Taming Adversarial Constraints in CMDPs

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

Anna Lunghi
Politecnico di Milano
anna.lunghi@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

In *constrained* MDPs (CMDPs) with *adversarial* rewards and constraints, a known impossibility result prevents any algorithm from attaining sublinear regret and constraint violation, when competing against a best-in-hindsight policy that satisfies the constraints on average. In this paper, we show how to ease such a negative result, by considering settings that generalize *both* stochastic CMDPs and adversarial ones. We provide algorithms whose performances smoothly degrade as the level of environment adverseness increases. Specifically, they attain $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *positive* constraint violation under *bandit* feedback, where C measures the adverseness of rewards and constraints. This is $C = \Theta(T)$ in the worst case, coherently with the impossibility result for adversarial CMDPs. First, we design an algorithm with the desired guarantees when C is known. Then, in the case C is *unknown*, we obtain the same results by embedding multiple instances of such an algorithm in a general *meta-procedure*, which suitably selects them so as to balance the trade-off between regret and constraint violation.

1 Introduction

Reinforcement learning [Sutton and Barto, 2018] is concerned with settings where a learner sequentially interacts with an environment modeled as a *Markov decision process* (MDP) [Puterman, 2014]. Most of the works in the field focus on learning policies that maximize learner's rewards. However, in most of the real-world applications of interest, the learner also has to meet some additional requirements. For instance, bidding agents in ad auctions must *not* deplete their budget [Wu et al., 2018, He et al., 2021], users of recommender systems must *not* be exposed to offending content [Singh et al., 2020], and online companies must ensure a minimum number of items are sold while maximizing the associated profits [Stradi et al., 2024a]. Requirements of this kind can be usually captured by means of *constrained* MDPs (CMDPs) [Altman, 1999], which generalize MDPs by specifying constraints that the learner has to satisfy while maximizing their rewards.

In this paper, we study *online learning* problems in *episodic* CMDPs (see, *e.g.*, [Efroni et al., 2020]), where the goal of the learner is twofold. On the one hand, the learner wants to minimize their *regret*, which measures how much reward they lost over the episodes compared to what they would have obtained by always using a best-in-hindsight constraint-satisfying policy. On the other hand, the learner wants to ensure that the *(cumulative) constraint violation* is minimized during the learning

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

process. Ideally, one seeks to design algorithms with both regret and constraint violation growing sublinearly in the number of episodes T. A crucial feature distinguishing online learning problems in CMDPs is whether rewards and constraints are selected *stochastically* or *adversarially*. Most of the works in the literature focus on the case in which constraints are stochastic (see, *e.g.*, [Wei et al., 2018, Zheng and Ratliff, 2020, Efroni et al., 2020, Qiu et al., 2020, Liu et al., 2021, Bai et al., 2023]), with only one exception addressing settings with adversarial constraints [Stradi et al., 2024b, 2025c]. This is primarily motivated by a well-known impossibility result by Mannor et al. [2009], which prevents any learning algorithm from attaining both sublinear regret and sublinear constraint violation, when competing against a best-in-hindsight policy that satisfies the constraints *on average*. However, dealing with adversarially-selected rewards and constraints is of paramount importance to cope with real-world environments, which are typically non-stationary. For instance, adversarial reward and constraints are present in those settings where the environment encompasses other agents.

1.1 Original Contributions

The main contribution of this paper is to show how to ease the negative result by Mannor et al. [2009]. In order to do so, we consider non-stationary settings that generalize *both* stochastic CMDPs and adversarial ones. Specifically, we address CMDPs where rewards and constraints are selected from probability distributions that are allowed to change *adversarially* from episode to episode. Thus, our CMDPs bridge the gap between fully-stochastic and fully-adversarial ones. We design algorithms whose performances—in terms of regret and constraint violation—smoothly degrade as a suitable measure of the adverseness of rewards and constraints increases. This is called *(adversarial) corruption*, and it intuitively quantifies how much the distributions of rewards and constraints vary over the episodes with respect to some suitable "fictitious" non-corrupted counterparts.

We propose algorithms that attain $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and constraint violation, where C denotes the corruption of the setting. We remark that $C = \Theta(T)$ in the worst case, and, thus, our bounds are coherent with the impossibility result by Mannor et al. [2009]. Moreover, in stochastic CMDPs, our bounds reduce to state-of-the-art $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds [Efroni et al., 2020]. Notably, our algorithms work under *bandit* feedback, namely by only observing rewards and constraint costs of the state-action pairs visited during episodes. Moreover, they are able to manage *positive* constraint violation. This means that they do *not* allow for a negative violation (*i.e.*, a constraint satisfaction) to cancel out a positive one across different episodes. This is a crucial for most of the practical applications. For instance, in autonomous driving, avoiding a collision does *not* "repair" a previous crash.

In the first part of the paper, we design an algorithm (`NS-SOPS`) that works assuming C is known. `NS-SOPS` achieves $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and positive constraint violation by using a *policy search* method *optimistic* in both reward maximization and constraint satisfaction. Specifically, `NS-SOPS` incorporates C in confidence bounds, so as to "boost" optimism and achieve the desired guarantees.

In the second part of the paper, we show how to embed the `NS-SOPS` algorithm in a *meta-procedure* that allows to achieve $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and positive constraint violation when C is *unknown*. The meta-procedure works by instantiating multiple instances of an algorithm for the case in which C is known, each one taking care of a different "guess" on the value of C. Specifically, the meta-procedure acts as a *master* by choosing which instance to follow in order to select a policy at each episode. To do so, it employs an adversarial online learning algorithm, which is fed with losses constructed starting from the Lagrangian of the CMDP problem, suitably modified to account for *positive* constraint violation.

1.2 Related Works

CMDPs with *stochastic* rewards and constraints have been widely investigated. However, their adversarial counterparts are still largely unexplored. In the following, we discuss the works that are most related to ours, while Appendix A provides a comprehensive survey of related works.

