145/3179415.

Lu Wen, Jingliang Duan, Shengbo Eben Li, Shaobing Xu, and Huei Peng. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020.

Di Wu, Xiujun Chen, Xun Yang, Hao Wang, Qing Tan, Xiaoxun Zhang, Jian Xu, and Kun Gai. Budget constrained bidding by model-free reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1443–1451, 2018.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 620–629. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/zheng20a.html.

Appendix

The appendix is organized as follows:

- In Appendix A we provide the omitted proofs related to the analysis of the clean event.
- In Appendix B we provide the omitted proofs related to the performances attained by Algorithm 2, namely, the one which guarantees bounded violation.
- In Appendix C we provide the omitted proofs related to the performances attained by Algorithm 3, namely, the one which guarantees the safety property.
- In Appendix D we provide useful lemmas from existing works.

A Omitted proofs for the clean event

In this section, we report the omitted proof related to the clean event. We start stating the following preliminary result.

Lemma A.1. *Given any $\delta \in (0,1)$, fix $i \in [m]$, $t \in [T]$ and $(x,a) \in X \times A$, it holds, with probability at least $1-\delta$:*

$$\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \zeta_t(x,a),$$

where $\zeta_t(x,a) := \sqrt{\frac{\ln(2/\delta)}{2N_t(x,a)}}$ and $\overline{g}_{t,i}(x,a)$ is the true mean value of the distribution.

Proof. Focus on specifics $i \in [m]$, $t \in [T]$ and $(x,a) \in X \times A$. By Hoeffding's inequality and noticing that constraints values are bounded in $[0,1]$, it holds that:

$$\mathbb{P}\left[\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \geq \frac{c}{N_t(x,a)}\right] \leq 2\exp\left(-\frac{2c^2}{N_t(x,a)}\right)$$

Setting $\delta = 2\exp\left(-\frac{2c^2}{N_t(x,a)}\right)$ and solving to find a proper value of c concludes the proof. □

Now we generalize the previous result in order to hold for every $i \in [m]$, $t \in [T]$ and $(x,a) \in X \times A$ at the same time.

Lemma 3.1. *Given a confidence parameter $\delta \in (0,1)$, with probability at least $1-\delta$, for every $i \in [m]$, episode $t \in [T]$, and pair $(x,a) \in X \times A$, it holds $|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)| \leq \xi_t(x,a)$, where we let the confidence bound $\xi_t(x,a) := \min\{1, \sqrt{4\ln(T|X||A|m/\delta)/\max\{1, N_t(x,a)\}}\}$.*

Proof. From Lemma 3.2, given $\delta' \in (0,1)$, we have for any $i \in [m]$, $t \in [T]$ and $(x,a) \in X \times A$:

$$\mathbb{P}\left[\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \zeta_t(x,a)\right] \geq 1-\delta'.$$

Now, we are interested in the intersection of all the events, namely,

$$\mathbb{P}\left[\bigcap_{x,a,m,t}\left\{\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \zeta_t(x,a)\right\}\right].$$

Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,m,t}\left\{\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \zeta_t(x,a)\right\}\right] \\
&\qquad = 1 - \mathbb{P}\left[\bigcup_{x,a,m,t}\left\{\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \zeta_t(x,a)\right\}^c\right] \\
&\qquad \geq 1 - \sum_{x,a,m,t}\mathbb{P}\left[\left\{\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \zeta_t(x,a)\right\}^c\right] \qquad (5)
\end{aligned}$$

$$\geq 1 - |X||A|mT\delta',$$

where Inequality (5) holds by Union Bound. Noticing that $g_{t,i}(x,a) \leq 1$, substituting δ' with $\delta := \delta'/|X||A|mT$ in $\zeta_t(x,a)$ with an additional Union Bound over the possible values of $N_t(x,a)$, and thus obtaining $\xi_t(x,a)$, concludes the proof. □

B Omitted proofs when Condition 2.5 does not hold

In this section we report the omitted proofs of the theoretical results for Algorithm 2.

B.1 Feasibility

We start by showing that Program (4) admits a feasible solution with arbitrarily large probability.

Lemma 4.1. *Given confidence $\delta \in (0,1)$, Algorithm 2 ensures that* PROJ$(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t)$ *is feasible at every episode $t \in [T]$ with probability at least $1 - 5\delta$.*

Proof. To prove the lemma we show that under the event $\mathcal{E}^{G,\Delta}(\delta)$, which holds the probability at least $1 - 5\delta$, Program (4) admits a feasible solution. Precisely, under the event $\mathcal{E}^{\Delta}(\delta)$, the true transition function P belongs to $\mathcal{P}_t$ at each episode. Moreover, under the event $\mathcal{E}^{G}(\delta)$, we have, for any feasible solution $q^{\square}$ of the offline optimization problem, for any $t \in [T]$,

$$\left(\widehat{G}_t - \Xi_t\right)^{\top} q^{\square} \preceq \overline{G}_t^{\top} q^{\square} \preceq \alpha,$$

where the first inequality holds by the definition of the event. The previous inequality shows that if $q^{\square}$ satisfies the constraints with respect to the true mean constraint matrix, it satisfies also the optimistic constraints. Thus, the feasible solutions to the offline problem are all available at every episode. Noticing that the clean event is defined as the intersection between $\mathcal{E}^{G}(\delta)$ and $\mathcal{E}^{\Delta}(\delta)$ concludes the proof. □

B.2 Violations

We proceed bounding the cumulative positive violation as follows.

Theorem 4.2. *Given $\delta \in (0,1)$, Algorithm 2 attains cumulative positive constraints violation $V_T \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(T|X||A|m/\delta\right)}\right)$ with probability at least $1 - 8\delta$.*

Proof. The key point of the problem is to relate the constraints satisfaction with the convergence rate of both the confidence bound on the constraints and the transitions.