Qiu et al. [2020] provide the first primal-dual approach to deal with episodic CMDPs with adversarial losses and stochastic constraints, achieving, under full feedback, both sublinear regret and sublinear (non-positive) constraint violation (*i.e.*, allowing for cancellations). Stradi et al. [2025a] are the first to tackle CMDPs with adversarial losses and stochastic constraints under bandit feedback, by proposing an algorithm that achieves sublinear regret and sublinear positive constraint violation. These works

do *not* consider settings where constraints may change over the episodes. Stradi et al. [2024b, 2025c] study CMDPs with adversarial constraints: the former studies CMDPs with full feedback, the latter focuses on the bandit-feedback setting. Given the impossibility result by Mannor et al. [2009], they propose algorithms that attain sublinear (non-positive) constraint violation (*i.e.*, with cancellations allowed) and a fraction of the optimal reward, thus resulting in a regret growing linearly in T. We show that sublinear regret and sublinear constraint violation can indeed be attained simultaneously if one takes into account the corruption C. Moreover, let us remark that our algorithms deal with *positive* constraint violation, and, thus, they are much more general than those in [Stradi et al., 2024b, 2025c].

The work is also closely related to *corruption-robust* online learning, which, while well-established in different settings, such as unconstrained MDPs with corrupted transitions, *remains largely unexplored for CMDPs*. Specifically, Lykouris et al. [2021] are the first to establish sublinear regret guarantees for MDPs with corrupted rewards and transitions under bandit feedback, achieving $\widetilde{O}(C\sqrt{T})$ regret without requiring prior knowledge of C. Chen et al. [2021] are the first to provide a regret bound that additively depends on C, namely of the order of $\widetilde{O}(\sqrt{T} + C^2)$. This result is improved by Wei et al. [2022], who show a regret bound of order $\widetilde{O}(\sqrt{T} + C)$ under the same conditions. Finally, very recently Jin et al. [2024] study MDPs with adversarial rewards and corrupted transitions, under bandit feedback and unknown corruption value, attaining regret $\widetilde{O}(\sqrt{T} + C^P)$, where C^P is the corruption associated with the transitions. While the techniques employed in these works share some similarities with ours, they *cannot* be easily extended to CMDPs. This is because CMDPs involve a dual objective: minimizing the regret while ensuring low constraint violation. This cannot be achieved through *standard* corralling techniques that are commonly used in the corruption-robust online learning [Agarwal et al., 2017], as these are designed to deal with single-objective settings.

Finally, there is also a related literature that focuses on *non-stationary* CMDPs. While in such settings the learner-environment interaction closely resembles ours, the performance metrics are different from ours and *not* easily comparable. Specifically, Ding and Lavaei [2023] and Wei et al. [2023] consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded. Our work differs from theirs in multiple aspects. First, we consider *positive* constraint violation, while they allow for cancellations. As concerns the definition of regret, ours and that by Ding and Lavaei [2023] and Wei et al. [2023] are *not* comparable. Indeed, they employ a dynamic regret baseline, which, in general, is harder than the static regret employed in our work. However, they compare learner's performances against a dynamic policy that satisfies the constraints at every round. Instead, we consider a policy that satisfies the constraints *on average*, which can perform arbitrarily better than a policy satisfying the constraints at every round. Furthermore, the dependence on T in their regret bound is much worse than ours, even when the non-stationarity is small, namely when it is a constant independent of T and does *not* affect our regret bound. Finally, we do *not* make any assumption on T, while the bounds in [Wei et al., 2023] only hold for large T.

2 Constrained Markov Decision Processes

We study *episodic constrained* MDPs (CMDPs) [Altman, 1999], in which a learner interacts with an unknown environment over T episodes, with the goal of maximizing long-term rewards subject to some constraints. X is a finite set of states of the environment, A is a finite set of actions available to the learner in each state, while the environment dynamics is governed by a transition function $P : X \times A \times X \to [0, 1]$, with $P(x'|x, a)$ denoting the probability of going from state $x \in X$ to $x' \in X$ by taking action $a \in A$.[1] At each episode $t \in [T]$,[2] a reward vector $r_t \in [0, 1]^{|X \times A|}$ is sampled according to a probability distribution $\mathcal{R}_t$, with $r_t(x, a)$ being the reward of taking action $a \in A$ in state $x \in X$ at episode t. Moreover, a constraint cost matrix $G_t \in [0, 1]^{|X \times A| \times m}$ is sampled according to a probability distribution $\mathcal{G}_t$, with $g_{t,i}(x, a)$ being the cost of constraint $i \in [m]$ when

[1] In this paper, we consider w.l.o.g. *loop-free* CMDPs. This means that X is partitioned into L layers $X_0, \dots, X_L$ such that the first and the last layers are singletons, *i.e.*, $X_0 = \{x_0\}$ and $X_L = \{x_L\}$. Moreover, the loop-free property implies that $P(x'|x, a) > 0$ only if $x' \in X_{k+1}$ and $x \in X_k$ for some $k \in [0 \dots L - 1]$. Notice that any episodic CMDP with horizon L that is *not* loop-free can be cast into a loop-free one by suitably duplicating the state space L times, *i.e.*, a state x is mapped to a set of new states (x, k), where $k \in [0 \dots L]$.

[2] In this paper, we denote by $[a \dots b]$ the set of all the natural numbers from $a \in \mathbb{N}$ to $b \in \mathbb{N}$ (both included), while $[b] \coloneqq [1 \dots b]$ is the set of the first $b \in \mathbb{N}$ natural numbers.

taking action $a \in A$ in state $x \in X$ at episode t. We also denote by $g_{t,i} \in [0,1]^{|X \times A|}$ the vector of all the costs $g_{t,i}(x,a)$ associated with constraint i at episode t. Each constraint requires that its corresponding expected cost is kept below a given threshold. The thresholds of all the m constraints are encoded in a vector $\alpha \in [0, L]^m$, with α_i denoting the threshold of the i-th constraint.

We consider a setting in which the sequences of probability distributions $\{\mathcal{R}_t\}_{t=1}^T$ and $\{\mathcal{G}_t\}_{t=1}^T$ are selected *adversarially*. Thus, reward vectors r_t and constraint cost matrices G_t are random variables whose distributions are allowed to change arbitrarily from episode to episode. In other terms, they exhibit non-stationarity. To measure how much such probability distributions change over the episodes, we introduce the notion of *(adversarial) corruption*. In particular, we define the adversarial corruption C_r for the rewards as:

$$C_r \coloneqq \min_{r \in [0,1]^{|X \times A|}} \sum_{t \in [T]} \|\mathbb{E}[r_t] - r\|_1 \,. \tag{1}$$

Intuitively, the corruption C_r encodes the sum over all episodes of the distances between the means $\mathbb{E}[r_t]$ of the adversarial distributions $\mathcal{R}_t$ and a "fictitious" non-corrupted reward vector r. Notice that a similar notion of corruption has been employed in unconstrained MDPs to measure the non-stationarity of transition probabilities; see [Jin et al., 2024]. In the following, we let $r^\circ \in [0,1]^{|X \times A|}$ be a reward vector that attains the minimum in the definition of C_r. Similarly, we introduce the adversarial corruption C_G for constraint costs, which is defined as follows:

$$C_G \coloneqq \min_{G \in [0,1]^{|X \times A| \times m}} \sum_{t \in [T]} \max_{i \in [m]} \|\mathbb{E}[g_{t,i}] - g_i\|_1, \tag{2}$$

where g_i is the i-th component of G. In the following, we let $G^\circ \in [0,1]^{|X \times A| \times m}$ be the constraint cost matrix that attains the minimum in the definition of C_G. Finally, the total adversarial corruption C is defined as $C \coloneqq \max\{C_G, C_r\}$.

Algorithm 1 summarizes how the learner interacts with the environment at episode $t \in [T]$. In particular, the learner chooses a *policy* $\pi : X \times A \to [0,1]$ at each episode, defining a probability distribution over actions to be employed in each state. For ease of notation, we denote by $\pi(\cdot|x)$ the probability distribution for a state $x \in X$, with $\pi(a|x)$ being the probability of selecting action $a \in A$. Let us remark that we assume that the learner knows X and A, but they do *not* know anything about P. Moreover, the *feedback* received by the learner after each episode is *bandit*, as they observe the realizations of rewards and costs only for the state-action pairs (x_k, a_k) actually visited during that episode.

Algorithm 1 Learner-Environment Interaction

$\mathcal{R}_t$ and $\mathcal{G}_t$ are chosen *adversarially*
Choose a policy $\pi_t : X \times A \to [0,1]$
Observe initial state x_0
for $k = 0, \dots, L-1$ **do**
    Play $a_k \sim \pi_t(\cdot|x_k)$
    Observe $r_t(x_k, a_k)$ and $g_{t,i}(x_k, a_k)$ for $i \in [m]$
    Observe new state $x_{k+1} \sim P(\cdot|x_k, a_k)$

Occupancy Measures Given a transition function P and a policy π, we let $q^{P,\pi} \in [0,1]^{|X \times A \times X|}$ be the *occupancy measure* induced by P and π. For $x \in X_k$, $a \in A$, $x' \in X_{k+1}$ with $k \in [0 \dots L-1]$:

$$q^{P,\pi}(x,a,x') \coloneqq \mathbb{P}[x_k = x, a_k = a, x_{k+1} = x' | P, \pi],$$

which is the probability that the learner reaches state x, plays action a, and gets to state x'. Moreover, we also define $q^{P,\pi}(x,a) \coloneqq \sum_{x' \in X_{k+1}} q^{P,\pi}(x,a,x')$ and $q^{P,\pi}(x) \coloneqq \sum_{a \in A} q^{P,\pi}(x,a)$. Following [Rosenberg and Mansour, 2019b], we say that $q \in [0,1]^{|X \times A \times X|}$ is a *valid* occupancy measure of an episodic loop-free CMDP if and only if it satisfies the following three conditions:

$$(i) \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x,a,x') = 1 \quad \forall k \in [0 \dots L-1],$$

$$(ii) \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x,a,x') = \sum_{x' \in X_{k-1}} \sum_{a \in A} q(x',a,x) \quad \forall k \in [1 \dots L-1], \forall x \in X_k,$$

$$(iii)\ P^q = P,$$

where P is the transition function of the CMDP and P^q is the one induced by q (as outlined next). Notice that any valid occupancy measure q induces a transition function P^q and a policy π^q, which are defined as $P^q(x'|x,a) = \frac{q(x,a,x')}{q(x,a)}$ and $\pi^q(a|x) = \frac{q(x,a)}{q(x)}$.

2.1 Performance Metrics

In order to define the performance metrics used to evaluate our *online* learning algorithms, we need to introduce an *offline* optimization problem. Given a CMDP with transition function P, we define the following parametric *linear program* (Program (3)), which is parametrized by a reward vector $r \in [0,1]^{|X \times A|}$, a constraint cost matrix $G \in [0,1]^{|X \times A| \times m}$ and a threshold vector $\alpha \in [0, L]^m$.

$$\text{OPT}_{r,G,\alpha} \coloneqq \begin{cases} \max_{q \in \Delta(P)} & r^\top q \quad \text{s.t.} \\ & G^\top q \leq \alpha, \end{cases} \tag{3}$$

where $q \in [0,1]^{|X \times A|}$ is a vector encoding an occupancy measure, and $\Delta(P)$ is the set of all valid occupancy measures given the transition function P. We say that an instance of Program (3) satisfies *Slater's condition* if the following condition holds.

Condition 1 (Slater)**.** *There exists an occupancy measure $q^\circ \in \Delta(P)$ such that $G^\top q^\circ < \alpha$.*

We also introduce a problem-specific *feasibility parameter* $\rho \in [0, L]$ related to Program (3), defined as $\rho \coloneqq \sup_{q \in \Delta(P)} \min_{i \in [m]} \left[\alpha - G^\top q\right]_i$.[3]. Intuitively, ρ represents by how much feasible solutions to Program (3) strictly satisfy the constraints. Condition 1 is equivalent to say $\rho > 0$, and, whenever $\rho = 0$, there is no occupancy measure that strictly satisfies the constraints in Program (3).