First, we notice that under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, all the following reasoning hold for every constraint $i \in [m]$. Thus, we focus on the bound of a single constraint violation problem defined as follows:

$$V_T := \sum_{t=1}^{T} \left[\overline{g}^{\top} q_t - \alpha\right]^{+}$$

By Lemma 4.1, under the clean event the $\mathcal{E}^{G,\Delta}(\delta)$, the convex program is feasible and it holds:

$$\overline{g} - 2\xi_t \preceq \widehat{g}_t - \xi_t$$

Thus, multiplying for the estimated occupancy measure and by construction of the convex program we obtain:

$$\left(\overline{g} - 2\xi_{t-1}\right)^{\top} \widehat{q}_t \leq \left(\widehat{g}_{t-1} - \xi_{t-1}\right)^{\top} \widehat{q}_t \leq \alpha.$$

Rearranging the equation, it holds:

$$\overline{g}^{\top} \widehat{q}_t \leq \alpha + 2\xi_{t-1}^{\top} \widehat{q}_t.$$

Now, in order to obtain the instantaneous violation definition we proceed as follows,

$$\overline{g}^{\top} \widehat{q}_t + \overline{g}^{\top} q_t - \overline{g}^{\top} q_t \leq \alpha + 2\xi_{t-1}^{\top} \widehat{q}_t,$$

from which we obtain:

$$\begin{aligned}\overline{g}^\top q_t - \alpha &\leq \overline{g}^\top (q_t - \widehat{q}_t) + 2\xi_{t-1}^\top \widehat{q}_t \\ &\leq \|\overline{g}\|_\infty \|q_t - \widehat{q}_t\|_1 + 2\xi_{t-1}^\top \widehat{q}_t,\end{aligned}$$

where the last step holds by the Hölder inequality. Notice that, since the RHS of the previous inequality is greater than zero, it holds,

$$[\overline{g}^\top q_t - \alpha]^+ \leq \|q_t - \widehat{q}_t\|_1 + 2\xi_{t-1}^\top \widehat{q}_t.$$

which leads to $V_T \leq \sum_{t=1}^T \|q_t - \widehat{q}_t\|_1 + 2\sum_{t=1}^T \xi_{t-1}^\top \widehat{q}_t$, where the first part of the equation refers to the estimate of the transitions while the second one to the estimate of the constraints. We will bound the two terms separately.

Bound on $\sum_{t=1}^T \|\widehat{q}_t - q_t\|_1$. The term of interest encodes the distance between the estimated occupancy measure and the real one chosen by the algorithm. Thus, it depends on the estimation of the true transition functions. To bound the quantity of interest, we proceed as follows:

$$\begin{aligned}\sum_{t=1}^T \|\widehat{q}_t - q_t\|_1 &= \sum_{t=1}^T \sum_{x,a} |\widehat{q}_t(x,a) - q_t(x,a)| \\ &\leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \qquad (6)\end{aligned}$$

where Inequality (6) holds since, by Lemma D.1, under the clean event, with probability at least $1-2\delta$, we have $\sum_{t=1}^T \sum_{x,a} |\widehat{q}_t(x,a) - q_t(x,a))| \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right)$, when $\widehat{q}_t \in \Delta(\mathcal{P}_t)$. Please notice that the condition $\widehat{q}_t \in \Delta(\mathcal{P}_t)$ is verified since the constrained space defined by Program (4) is contained in $\Delta(\mathcal{P}_t)$.

Bound on $\sum_{t=1}^T \xi_{t-1}^\top \widehat{q}_t$. This term encodes the estimation of the constraints functions obtained following the estimated occupancy measure. Nevertheless, since the confidence bounds converge only for the paths traversed by the learner, it is necessary to relate ξ_t to the real occupancy measure chosen by the algorithm. To do so, we notice that by Hölder inequality and since $\xi_t(x,a) \leq 1$, it holds:

$$\begin{aligned}\sum_{t=1}^T \xi_{t-1}^\top \widehat{q}_t &\leq \sum_{t=1}^T \xi_{t-1}^\top q_t + \sum_{t=1}^T \xi_{t-1}^\top (\widehat{q}_t - q_t) \\ &\leq \sum_{t=1}^T \xi_{t-1}^\top q_t + \sum_{t=1}^T \|\xi_{t-1}\|_\infty \|\widehat{q}_t - q_t\|_1 \\ &\leq \sum_{t=1}^T \xi_{t-1}^\top q_t + \sum_{t=1}^T \|\widehat{q}_t - q_t\|_1.\end{aligned}$$

The second term of the inequality is bounded by the previous analysis, while for the first term we proceed as follows:

$$\begin{aligned}\sum_{t=1}^T \xi_{t-1}^\top q_t &= \sum_{t=1}^T \sum_{x,a} \xi_{t-1}(x,a) q_t(x,a) \\ &\leq \sum_{t=1}^T \sum_{x,a} \xi_{t-1}(x,a) \mathbb{1}_t\{x,a\} + L\sqrt{2T\ln\frac{1}{\delta}} \qquad (7) \\ &= \sqrt{4\ln\left(\frac{T|X||A|m}{\delta}\right)} \sum_{t=1}^T \sum_{x,a} \sqrt{\frac{1}{\max\{1, N_{t-1}(x,a)\}}} \mathbb{1}_t\{x,a\} + L\sqrt{2T\ln\frac{1}{\delta}}\end{aligned}$$

$$\leq 3\sqrt{4\ln\left(\frac{T|X||A|m}{\delta}\right)}\sum_{x,a}\sqrt{N_T(x,a)} + L\sqrt{2T\ln\frac{1}{\delta}} \tag{8}$$

$$\leq 6\sqrt{L|X||A|T\ln\left(\frac{T|X||A|m}{\delta}\right)} + L\sqrt{2T\ln\frac{1}{\delta}}, \tag{9}$$

where Inequality (7) follows from Azuma inequality and noticing that $\sum_{x,a}\xi_{t-1}(x,a)q_t(x,a) \leq L$ (with probability at least $1-\delta$), Inequality (8) holds since $1+\sum_{t=1}^{T}\frac{1}{t} \leq 2\sqrt{T}+1 \leq 3\sqrt{T}$ and Inequality (9) follows from Cauchy-Schwarz inequality and noticing that $\sqrt{\sum_{x,a}N_T(x,a)} \leq \sqrt{LT}$.

We combine the previous bounds as follows:

$$\begin{aligned} V_T &\leq \sum_{t=1}^{T}\|q_t - \widehat{q}_t\|_1 + 2\sum_{t=1}^{T}\xi_{t-1}^{\top}\widehat{q}_t \\ &\leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|m}{\delta}\right)}\right). \end{aligned}$$

The results holds with probability at least at least $1-8\delta$ by union bound over the clean event, Lemma D.1 and the Azuma-Hoeffding inequality. This concludes the proof. □

B.3 Regret

In this section, we prove the regret bound of Algorithm 2. Precisely, the bound follows from noticing that, under the clean event, the optimal safe solution is included in the decision space for every episode $t \in [T]$.