Now, we introduce the notion of *(cumulative) regret* and *(cumulative) positive constraint violation*, which are the performance metrics used to evaluate algorithms. The regret over T episodes is

$$R_T \coloneqq T \cdot \text{OPT}_{\overline{r},\overline{G},\alpha} - \sum_{t \in [T]} \mathbb{E}[r_t]^\top q^{P,\pi_t},$$

where $\overline{r} \coloneqq \frac{1}{T}\sum_{t=1}^T \mathbb{E}[r_t]$ and $\overline{G} \coloneqq \frac{1}{T}\sum_{t=1}^T \mathbb{E}[G_t]$. In the following, we denote by q^* an occupancy measure solving Program (3) instantiated with $\overline{r}$, $\overline{G}$, and α, while its corresponding policy is π^*. Thus, $\text{OPT}_{\overline{r},\overline{G},\alpha} = \overline{r}^\top q^*$ and the regret can be written as $R_T \coloneqq \sum_{t=1}^T \mathbb{E}[r_t]^\top (q^* - q^{P,\pi_t})$. Furthermore, the cumulative positive constraint violation over T episodes is defined as

$$V_T \coloneqq \max_{i \in [m]} \sum_{t \in [T]} \left[\mathbb{E}[G_t]^\top q^{P,\pi_t} - \alpha\right]_i^+, \text{ where we let } [\cdot]^+ \coloneqq \max\{0, \cdot\}.$$

For ease of notation, we refer to q^{P,π_t} as q_t, thus omitting the dependency on P and π.

Remark 1 (Relation with adversarial/stochastic CMDPs)**.** *Our setting naturally encompasses both stochastic and adversarial CMDPs. Indeed, if the distributions $\mathcal{R}_t$ and $\mathcal{G}_t$ do not change over the episodes, then we recover a CMDP with stochastic rewards and constraints. Moreover, when the supports of $\mathcal{R}_t$ and $\mathcal{G}_t$ are singletons (and, thus, mean values are fully revealed), our setting reduces to a CMDP with adversarial rewards and constraints, since $\mathcal{R}_t$ and $\mathcal{G}_t$ are selected adversarially.*

Remark 2 (Impossibility results carrying over from adversarial CMDPs)**.** *Mannor et al. [2009] show that, in online learning problems with constraints selected adversarially, it is impossible to achieve both regret and constraint violation growing sublinearly in T. This result holds for a regret definition that corresponds to ours. Thus, it carries over to our setting. This is why we look for algorithms whose regret and positive constraint violation scale as $\widetilde{\mathcal{O}}(\sqrt{T} + C)$, with a linear dependency on the adversarial corruption C. Notice that the impossibility result by Mannor et al. [2009] does not rule out the possibility of achieving such a guarantee, since regret and positive constraint violation are not sublinear when C grows linearly in T, as it could be the case in a classical adversarial setting.*

3 Learning When C is Known: More Optimism is All You Need

We start studying the case in which the learner *knows* the adversarial corruption C. We propose an algorithm (called `NS-SOPS`, see Algorithm 2), which adopts a suitably-designed UCB-like approach encompassing the adversarial corruption C in the confidence bounds of rewards and constraint costs. This effectively results in "boosting" the *optimism* of the algorithm, and it allows to achieve regret and positive constraint violation of the order of $\tilde{\mathcal{O}}(\sqrt{T} + C)$. The `NS-SOPS` algorithm is a crucial building block to deal with the case in which C is *not* known, as we show in the following section.

[3] Given a vector y, we denote by $[y]_i$ its i-th component.

3.1 NS-SOPS: Non-Stationary Safe Optimistic Policy Search

Algorithm 2 provides the pseudocode of the *non-stationary safe optimistic policy search* (NS-SOPS) algorithm. The algorithm keeps track of suitably-defined confidence bounds for transitions, rewards, and constraint costs. At each episode $t \in [T]$, the algorithm builds a confidence set $\mathcal{P}_t$ for the transition function P by following the same approach as Jin et al. [2020] (see Appendix G for its definition). Instead, for rewards and constraint costs, the algorithm adopts novel *enlarged* confidence bounds, which are suitably designed to tackle non-stationarity.

Given any confidence parameter $\delta \in (0,1)$, by letting $N_t(x,a)$ be the total number of visits to the state-action pair $(x,a) \in X \times A$ up to episode t, the confidence bound for the reward $r_t(x,a)$ is $\phi_t(x,a) \coloneqq \min\left\{1, \sqrt{\frac{\ln(2T|X||A|/\delta)}{2\max\{N_t(x,a),1\}}} + \frac{C}{\max\{N_t(x,a),1\}} + \frac{C}{T}\right\}$, while the bound for the constraint cost $g_{t,i}(x,a)$ is $\xi_t(x,a) \coloneqq \min\left\{1, \sqrt{\frac{\ln(2mT|X||A|/\delta)}{2\max\{N_t(x,a),1\}}} + \frac{C}{\max\{N_t(x,a),1\}} + \frac{C}{T}\right\}$. Intuitively, the first term in the expressions above is derived from Azuma-Hoeffding inequality, the second term allows to deal with the non-stationarity of rewards and constraint costs, while the third term is needed to bound how much the averages $\overline{r}$ and $[\overline{G}]_i$ differ from their "fictitious" non-corrupted counterparts r° and $[G^\circ]_i$, respectively. Algorithm 2 also computes empirical rewards and constraint costs. At each episode $t \in [T]$, for any state-action pair $(x,a) \in X \times A$ and constraint $i \in [m]$, such estimates are defined as $\widehat{r}_t(x,a) \coloneqq \frac{\sum_{\tau \in [t]} \mathbb{I}_\tau(x,a) r_\tau(x,a)}{\max\{N_t(x,a),1\}}$ and $\widehat{g}_{t,i}(x,a) \coloneqq \frac{\sum_{\tau \in [t]} \mathbb{I}_\tau(x,a) g_{\tau,i}(x,a)}{\max\{N_t(x,a),1\}}$, where $\mathbb{I}_\tau(x,a) = 1$ if and only if the pair (x,a) is visited during episode τ, while $\mathbb{I}_\tau(x,a) = 0$ otherwise. For ease of notation, we let $\widehat{G}_t \in [0,1]^{|X \times A| \times m}$ be the matrix with components $\widehat{g}_{t,i}(x,a)$. We refer to Appendix C for a detailed treatment of all the results related to confidence bounds.

Algorithm 2 selects policies with an UCB-like approach encompassing *optimism* in both rewards and constraints satisfaction, following an approach similar to that employed by Efroni et al. [2020]. Specifically, at each episode $t \in [T]$ and for any state-action pair $(x,a) \in X \times A$, the algorithm employs an *upper* confidence bound for the reward $r_t(x,a)$, defined as $\overline{r}_t(x,a) \coloneqq \widehat{r}_t(x,a) + \phi_t(x,a)$, while it uses *lower* confidence bounds for the constraint costs $g_{t,i}(x,a)$, defined as $\underline{g}_{t,i}(x,a) \coloneqq \widehat{g}_{t,i}(x,a) - \xi_t(x,a)$ for every constraint $i \in [m]$. Then, by letting $\overline{r}_t \in [0,1]^{|X \times A|}$ be the vector with components $\overline{r}_t(x,a)$ and $\underline{G}_t$ be the matrix with entries $\underline{g}_{t,i}(x,a)$, Algorithm 2 chooses the policy to be employed in the next episode $t+1$ by solving the following linear program:

Algorithm 2 NS-SOPS

Require: $C, \delta \in (0,1)$

$\pi_1 \leftarrow$ select any policy
for $t \in [T]$ **do**
  Choose policy π_t in Algorithm 1 and observe feedback from interaction
  Compute $\mathcal{P}_t$, $\overline{r}_t$, and $\underline{G}_t$
  $q \leftarrow$ solution to $\texttt{OPT-CB}_{\Delta(\mathcal{P}_t),\overline{r}_t,\underline{G}_t,\alpha}$
  if problem is *feasible* **then**
    $\widehat{q}_{t+1} \leftarrow q$
  else
    $\widehat{q}_{t+1} \leftarrow$ take any $q \in \Delta(\mathcal{P}_t)$
  $\pi_{t+1} \leftarrow \pi^{\widehat{q}_{t+1}}$

$$\texttt{OPT-CB}_{\Delta(\mathcal{P}_t),\overline{r}_t,\underline{G}_t,\alpha} \coloneqq \begin{cases} \arg\max_{q \in \Delta(\mathcal{P}_t)} & \overline{r}_t^\top q \quad \text{s.t.} \\ & \underline{G}_t^\top q \leq \alpha, \end{cases} \tag{4}$$

where $\Delta(\mathcal{P}_t)$ is the set of all the possible valid occupancy measures given the confidence set $\mathcal{P}_t$ (see Appendix G). If $\texttt{OPT-CB}_{\Delta(\mathcal{P}_t),\overline{r}_t,\underline{G}_t,\alpha}$ is feasible, its solution is used to compute a policy to be employed in the next episode, otherwise the algorithm uses any occupancy measure in $\Delta(\mathcal{P}_t)$.

3.2 Theoretical Guarantees of NS-SOPS

Next, we prove the theoretical guarantees attained by Algorithm 2 (see Appendix D for complete proofs of the theorems and associated lemmas). First, we analyze the positive cumulative violation incurred by the algorithm. Formally, we can state the following result.

Theorem 2. *Given any $\delta \in (0,1)$, with probability at least $1 - 8\delta$, Algorithm 2 attains positive violation $V_T = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(mT|X||A|/\delta\right)} + \ln(T)|X||A|C\right)$.*

Intuitively, Theorem 2 is proved by showing that every constraint-satisfying occupancy measure is also feasible for Program (4) with high probability. This holds since Program (4) employs lower confidence bounds for constraint costs. Thus, in order to bound V_T, it is sufficient to analyze at which

rate the feasible region of Program (4) concentrates to the *true* one (*i.e.*, the one defined by $\overline{G}$ in Program (3)). Since by definition of $\xi_t(x,a)$ the feasibility region of Program (4) concentrates as $1/\sqrt{t} + C/t$, the resulting bound for the positive violation V_T is of the order of $\widetilde{\mathcal{O}}(\sqrt{T} + C)$.

The regret guaranteed by Algorithm 2 is formalized by the following theorem.

Theorem 3. *Given any $\delta \in (0,1)$, with probability at least $1 - 9\delta$, Algorithm 2 attains regret* $R_T = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(T|X||A|/\delta\right)} + \ln(T)|X||A|C\right)$.

Theorem 3 is proved similarly to Theorem 2. Indeed, since every constraint-satisfying occupancy measure is feasible for Program (4) with high probability, this also holds for q^*, as it satisfies constraints by definition. Thus, since by definition of $\phi_t(x,a)$ the upper confidence bound for the rewards maximized by Program (4) concentrates as $1/\sqrt{t} + C/t$, the regret bound follows.

Remark 3 (What if some under/overestimate of C is available)**.** *We also study what happens if the learner runs Algorithm 2 with an under/overestimate on the adversarial corruption as input. We defer to Appendix E all the technical results related to this analysis. In particular, it is possible to show that any underestimate on C does not detriment the bound on V_T, which remains the one in Theorem 2. On the other hand, an overestimate on C, say $\widehat{C} > C$, results in a bound on V_T of the order of $\widetilde{\mathcal{O}}(\sqrt{T} + \widehat{C})$, which is worse than the one in Theorem 2. Intuitively, this is because using an overestimate makes Algorithm 2 too conservative. As a result, one could be tempted to conclude that running Algorithm 2 with an underestimate of C as input is satisfactory when the true value of C is unknown. However, this would lead to a regret R_T growing linearly in T, since, intuitively, a regret-minimizing policy could be cut off from the algorithm decision space. This motivates the introduction of additional tools to deal with the case in which C is unknown, as we do in Section 4.*

4 Learning When C is *Not* Known: A Lagrangified Meta-Procedure

In this section, we go beyond Section 3 by studying the more relevant case in which the learner does *not* know the value of the adversarial corruption C. In order to tackle this challenging scenario, we develop a *meta-procedure* (called `Lag-FTRL`, see Algorithm 3) that instantiates multiple instances of an algorithm working for the case in which C is known, with each instance taking care of a different "guess" on the value of C. The `Lag-FTRL` algorithm is inspired by the work of Agarwal et al. [2017] in the context of classical (unconstrained) multi-armed bandit problems. Let us remark that standard "coralling" techniques, such as the one proposed by Agarwal et al. [2017], cannot be easily generalized to our setting, since our objective is twofold: minimizing the regret while simultaneously ensuring small constraint violation. In this section, to deal with our non-stationary CMDP setting, we let `Lag-FTRL` instantiate multiple instances of the `NS-SOPS` algorithm in Section 3.