Theorem 4.3. *Given $\delta \in (0,1)$, by setting $\eta = \gamma = \sqrt{L\ln(L|X||A|/\delta)/T|X||A|}$ in Algorithm 2, the algorithm attains regret $R_T \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(T|X||A|/\delta\right)}\right)$ with probability at least $1-10\delta$.*

Proof. We first rewrite the regret definition as follows:

$$\begin{aligned} R_T &= \sum_{t=1}^{T}\ell_t^{\top}q_t - \sum_{t=1}^{T}\ell_t^{\top}q^* \\ &= \underbrace{\sum_{t=1}^{T}\ell_t^{\top}(q_t-\widehat{q}_t)}_{①} + \underbrace{\sum_{t=1}^{T}\widehat{\ell}_t^{\top}(\widehat{q}_t-q^*)}_{②} + \underbrace{\sum_{t=1}^{T}(\ell_t-\widehat{\ell}_t)^{\top}\widehat{q}_t}_{③} + \underbrace{\sum_{t=1}^{T}(\widehat{\ell}_t-\ell_t)^{\top}q^*}_{④}. \end{aligned}$$

Precisely, the first term encompasses the distance between the true transitions and the estimated ones, the second concerns the optimization performed by online mirror descent and the last ones encompass the bias of the estimators.

Bound on ①. We start bounding the first term, namely, the cumulative distance between the estimated occupancy measure and the real one, as follows:

$$\begin{aligned} ① &= \sum_{t=1}^{T}\ell_t^{\top}(q_t-\widehat{q}_t) \\ &= \sum_{t=1}^{T}\sum_{x,a}\ell_t(x,a)(q_t(x,a)-\widehat{q}_t(x,a)) \\ &\leq \sum_{t=1}^{T}\sum_{x,a}|(q_t(x,a)-\widehat{q}_t(x,a)|, \end{aligned} \tag{10}$$

where the Inequality (10) holds by Hölder inequality noticing that $\|\ell_t\|_\infty \leq 1$ for all $t \in [T]$. Then, noticing that the projection of Algorithm 2 is performed over a subset of $\Delta(\mathcal{P}_t)$ and employing Lemma D.1, we obtain:

$$① \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \tag{11}$$

with probability at least $1-2\delta$, under the clean event.

Bound on ②. To bound the second term, we underline that, under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, the estimated safe occupancy $\widehat{q}_t$ belongs to $\Delta(\mathcal{P}_t)$ and the optimal safe solution q^* is included in the constrained decision space for each $t \in [T]$. Moreover we notice that, for each $t \in [T]$, the constrained space is convex and linear, by construction of Program (4). Thus, following the standard analysis of online mirror descent Orabona [2019] and from Lemma D.5, we have, under the clean event:

$$② \leq \frac{L\ln\left(|X|^2|A|\right)}{\eta} + \eta\sum_{t,x,a}\widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2.$$

Thus, to bound the biased estimator, we notice that $\widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2 \leq \frac{\widehat{q}_t(x,a)}{u_t(x,a)+\gamma}\widehat{\ell}_t(x,a) \leq \widehat{\ell}_t(x,a)$. We then apply Lemma D.2 with $\alpha_t(x,a) = 2\gamma$ and obtain $\sum_{t,x,a}\widehat{q}_t(x,a)\widehat{\ell}_t(x,a)^2 \leq \sum_{t,x,a}\frac{q_t(x,a)}{u_t(x,a)}\ell_t(x,a) + \frac{L\ln\frac{L}{\delta}}{2\gamma}$. Finally, we notice that, under the clean event, $q_t(x,a) \leq u_t(x,a)$, obtaining, with probability at least $1-\delta$:

$$② \leq \frac{L\ln\left(|X|^2|A|\right)}{\eta} + \eta|X||A|T + \frac{\eta L\ln(L/\delta)}{2\gamma}.$$

Setting $\eta = \gamma = \sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$, we obtain:

$$② \leq \mathcal{O}\left(L\sqrt{|X||A|T\ln\left(\frac{|X||A|}{\delta}\right)}\right), \tag{12}$$

with probability at least $1-\delta$, under the clean event.

Bound on ③. The third term follows from Lemma D.4, from which, under the clean event, with probability at least $1-3\delta$ and setting $\gamma = \sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$, we obtain:

$$③ \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{13}$$

Bound on ④. We bound the fourth term employing Corollary D.3 and obtaining,

$$\begin{aligned}
\sum_{t=1}^{T}\left(\widehat{\ell}_t - \ell_t\right)^\top q^* &= \sum_{t,x,a} q^*(x,a)\left(\widehat{\ell}_t(x,a) - \ell_t(x,a)\right) \\
&\leq \sum_{t,x,a} q^*(x,a)\ell_t(x,a)\left(\frac{q_t(x,a)}{u_t(x,a)} - 1\right) + \sum_{x,a}\frac{q^*(x,a)\ln\frac{|X||A|}{\delta}}{2\gamma} \\
&= \sum_{t,x,a} q^*(x,a)\ell_t(x,a)\left(\frac{q_t(x,a)}{u_t(x,a)} - 1\right) + \frac{L\ln\frac{|X||A|}{\delta}}{2\gamma}.
\end{aligned}$$

Noticing that, under the clean event, $q_t(x,a) \leq u_t(x,a)$ and setting $\gamma = \sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$, we obtain, with probability at least $1-\delta$:

$$④ \leq \mathcal{O}\left(L\sqrt{|X||A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{14}$$

Final result. Finally, combining Equation (11), Equation (12), Equation (13) and Equation (14) and applying a union bound, we obtain, with probability at least $1 - 10\delta$,

$$R_T \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right).$$

□

C Omitted proofs when Condition 2.5 holds

In this section we report the omitted proofs of the theoretical results for Algorithm 3.

C.1 Safety

We start by showing that Algorithm 3 is safe with high probability.

Theorem 5.1. *Given a confidence $\delta \in (0,1)$, Algorithm 3 is safe with probability at least $1 - 5\delta$.*

Proof. We show that, under event $\mathcal{E}^{G,\Delta}(\delta)$, the *non-Markovian* policy defined by the probability λ_t satisfies the constraints. Intuitively, the result follows from the construction of the convex combination parameter λ_t. Indeed, λ_t is built using a pessimist estimated of the constraints cost, namely, $\widehat{g}_{t,i} + \xi_t$. Moreover, the upper occupancy bound $\widehat{u}_t$ introduces pessimism in the choice of the transition function. Finally, the $\max_{i\in[m]}$ operator allows to be conservative for all the m constraints.

We split the analysis in the three possible cases defined by λ_t, namely, $\lambda_t = 0$ and $\lambda_t \in (0,1)$. Please notice that $\lambda_t < 1$, by construction.