Algorithm 3 `Lag-FTRL`

Require: $\delta \in (0,1)$

$\Lambda \leftarrow \frac{Lm+1}{\rho}$, $M \leftarrow \lceil \log_2 T \rceil$

$\gamma \leftarrow \sqrt{\ln(M/\delta)/TM}$, $\eta \leftarrow \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T} + \beta_2 + \beta_5 + \sqrt{\beta_4 T}\right)}$

for $j \in [M]$ **do**

  $\texttt{Alg}^j \leftarrow$ stabilized Algorithm 2 with $C = 2^j$

$w_{1,j} \leftarrow 1/M$ for all $j \in [M]$

for $t \in [T]$ **do**

  Sample index $j_t \sim w_t$

  $\pi_t^{j_t} \leftarrow$ policy that $\texttt{Alg}^{j_t}$ would choose

  Choose policy $\pi_t^{j_t}$ in Algorithm 1 and observe feedback from interaction

  Let $\texttt{Alg}^{j_t}$ observe received feedback

  for $j \in [M]$ **do**

    Build $\ell_{t,j}$ as in Equation (5)

    Build $b_{t,j}$ as in Equation (6)

  $w_{t+1} \leftarrow \underset{\substack{w \in \Delta_M, \\ w_j \geq 1/T}}{\arg\min}\ w^\top \sum_{\tau \in [t]} (\ell_t - b_t) + \frac{1}{\eta} \sum_{j \in [M]} \ln \frac{1}{w_j}$

4.1 `Lag-FTRL`: Lagrangified FTRL

At a high level, the *Lagrangified follow-the-regularized-leader* (`Lag-FTRL` for short) algorithm works by instantiating several instances of Algorithm 2, suitably stabilized (see section H), with each instance $\texttt{Alg}^j$ being run for a different "guess" of the (unknown) adversarial corruption value C.

The algorithm plays the role of a *master* by choosing which instance $\texttt{Alg}^j$ to use at each episode. The selection is done by employing an FTRL approach with a suitable log-barrier regularization. In particular, at each episode $t \in [T]$, by letting $\texttt{Alg}^{j_t}$ be the selected instance, the `Lag-FTRL` algorithm employs the policy $\pi_t^{j_i}$ prescribed by $\texttt{Alg}^{j_t}$ and provides feedback to instance $\texttt{Alg}^{j_t}$ only.

The `Lag-FTRL` algorithm faces two main challenges. First, the feedback available to the FTRL procedure implemented at the master level is *partial*. This is because, at each episode $t \in [T]$, the algorithm only observes the result of using the policy $\pi_t^{j_i}$ prescribed by the chosen instance $\texttt{Alg}^{j_t}$, and *not* those of the policies suggested by other instances. The algorithm tackles this challenge by employing *optimistic loss estimators* in the FTRL selection procedure, following an approach originally introduced by Neu [2015]. The second challenge originates from the fact that the goal of the algorithm is to keep under control both the regret and the positive constraint violation. This is accomplished by feeding the FTRL procedure with losses constructed starting from the Lagrangian of the offline optimization problem in Program (3), and suitably modified to manage *positive* violations.

The pseudocode of the `Lag-FTRL` algorithm is provided in Algorithm 3. At Line 4, it instantiates $M \coloneqq \lceil \log_2 T \rceil$ instances of Algorithm 2, with each instance $\texttt{Alg}^j$, for $j \in [M]$, receiving as input a "guess" on the adversarial corruption $C = 2^j$. Notice that, to every instance of Algorithm 2, a standard doubling trick and a stabilization procedure is applied (see Algorithm 4 for additional details). This modification to Algorithm 2 is necessary to guarantee that each instance j attains a regret and positive cumulative constraints violation which linearly degrade in $\nu_{T,j} = {}^1\!/_{\min_{t \in [T]} w_{t,j}}$ and C, when employed by the master algorithm. The algorithm assigns weights defining a probability distribution to instances $\texttt{Alg}^j$, with $w_{t,j} \in [0,1]$ denoting the weight of instance $\texttt{Alg}^j$ at episode $t \in [T]$. We denote by $w_t \in \Delta_M$ the weight vector at episode t, with Δ_M being the M-dimensional simplex. At the first episode, all the weights $w_{1,j}$ are initialized to the value ${}^1\!/_M$ (Line 5). Then, at each episode $t \in [T]$, the algorithm samples an instance index $j_t \in [M]$ according to the probability distribution defined by the weight vector w_t (Line 7), and it employs the policy $\pi_t^{j_t}$ prescribed by $\texttt{Alg}^{j_t}$ (Line 8). The algorithm observes the feedback from the interaction described in Algorithm 1 and it sends such a feedback to instance $\texttt{Alg}^{j_t}$ (Line 10). Then, at Line 12, the algorithm builds an *optimistic* loss estimator to be fed into each instance $\texttt{Alg}^j$. In particular, at episode $t \in [T]$ and for every $j \in [M]$, the optimistic loss estimator is defined as:

$$\ell_{t,j} \coloneqq \frac{\mathbb{I}(j_t = j)}{w_{t,j} + \gamma} \left(L - \sum_{k \in [0 \dots L-1]} r_t(x_k^t, a_k^t) + \Lambda \sum_{i \in [m]} \left[\left(\widehat{G}_t^j \right)^\top \widehat{q}_t^j - \alpha \right]_i^+ \right), \tag{5}$$

where γ is a suitably-defined implicit exploration factor, (x_k^t, a_k^t) is the state-action pair visited at layer k during episode t, Λ is a suitably-defined upper bound on the optimal values of Lagrangian multipliers,[4] $\widehat{G}_t^j$ is the matrix of empirical constraint costs built by the instance $\texttt{Alg}^j$ of Algorithm 2 at episode t, while $\widehat{q}_t^j$ is the occupancy measure computed by instance $\texttt{Alg}^j$ of Algorithm 2 at t. Finally, the algorithm updates the weight vector according to an FTRL update on a cut decision space with a suitable log-barrier regularization and a bonus term b_t defined as:

$$b_{t,j} \coloneqq \left((m\Lambda\beta_5 + \beta_2) + \left(\sqrt{\beta_1} + m\Lambda\sqrt{\beta_4} \right) \sqrt{T} \right) \cdot (\nu_{t,j} - \nu_{t-1,j}), \tag{6}$$

where $\nu_{t,j} = \max_{\tau \leq t} \frac{1}{w_{\tau,j}}$ and the parameters β are linked to the performance of Algorithm 2 (see Line 13 and Section F.2.1 for additional details). See Line 14 for the complete definition of the update. The bonus term purpose is to balance out the term related to the difference between the performance of Algorithm 2 updated at each episode and the performance of its stabilized version, which works under the condition imposed by the master algorithm.