Analysis when $\lambda_t = 0$. When $\lambda_t = 0$, it holds, by construction, that $\forall i \in [m] : (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{u}_t \leq \alpha_i$. Thus, under the event $\mathcal{E}^{G,\Delta}(\delta)$, it holds, $\forall i \in [m]$:

$$\begin{aligned}
\alpha_i &\geq (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{u}_t \\
&\geq (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{q}_t && (15)\\
&= (\widehat{g}_{t-1,i} + \xi_{t-1})^\top q_t \\
&\geq \overline{g}_i^\top q_t, && (16)
\end{aligned}$$

where Inequality (15) holds by definition of $\widehat{u}_t$ and Inequality (16) by the pessimistic definition of the constraints.

Analysis when $\lambda_t \in (0,1)$. We focus on a single constraint $i \in [m]$, then we generalize the analysis for the entire set of constraints. First we notice that the constraints cost, for a single constraint $i \in [m]$, attained by the *non-Markovian* policy π_t, is equal to $\lambda_{t-1}\overline{g}_i^\top q^\diamond + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t}$. Thus, it holds by definition of the known strictly feasible $\pi^\diamond$,

$$\lambda_{t-1}\overline{g}_i^\top q^\diamond + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t} = \lambda_{t-1}\beta_i + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t}. \tag{17}$$

Then, we consider both the cases when $L < (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{u}_t$ (first case) and $L > (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{u}_t$ (second case). If the two quantities are equivalent, the proof still holds breaking the ties arbitrarily.

First case. It holds that:

$$\begin{aligned}
\lambda_{t-1}\beta_i + (1-\lambda_{t-1})\overline{g}_i^\top q^{\widehat{\pi}_t,P} &\leq \lambda_{t-1}\beta_i + (1-\lambda_{t-1})L && (18)\\
&= \frac{L-\alpha_i}{L-\beta_i}(\beta_i - L) + L \\
&= \frac{\alpha_i - L}{\beta_i - L}(\beta_i - L) + L \\
&= \alpha_i,
\end{aligned}$$

where Inequality (18) holds by definition of the constraints.

Second case. It holds that:

$$\begin{aligned}
&\lambda_{t-1}\beta_i+(1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t} \\
&\qquad \leq \lambda_{t-1}\beta_i + (1-\lambda_{t-1})\left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top q^{P,\widehat{\pi}_t} && (19)\\
&\qquad \leq \lambda_{t-1}\beta_i + (1-\lambda_{t-1})\left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t && (20)\\
&\qquad = \lambda_{t-1}\beta_i - \lambda_{t-1}\left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t + \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t \\
&\qquad = \lambda_{t-1}(\beta_i - \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t) + \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t \\
&\qquad \leq \frac{\left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t - \alpha_i}{\left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t - \beta_i}(\beta_i - \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t) + \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t \\
&\qquad = \frac{\alpha_i - \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t}{\beta_i - \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t}(\beta_i - \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t) + \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t \\
&\qquad = \alpha_i - \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t + \left(\widehat{g}_{t-1,i}+\xi_{t-1}\right)^\top \widehat{u}_t \\
&\qquad = \alpha_i,
\end{aligned}$$

where Inequality (19) holds by the definition of the event and Inequality (20) holds by the definition of $\widehat{u}_t$.

To conclude the proof, we underline that λ_t is chosen taking the maximum over the constraints, which implies that the more conservative λ_t (the one which takes the combination nearer to the strictly feasible solution) is chosen. Thus, all the constraints are satisfied. □

C.2 Regret

We start by the statement of the following Lemma, which is a generalization of the results from Jin et al. [2020]. Intuitively, the following result states that the distance between the estimated *non-safe* occupancy measure $\widehat{q}_t$ and the real one reduces as the number of episodes increases, paying a $1-\lambda_t$ factor. This is reasonable since, from the update of the *non-Markovian* policy π_t (see Algorithm 3), policy $\widehat{\pi}_t \leftarrow \widehat{q}_t$ is played with probability $1-\lambda_{t-1}$.

Lemma C.1. *Under the clean event, with probability at least $1-2\delta$, for any collection of transition functions $\{P_t^x\}_{x\in X}$ such that $P_t^x \in \mathcal{P}_t$, and for any collection of $\{\lambda_t\}_{t=0}^{T-1}$ used to select policy π_{t+1}, we have, for all x,*

$$\sum_{t=1}^{T}(1-\lambda_{t-1})\sum_{x\in X,a\in A}\left|q^{P_t^x,\widehat{\pi}_t}(x,a)-q^{P,\widehat{\pi}_t}(x,a)\right| \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right).$$

Proof. We will refer as q_t^x to $q^{P_t^x,\pi_t}$ and as $\widehat{q}_t{}^x$ to $q^{P_t^x,\widehat{\pi}_t}$. Moreover, we define:

$$\epsilon_t^*(x'|x,a) = \sqrt{\frac{P\left(x'|x,a\right)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1,N_t(x,a)\right\}}} + \frac{\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1,N_t(x,a)\right\}}.$$

Now following standard analysis by Lemma D.1 from Jin et al. [2020], we have that,

$$\begin{aligned}
&\sum_{t=1}^{T}(1-\lambda_{t-1})\sum_{x\in X,a\in A}\left|q^{P_t^x,\widehat{\pi}_t}(x,a)-q^{P,\widehat{\pi}_t}(x,a)\right| \leq \\
&\qquad \sum_{0\leq m<k<L}\sum_{t,w_m}(1-\lambda_{t-1})\epsilon_t^*(x_{m+1}|x_m,a_m)q^{P,\widehat{\pi}_t}(x_m,a_m) + |X|\sum_{0\leq m<h<L}\sum_{t,w_m,w'_h}(1-\lambda_{t-1})\cdot \\
&\qquad\qquad \cdot\epsilon_{i_t}^*\left(x_{m+1}\mid x_m,a_m\right)q^{P,\widehat{\pi}_t}\left(x_m,a_m\right)\epsilon_t^*\left(x'_{h+1}\mid x'_h,a'_h\right)q^{P,\widehat{\pi}_t}\left(x'_h,a'_h\mid x_{m+1}\right),
\end{aligned}$$

where $w_m = (x_m,a_m,x_{m+1})$.

Bound on the first term. To bound the first term we notice that, by definition of $q^{P,\widehat{\pi}_t}$ it holds:

$$\begin{aligned}
&\sum_{0\le m<k<L}\sum_{t,w_m}(1-\lambda_{t-1})\epsilon_t^*(x_{m+1}|x_m,a_m)q^{P,\widehat{\pi}_t}(x_m,a_m)\\
&\qquad=\sum_{0\le m<k<L}\sum_{t,w_m}\epsilon_t^*(x_{m+1}|x_m,a_m)\left(q^{P,\pi_t}(x_m,a_m)-\lambda_{t-1}q^{P,\pi^\diamond}(x_m,a_m)\right)\\
&\qquad\le\sum_{0\le m<k<L}\sum_{t,w_m}\epsilon_t^*(x_{m+1}|x_m,a_m)q^{P,\pi_t}(x_m,a_m)\\
&\qquad\le\mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right),
\end{aligned}$$

where the last step holds following Lemma D.1 from Jin et al. [2020].