4.2 Theoretical Guarantees of `Lag-FTRL`

Next, we prove the theoretical guarantees attained by Algorithm 3 (see Appendix F for complete proofs of the theorems and associated lemmas). As a first preliminary step, we extend the well-known strong duality result for CMDPs [Altman, 1999] to the case of bounded Lagrangian multipliers.

[4]Notice that, in the definition of Λ, ρ is the feasibility parameter of Program (3) for the reward vector $\overline{r}$, the constraint cost matrix $\overline{G}$, and the threshold vector α. In order to compute Λ, Algorithm 3 needs knowledge of ρ. Nevertheless, our results continue to hold even if Algorithm 3 is only given access to a lower bound on ρ.

Lemma 1. *Given a CMDP with a transition function P, for every reward vector $r \in [0,1]^{|X\times A|}$, constraint cost matrix $G \in [0,1]^{|X\times A|\times m}$, and threshold vector $\alpha \in [0,L]^m$, if Program* (3) *satisfies Slater's condition (Condition 1):*

$$\min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{q\in\Delta(P)} r^\top q - \sum_{i\in[m]} \lambda_i \left[G^\top q - \alpha\right]_i = \max_{q\in\Delta(P)} \min_{\|\lambda\|_1 \in [0, L/\rho]} r^\top q - \sum_{i\in[m]} \lambda_i \left[G^\top q - \alpha\right]_i$$
$$= \text{OPT}_{r,G,\alpha},$$

where $\lambda \in \mathbb{R}^m_{\geq 0}$ is a vector of Lagrangian multipliers and ρ is the feasibility parameter of Program (3).

Intuitively, Lemma 1 states that, under Slater's condition, strong duality continues to hold even when restricting the set of Lagrangian multipliers to the $\lambda \in \mathbb{R}^m_{\geq 0}$ having $\|\lambda\|_1$ bounded by L/ρ. Furthermore, we extend the result in Lemma 1 to the case of a Lagrangian function suitably-modified to encompass *positive* violations. We call it *positive Lagrangian* of Program (3), defined as follows.

Definition 1 (Positive Lagrangian)**.** *Given a CMDP with a transitions P, for every reward vector $r \in [0,1]^{|X\times A|}$, constraint cost matrix $G \in [0,1]^{|X\times A|\times m}$, and threshold vector $\alpha \in [0,L]^m$, the* positive Lagrangian *of Program* (3) *is defined as a function $\mathcal{L} : \mathbb{R}_+ \times \Delta(P) \to \mathbb{R}$ such that $\mathcal{L}(\beta, q) \coloneqq r^\top q - \beta \sum_{i\in[m]} \left[G^\top q - \alpha\right]^+_i$ for every $\beta \geq 0$ and $q \in \Delta(P)$.*

The positive Lagrangian is related to the Lagrangian of a variation of Program (3) in which the $[\cdot]^+$ operator is applied to the constraints. Notice that such a problem does *not* meet Slater's condition, since, by definition of $[\cdot]^+$, it does *not* exist an occupancy measure q° such that $\left[G^\top q^\circ - \alpha\right]^+_i < 0$ for every $i \in [m]$. Nevertheless, we show that some sort of strong duality result still holds for $\mathcal{L}(L/\rho, q)$, when Slater's condition is met by Program (3). This is made formal by the following theorem.

Theorem 4. *Given a CMDP with a transition function P, for every reward vector $r \in [0,1]^{|X\times A|}$, constraint cost matrix $G \in [0,1]^{|X\times A|\times m}$, and threshold vector $\alpha \in [0,L]^m$, if Program* (3) *satisfies Slater's condition (Condition 1), then the following holds:*

$$\max_{q\in\Delta(P)} \mathcal{L}(L/\rho, q) = \max_{q\in\Delta(P)} r^\top q - \frac{L}{\rho} \sum_{i\in[m]} \left[G^\top q - \alpha\right]^+_i = \text{OPT}_{r,G,\alpha},$$

where ρ is the feasibility parameter of Program (3)*.*

Theorem 4 intuitively shows that a L/ρ multiplicative factor on the positive constraint violation is enough to compensate the large rewards that non-feasible policies would attain when employed by the learner. This result is crucial since, without properly defining the Lagrangian function optimized by Algorithm 3, the FTRL optimization procedure would choose instances with both large rewards and large constraint violation, thus preventing the violation bound from being sublinear. By means of Theorem 4, it is possible to provide the following result.

Theorem 5. *If Program* (3) *instantiated with $\overline{r}$, $\overline{G}$ and α satisfies Slater's condition (Condition 1), then, given any $\delta \in (0,1)$, with probability at least $1 - 34\delta$, Algorithm 3 attains positive constraint violation $V_T = \mathcal{O}(m^2L^2|X|\sqrt{|A|T\log\left(mT|X||A|/\delta\right)}\log(T)^2 + m^2L|X|^2|A|^2\log(T)^3\log\left(\log(T)/\delta\right) + m^2L\log(T)^2|X||A|C)$.*

Intuitively, to prove Theorem 5, it is necessary to bound the negative regret attained by the algorithm, *i.e.*, how better Algorithm 3 can perform in terms of rewards with respect to an optimal occupancy in hindsight q^*. Notice that this is equivalent to showing that the FTRL procedure cannot gain more than $\text{OPT}_{\overline{r},\overline{G},\alpha}$ by playing policies that are *not* feasible, or, equivalently, by choosing instances $\texttt{Alg}^j$ with a large corruption guess, which, by definition of the confidence sets employed by Algorithm 2, may play non-feasible policies attaining large rewards. This is done by employing Theorem 4, which shows that the positive Lagrangian does *not* allow the algorithm to achieve too large rewards with respect to q^*. Thus, the violations are still upper bounded by $\tilde{\mathcal{O}}(\sqrt{T} + C)$.