Bound on the second term. Following Lemma D.1 from Jin et al. [2020], the second term is bounded by (ignoring constants),

$$\begin{aligned}
&\sum_{0\le m<h<L}\sum_{t,w_m,w'_h}(1-\lambda_{t-1})\sqrt{\frac{P\left(x_{m+1}\mid x_m,a_m\right)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1,N_t\left(x_m,a_m\right)\right\}}}\cdot\\
&\qquad\cdot q^{P,\widehat{\pi}_t}\left(x_m,a_m\right)\sqrt{\frac{P\left(x'_{h+1}\mid x'_h,a'_h\right)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1,N_t\left(x'_h,a'_h\right)\right\}}}q^{P,\widehat{\pi}_t}\left(x'_h,a'_h\mid x_{m+1}\right)\\
&+\sum_{0\le m<h<L}\sum_{t,w_m,w'_h}(1-\lambda_{t-1})\frac{q^{P,\widehat{\pi}_t}\left(x_m,a_m\right)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1,N_t\left(x_m,a_m\right)\right\}}+\\
&\qquad+\sum_{0\le m<h<L}\sum_{t,w_m,w'_h}(1-\lambda_{t-1})\frac{q^{P,\widehat{\pi}_t}\left(x'_h,a'_h\right)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1,N_t\left(x'_h,a'_h\right)\right\}}.
\end{aligned}$$

The last two terms are bounded logarithmically in T, employing the definition of $q^{P,\widehat{\pi}_t}$ and following Lemma D.1 from Jin et al. [2020], while, similarly, the first term is bounded by:

$$\sum_{0\le m<h<L}\sqrt{|X_{m+1}|\sum_{t,x_m,a_m}\frac{(1-\lambda_{t-1})q^{P,\widehat{\pi}_t}\left(x_m,a_m\right)}{\max\left\{1,N_t\left(x_m,a_m\right)\right\}}}\sqrt{|X_{h+1}|\sum_{t,x'_h,a'_h}\frac{(1-\lambda_{t-1})q^{P,\widehat{\pi}_t}\left(x'_h,a'_h\right)}{\max\left\{1,N_t\left(x'_h,a'_h\right)\right\}}},$$

which is upper bounded by:

$$\sum_{0\le m<h<L}\sqrt{|X_{m+1}|\sum_{t,x_m,a_m}\frac{q_t\left(x_m,a_m\right)}{\max\left\{1,N_t\left(x_m,a_m\right)\right\}}}\sqrt{|X_{h+1}|\sum_{t,x'_h,a'_h}\frac{q_t\left(x'_h,a'_h\right)}{\max\left\{1,N_t\left(x'_h,a'_h\right)\right\}}}.$$

Employing the same argument as Lemma D.1 from Jin et al. [2020] shows that the previous term is bounded logarithmically in T and concludes the proof. □

We are now ready to prove the regret bound attained by Algorithm 3.

Theorem 5.2. *Given $\delta\in(0,1)$, by setting $\eta=\gamma=\sqrt{L\ln(L|X||A|/\delta)/T|X||A|}$ in Algorithm 3, the algorithm attains regret $R_T\le\mathcal{O}\left(\Psi L^3|X|\sqrt{|A|T\ln\left(T|X||A|m/\delta\right)}\right)$ with probability at least $1-11\delta$, where $\Psi:=\max_{i\in[m]}\{1/\min\{(\alpha_i-\beta_i),(\alpha_i-\beta_i)^2\}\}$.*

Proof. We start decomposing the $R_T:=\sum_{t=1}^T\ell_t^\top(q_t-q^*)$ definition as:

$$\underbrace{\sum_{t=1}^T\ell_t^\top\left(q_t-q^{P_t,\pi_t}\right)}_{①}+\underbrace{\sum_{t=1}^T\widehat{\ell}_t^\top\left(q^{P_t,\widehat{\pi}_t}-q^*\right)}_{②}+\underbrace{\sum_{t=1}^T\ell_t^\top\left(q^{P_t,\pi_t}-q^{P_t,\widehat{\pi}_t}\right)}_{③}+$$

$$+\underbrace{\sum_{t=1}^{T}\left(\ell_t-\widehat{\ell}_t\right)^\top q^{P_t,\widehat{\pi}_t}}_{④}+\underbrace{\sum_{t=1}^{T}\left(\widehat{\ell}_t-\ell_t\right)^\top q^*}_{⑤},$$

where P_t is the transition chosen by the algorithm at episode t. Precisely, the first term encompasses the estimation of the transition functions, the second term concerns the optimization performed by the algorithm, the third term encompasses the regret accumulated by performing the convex combination of policies and the last two terms concern the bias of the optimistic estimators.

We proceed bounding the five terms separately.

Bound on ① We bound the first term as follows:

$$\begin{aligned}
① &= \sum_{t=1}^{T}\ell_t^\top\left(q_t-q^{P_t,\pi_t}\right)\\
&= \sum_{t=1}^{T}\sum_{x,a}\ell_t(x,a)\left(q_t(x,a)-q^{P_t,\pi_t}(x,a)\right)\\
&\le \sum_{t=1}^{T}\sum_{x,a}\left|q_t(x,a)-q^{P_t,\pi_t}(x,a)\right|,
\end{aligned}$$

where the last inequality holds by Hölder inequality noticing that $\|\ell_t\|_\infty \le 1$ for all $t \in [T]$. Then we can employ Lemmas D.1, since π_t is the policy that guides the exploration and $P_t \in \mathcal{P}_t$, obtaining:

$$① \le \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \tag{21}$$

with probability at least $1-2\delta$, under the clean event.

Bound on ② The second term is bounded similarly to the second part of Theorem 4.3. Precisely, we notice that under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, the optimal safe solution q^* is included in the constrained decision space for each $t \in [T]$. Moreover we notice that, for each $t \in [T]$, the constrained space is convex and linear, by construction of the convex program. Thus, following the standard analysis of online mirror descent Orabona [2019] and from Lemma D.5, we have, under the clean event:

$$② \le \frac{L\ln\left(|X|^2|A|\right)}{\eta}+\eta\sum_{t,x,a}q^{P_t,\widehat{\pi}_t}(x,a)\widehat{\ell}_t(x,a)^2.$$

Guaranteeing the safety property makes bounding the biased estimator more complex with respect to Theorem 4.3. Thus, noticing that $\lambda_t \le \max_{i\in[m]}\left\{\frac{L-\alpha_i}{L-\beta_i}\right\}$ and by definition of π_t, we proceed as follows:

$$\begin{aligned}
\eta\sum_{t,x,a}q^{P_t,\widehat{\pi}_t}(x,a)\widehat{\ell}_t(x,a)^2 &\le \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\eta\sum_{t,x,a}(1-\lambda_{t-1})q^{P_t,\widehat{\pi}_t}(x,a)\widehat{\ell}_t(x,a)^2\\
&\le \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\eta\sum_{t,x,a}\left(q^{P_t,\pi_t}(x,a)-\lambda_t q^{P_t,\pi^\diamond}(x,a)\right)\widehat{\ell}_t(x,a)^2\\
&\le \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\eta\sum_{t,x,a}q^{P_t,\pi_t}(x,a)\widehat{\ell}_t(x,a)^2
\end{aligned}$$

The previous result is intuitive. Paying an additional $\max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}$ factor allows to relate the loss estimator $\widehat{\ell}_t$ with the policy that guides the exploration, namely, π_t. Thus, following the same steps as Theorem 4.3 we obtain, with probability $1-\delta$, under the clean event:

$$② \le \frac{L\ln\left(|X|^2|A|\right)}{\eta}+\max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\eta|X||A|T+\max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\frac{\eta L\ln(L/\delta)}{2\gamma}.$$

Setting $\eta = \gamma = \sqrt{\frac{L \ln(L|X||A|/\delta)}{T|X||A|}}$, we obtain:

$$② \leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{1}{\alpha_i - \beta_i}\right\} L\sqrt{L|X||A|T\ln\left(\frac{|X|^2|A|}{\delta}\right)}\right), \tag{22}$$

with probability at least $1-\delta$, under the clean event.

Bound on ③ In the following, we show how to rewrite the third term so that the dependence on the convex combination parameter is explicit. Intuitively, the third term is the regret payed to guarantee the safety property. Thus, we rewrite the third term as follows:

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t^\top \left(q^{P_t,\pi_t} - q^{P_t,\widehat{\pi}_t}\right) &= \sum_{t=1}^{T} \ell_t^\top \left(\lambda_{t-1} q^{P_t,\pi^\diamond} + (1-\lambda_{t-1}) q^{P_t,\widehat{\pi}_t} - q^{P_t,\widehat{\pi}_t}\right) \\
&\leq \sum_{t=1}^{T} \lambda_{t-1} \ell_t^\top q^{P_t,\pi^\diamond} \\
&\leq L \sum_{t=1}^{T} \lambda_{t-1}
\end{aligned}$$

where we used that $\ell_t^\top q^{P_t,\pi^\diamond} \leq L$ for any $t \in [T]$. Thus, we proceed bounding $\sum_{t=1}^{T} \lambda_{t-1}$.

We focus on a single episode $t \in [T]$, in which we assume without loss of generality that the i-th constraint is the hardest to satisfy.

Precisely,

$$\begin{aligned}
\lambda_t &= \frac{\min\left\{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1}, L\right\} - \alpha_i}{\min\left\{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1}, L\right\} - \beta_i} \\
&\leq \frac{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} - \alpha_i}{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} - \beta_i} \\
&\leq \frac{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} - \alpha_i}{\alpha_i - \beta_i} \qquad (23) \\
&= \frac{(\widehat{g}_{t,i} - \xi_t)^\top \widehat{u}_{t+1} + 2\xi_t^\top \widehat{u}_{t+1} - \alpha_i}{\alpha_i - \beta_i} \\
&= \frac{(\widehat{g}_{t,i} - \xi_t)^\top \widehat{q}_{t+1} + (\widehat{g}_{t,i} - \xi_t)^\top (\widehat{u}_{t+1} - \widehat{q}_{t+1}) + 2\xi_t^\top \widehat{u}_{t+1} - \alpha_i}{\alpha_i - \beta_i} \\
&\leq \frac{(\widehat{g}_{t,i} - \xi_t)^\top \widehat{q}_{t+1} + \widehat{g}_{t,i}^\top (\widehat{u}_{t+1} - \widehat{q}_{t+1}) + 2\xi_t^\top \widehat{u}_{t+1} - \alpha_i}{\alpha_i - \beta_i} \\
&\leq \frac{\widehat{g}_{t,i}^\top (\widehat{u}_{t+1} - \widehat{q}_{t+1}) + 2\xi_t^\top \widehat{u}_{t+1}}{\alpha_i - \beta_i} \qquad (24) \\
&= \frac{\widehat{g}_{t,i}^\top (\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}) + \widehat{g}_{t,i}^\top (q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}) + 2\xi_t^\top \widehat{u}_{t+1}}{\alpha_i - \beta_i} \\
&\leq \frac{\|\widehat{g}_{t,i}\|_\infty ||\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}||_1 + \|\widehat{g}_{t,i}\|_\infty \|q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}\|_1 + 2\xi_t^\top \widehat{u}_{t+1}}{\alpha_i - \beta_i} \\
&\leq \frac{||\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}||_1 + \|q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}\|_1 + 2\xi_t^\top \widehat{u}_{t+1}}{\alpha_i - \beta_i} \\
&\leq \frac{L(1-\lambda_t)\|\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}\|_1 + L(1-\lambda_t)\|q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}\|_1 + 2L(1-\lambda_t)\xi_t^\top \widehat{u}_{t+1}}{\min\left\{(\alpha_i - \beta_i), (\alpha_i - \beta_i)^2\right\}} \qquad (25)
\end{aligned}$$

where Inequality (23) holds since, for the hardest constraint, when $\lambda_t \neq 0$, $(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} > \alpha_i$, Inequality (24) holds since, under the clean event, $(\widehat{g}_{t,i} - \xi_t)^\top \widehat{q}_{t+1} \leq \alpha_i$ and Inequality (25) holds

since $\lambda_t \leq \frac{L-\alpha_i}{L-\beta_i}$. Intuitively, Inequality (25) shows that, to guarantee the safety property, Algorithm 3 has to pay a factor proportional to the pessimism introduced on the transition and cost functions, plus the constraints satisfaction gap of the strictly feasible solution given as input to the algorithm.

We need to generalize the result summing over t, taking into account that the hardest constraints may vary. Thus, we bound the summation as follows,

$$\sum_{t=1}^{T} \lambda_{t-1} \leq \max_{i \in [m]} \left\{ \frac{2L}{\min\{(\alpha_i - \beta_i), (\alpha_i - \beta_i)^2\}} \right\} \cdot \\ \cdot \sum_{t=1}^{T} \left((1-\lambda_{t-1}) \left(\|\widehat{u}_t - q^{P,\widehat{\pi}_t}\|_1 + \|q^{P,\widehat{\pi}_t} - q^{P_t,\widehat{\pi}_t}\|_1 + \xi_{t-1}^\top \widehat{u}_t \right) \right)$$

The first two terms of the equation are bounded applying Lemma C.1, which holds with probability at least $1-2\delta$, under the clean event, while, to bound $\sum_{t=1}^{T}(1-\lambda_{t-1})\xi_{t-1}^\top \widehat{u}_t$, we proceed as follows:

$$\sum_{t=1}^{T}(1-\lambda_{t-1})\xi_{t-1}^\top \widehat{u}_t = \sum_{t=1}^{T}(1-\lambda_{t-1})\xi_{t-1}^\top q^{P,\widehat{\pi}_t} + \sum_{t=1}^{T}(1-\lambda_{t-1})\xi_{t-1}^\top (\widehat{u}_t - q^{P,\widehat{\pi}_t}),$$

where the second term is bounded employing Hölder inequality and Lemma C.1. Next, we focus on the first term, proceeding as follows,

$$\begin{aligned}
&\sum_{t=1}^{T}(1-\lambda_{t-1})\xi_{t-1}^\top q^{P,\widehat{\pi}_t} \\
&\qquad \leq \sum_{t=1}^{T} \xi_{t-1}^\top q_t && (26)\\
&\qquad \leq \sum_{t=1}^{T}\sum_{x,a} \xi_{t-1}(x,a)\mathbb{1}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} && (27)\\
&\qquad = \sqrt{4\ln\left(\frac{T|X||A|m}{\delta}\right)} \sum_{t=1}^{T}\sum_{x,a} \sqrt{\frac{1}{\max\{1, N_{t-1}(x,a)\}}}\mathbb{1}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} \\
&\qquad \leq 6\sqrt{\ln\left(\frac{T|X||A|m}{\delta}\right)} \sqrt{|X||A|\sum_{x,a} N_T(x,a)} + L\sqrt{2T\ln\frac{1}{\delta}} && (28)\\
&\qquad \leq 6\sqrt{L|X||A|T\ln\left(\frac{T|X||A|m}{\delta}\right)} + L\sqrt{2T\ln\frac{1}{\delta}},
\end{aligned}$$

where Inequality (26) follows from the definition of π_t, Inequality (27) follows from Azuma-Hoeffding inequality and Inequality (28) holds since $1 + \sum_{t=1}^{T}\frac{1}{t} \leq 2\sqrt{T}+1 \leq 3\sqrt{T}$ and Cauchy-Schwarz inequality.

Thus, we obtain,

$$③ \leq \mathcal{O}\left(\max_{i\in[m]} \left\{ \frac{1}{\min\{(\alpha_i-\beta_i),(\alpha_i-\beta_i)^2\}} \right\} L^3|X| \sqrt{|A|T\ln\left(\frac{T|X||A|m}{\delta}\right)} \right), \tag{29}$$

with probability at least $1-3\delta$, under the clean event.

Bound on ④ We first notice that ④ presents an additional challenge with respect to the bounded violation case. Indeed, since $\widehat{\pi}_t$ is not the policy that drives the exploration, $\widehat{\ell}_t$ cannot be directly bounded employing results from the unconstrained adversarial MDPs literature. First, we rewrite the fourth term as follows,

$$\sum_{t=1}^{T}\left(\ell_t - \widehat{\ell}_t\right)^\top q^{P_t,\widehat{\pi}_t} \leq \sum_{t=1}^{T}\left(\mathbb{E}_t[\widehat{\ell}_t] - \widehat{\ell}_t\right)^\top q^{P_t,\widehat{\pi}_t} + \sum_{t=1}^{T}\left(\ell_t - \mathbb{E}_t[\widehat{\ell}_t]\right)^\top q^{P_t,\widehat{\pi}_t},$$

where $\mathbb{E}_t[\cdot]$ is the expectation given the filtration up to time t. To bound the first term we employ the Azuma-Hoeffding inequality noticing that, the martingale difference sequence is bounded by:

$$\begin{aligned}
\widehat{\ell}_t^\top q^{P_t,\widehat{\pi}_t} &\leq \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\widehat{\ell}_t^\top(1-\lambda_t)q^{P_t,\widehat{\pi}_t} \\
&= \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\widehat{\ell}_t^\top\left(q^{P_t,\pi_t}-\lambda_t q^{P_t,\pi^\diamond}\right) \\
&\leq \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\widehat{\ell}_t^\top q^{P_t,\pi_t} \\
&\leq \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}L,
\end{aligned}$$

where the first inequality holds since $\lambda_t \leq \lambda_0$. Thus, the first term is bounded by $\max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}L\sqrt{2T\ln\frac{1}{\delta}}$. To bound the second term, we employ the definition of π_t and the upper-bound to λ_t, proceeding as follows:

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\ell_t-\mathbb{E}_t[\widehat{\ell}_t]\right)^\top q^{P_t,\widehat{\pi}_t} \\
&\quad= \sum_{t,x,a} q^{P_t,\widehat{\pi}_t}(x,a)\ell_t(x,a)\left(1-\frac{\mathbb{E}_t\left[\mathbb{1}_t(x,a)\right]}{u_t(x,a)+\gamma}\right) \\
&\quad= \sum_{t,x,a} q^{P_t,\widehat{\pi}_t}(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right) \\
&\quad\leq \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\sum_{t,x,a}(1-\lambda_t)q^{P_t,\widehat{\pi}_t}(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right) \\
&\quad\leq \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\sum_{t,x,a} q^{P_t,\pi_t}(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right) \\
&\quad= \max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\sum_{t,x,a}\frac{q^{P_t,\pi_t}(x,a)}{u_t(x,a)+\gamma}\left(u_t(x,a)-q_t(x,a)+\gamma\right) \\
&\quad\leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right)+\max_{i\in[m]}\left\{\frac{L}{\alpha_i-\beta_i}\right\}\gamma|X||A|T,
\end{aligned}$$

where the last steps holds by Lemma D.1. Thus, combining the previous equations, we have, with probability at least $1-3\delta$, under the clean event:

$$④ \leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{1}{\alpha_i-\beta_i}\right\}L^2|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right) \tag{30}$$

Bound on ⑤ The last term is bounded as in Theorem 4.3. Thus, setting $\gamma=\sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$, we obtain, with probability at least $1-\delta$, under the clean event:

$$⑤ \leq \mathcal{O}\left(L\sqrt{|X||A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{31}$$

Final result Finally, we combine the bounds on ①, ②, ③, ④ and ⑤. Applying a union bound, we obtain, with probability at least $1-11\delta$,

$$R_T \leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{1}{\min\left\{(\alpha_i-\beta_i),(\alpha_i-\beta_i)^2\right\}}\right\}L^3|X|\sqrt{|A|T\ln\left(\frac{T|X||A|m}{\delta}\right)}\right),$$

which concludes the proof. □

D Auxiliary lemmas from existing works

D.1 Auxiliary lemmas for the transitions estimation

Similarly to Jin et al. [2020], the estimated occupancy measure space $\Delta(\mathcal{P}_t)$ is characterized as follows:

$$\Delta(\mathcal{P}_t) := \begin{cases} \forall k, & \sum_{x\in X_k, a\in A, x'\in X_{k+1}} q(x,a,x') = 1 \\ \forall k, \forall x, & \sum_{a\in A, x'\in X_{k+1}} q(x,a,x') = \sum_{x'\in X_{k-1}, a\in A} q(x',a,x) \\ \forall k, \forall (x,a,x'), & q(x,a,x') \le \left[\widehat{P}_t(x' \mid x,a) + \epsilon_t(x' \mid x,a)\right] \sum_{y\in X_{k+1}} q(x,a,y) \\ & q(x,a,x') \ge \left[\widehat{P}_t(x' \mid x,a) - \epsilon_t(x' \mid x,a)\right] \sum_{y\in X_{k+1}} q(x,a,y) \\ & q(x,a,x') \ge 0 \end{cases}.$$

Given the estimation of the occupancy measure space, it is possible to derive the following lemma.

Lemma D.1. *Jin et al. [2020] With probability at least $1-6\delta$, for any collection of transition functions $\{P_t^x\}_{x\in X}$ such that $P_t^x \in \mathcal{P}_t$, we have, for all x,*

$$\sum_{t=1}^{T} \sum_{x\in X, a\in A} \left| q^{P_t^x,\pi_t}(x,a) - q_t(x,a) \right| \le \mathcal{O}\left(L|X| \sqrt{|A|T \ln\left(\frac{T|X||A|}{\delta}\right)} \right).$$

We underline that the constrained space defined by Program (4) is a subset of $\Delta(\mathcal{P}_t)$. This implies that, in Algorithm 2, it holds $\widehat{q}_t \in \Delta(\mathcal{P}_t)$ and Lemma D.1 is valid.

D.2 Auxiliary lemmas for the optimistic loss estimator

We will make use of the optimistic biased estimator with implicit exploration factor (see, Neu [2015]). Precisely, we define the loss estimator as follows, for all $t \in [T]$:

$$\widehat{\ell}_t(x,a) := \frac{\ell_t(x,a)}{u_t(x,a)+\gamma} \mathbb{1}_t\{x,a\}, \quad \forall (x,a) \in X \times A,$$

where $u_t(x,a) := \max_{\overline{P}\in\mathcal{P}_t} q^{\overline{P},\pi_t}(x,a)$. Thus, the following lemmas holds.

Lemma D.2. *Jin et al. [2020] For any sequence of functions $\alpha_1, \dots, \alpha_T$ such that $\alpha_t \in [0, 2\gamma]^{X\times A}$ is $\mathcal{F}_t$-measurable for all t, we have with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \sum_{x,a} \alpha_t(x,a) \left(\widehat{\ell}_t(x,a) - \frac{q_t(x,a)}{u_t(x,a)} \ell_t(x,a) \right) \le L \ln \frac{L}{\delta}.$$

Following the analysis of Lemma D.2, with $\alpha_t(x,a) = 2\gamma \mathbb{1}_t(x,a)$ and union bound, the following corollary holds.

Corollary D.3. *Jin et al. [2020] With probability at least $1-\delta$:*

$$\sum_{t=1}^{T} \left(\widehat{\ell}_t(x,a) - \frac{q_t(x,a)}{u_t(x,a)} \ell_t(x,a) \right) \le \frac{1}{2\gamma} \ln\left(\frac{|X||A|}{\delta}\right).$$

Furthermore, when $\pi_t \leftarrow \widehat{q}_t$, the following lemma holds.

Lemma D.4. *Jin et al. [2020] With probability at least $1-7\delta$,*

$$\sum_{t=1}^{T} \left(\ell_t - \widehat{\ell}_t\right)^\top \widehat{q}_t \le \mathcal{O}\left(L|X| \sqrt{|A|T \ln\left(\frac{T|X||A|}{\delta}\right)} + \gamma |X||A|T \right).$$

We notice that $\pi_t \leftarrow \widehat{q}_t$ holds only for Algorithm 2, since in Algorithm 3, $\pi_t \leftarrow \widehat{q}_t$ with probability $1-\lambda_{t-1}$.

D.3 Auxiliary lemmas for online mirror descent

We will employ the following results for OMD (see, Orabona [2019]) with uniform initialization over the estimated occupancy measure space.

Lemma D.5. *Jin et al. [2020] The OMD update with* $\widehat{q}_1(x,a,x') = \frac{1}{|X_k||A||X_{k+1}|}$ *for all* $k < L$ *and* $(x,a,x') \in X_k \times A \times X_{k+1}$*, and*

$$\widehat{q}_{t+1} = \arg\min_{q \in \Delta(\mathcal{P}_t)} \widehat{\ell}_t^\top q + \frac{1}{\eta} D\left(q \| \widehat{q}_t\right),$$

where $D(q\|q') = \sum_{x,a,x'} q(x,a,x') \ln \frac{q\left(x,a,x'\right)}{q'(x,a,x')} - \sum_{x,a,x'} \left(q(x,a,x') - q'(x,a,x')\right)$ *ensures*

$$\sum_{t=1}^{T} \widehat{\ell}_t^\top (\widehat{q}_t - q) \leq \frac{L \ln\left(|X|^2|A|\right)}{\eta} + \eta \sum_{t,x,a} \widehat{q}_t(x,a) \widehat{\ell}_t(x,a)^2,$$

for any $q \in \cap_t \Delta(\mathcal{P}_t)$*, as long as* $\widehat{\ell}_t(x,a) \geq 0$ *for all* t, x, a*.*

EWRL Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state all the main contributions made by the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All the assumptions are clearly stated in Section 2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the theoretical results clearly state their assumptions, while all their proofs are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does *not* include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does *not* include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed, since the work is mainly theoretical.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.