Finally, we prove the regret bound attained by Algorithm 3.

Theorem 6. *If Program* (3) *instantiated with $\overline{r}$, $\overline{G}$ and α satisfies Slater's condition (Condition 1), then, given any $\delta \in (0,1)$, with probability at least $1 - 30\delta$, Algorithm 3 attains regret $R_T = \mathcal{O}(m^2L^2|X|\sqrt{|A|T\log\left(mT|X||A|/\delta\right)}\log(T)^2 + m^2L|X|^2|A|^2\log(T)^3\log\left(\log(T)/\delta\right) + m^2L\log(T)^2|X||A|C)$.*

Bounding the regret attained by Algorithm 3 requires different techniques with respect to bounding constraint violation. Indeed, strong duality is *not* needed, since, even if Λ is set to a too small value and thus the algorithm plays non-feasible policies, then the regret would still be sublinear. The regret bound is strongly related to the optimal value of the problem associated with the positive Lagrangian, which, by definition of $[\cdot]^+$ cannot perform worse than the optimum of Program (3), in terms of rewards gained. Thus, by letting j^* be the index of the instance associated with true corruption value C, proving Theorem 6 reduces to bounding the regret and the constraint violation of instance $\texttt{Alg}^{j^*}$, with the additional challenge of bounding the estimation error of the optimistic loss estimator. Finally, by means of the results for the *known* C case derived in Section 3, we are able to show that the regret is at most $\tilde{\mathcal{O}}(\sqrt{T} + C)$, which is the desired bound.

Acknowledgments

This paper is supported by the Italian MIUR PRIN 2022 Project "Targeted Learning Dynamics: Computing Efficient and Fair Equilibria through No-Regret Algorithms", by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence), and by the EU Horizon project ELIAS (European Lighthouse of AI for Sustainability, No. 101120237).

References

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

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 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 efficient model-free algorithm for mdps with peak constraints. *arXiv preprint arXiv:2003.05555*, 2020.

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.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. 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 2767–2783. PMLR, 17–23 Jul 2022a. URL `https://proceedings.mlr.press/v162/castiglioni22a.html`.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Giulia Romano, and Nicola Gatti. A unifying framework for online optimization with long-term constraints. *Advances in Neural Information Processing Systems*, 35:33589–33602, 2022b.

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

Yifang Chen, Simon Du, and Kevin Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, pages 1561–1570. PMLR, 2021.

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.

Gianmarco Genalti, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Data-dependent regret bounds for constrained mabs. *arXiv preprint arXiv:2505.20010*, 2025.

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.

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.

Tiancheng Jin, Junyan Liu, Chloé Rouyer, William Chang, Chen-Yu Wei, and Haipeng Luo. No-regret online reinforcement learning with adversarial losses and transitions. *Advances in Neural Information Processing Systems*, 36, 2024.

Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In *International Conference on Machine Learning*, pages 3944–3952. PMLR, 2019.

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.

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

Shie Mannor, John N. Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(20):569–590, 2009. URL http://jmlr.org/papers/v10/mannor09a.html.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e5a4d6bf330f23a8707bb0d6001dfbe8-Paper.pdf.

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

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.

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, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning adversarial mdps with stochastic hard constraints. In *Forty-second International Conference on Machine Learning*, 2025a.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained mdps via policy optimization. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Policy optimization for cmdps with bandit feedback: Learning stochastic and adversarial constraints. In *Forty-second International Conference on Machine Learning*, 2025c.

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

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

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.

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.

NeurIPS 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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

The appendix is structured as follows:

- In Appendix A we provide the complete related works.
- In Appendix B we provide the events dictionary.
- In Appendix C we provide the preliminary results on the confidence sets employed to estimate the unknown parameters of the environment.
- In Appendix D we provide the omitted proofs related to the theoretical guarantees when the corruption value is known by the learner, namely, the results attained by Algorithm 2.
- In Appendix E we provide the omitted proofs of the theoretical guarantees attained by Algorithm 2, when a guess on the corruption is given as input to the algorithm.
- In Appendix F we provide the omitted proofs related to the theoretical guarantees when the corruption value is *not* known by the learner, namely, the results attained by Algorithm 3.
- In Appendix G we restate useful results from existing works.
- In Appendix H we provide the results related to stability a corruption-robustness.

A Related works

In the following, we discuss some works that are tightly related to ours. In particular, we first describe works dealing with the online learning problem in MDPs, and, then, we discuss some works studying the constrained version of the classical online learning problem.

Online learning in MDPs The literature on online learning problems [Cesa-Bianchi and Lugosi, 2006] in MDPs is wide (see [Auer et al., 2008, Even-Dar et al., 2009, Neu et al., 2010] for some initial results on the topic). In such settings, two types of feedback are usually studied: in the *full-information feedback* model, the entire loss function is observed after the learner's choice, while in the *bandit feedback* model, the learner only observes the loss due to the chosen action. Azar et al. [2017] study the problem of optimal exploration in episodic MDPs with unknown transitions and stochastic losses when the feedback is bandit. The authors present an algorithm whose regret upper bound is $\tilde{\mathcal{O}}(\sqrt{T})$, thus matching the lower bound for this class of MDPs and improving the previous result by Auer et al. [2008].

Online learning in non-stationary MDPs The literature on non-stationary MDPs encompasses both works on non-stationary rewards and non-stationary transitions. As concerns the first research line, Rosenberg and Mansour [2019b] study the online learning problem in episodic MDPs with adversarial losses and unknown transitions when the feedback is full information. The authors present an online algorithm exploiting entropic regularization and providing a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$. The same setting is investigated by Rosenberg and Mansour [2019a] when the feedback is bandit. In such a case, the authors provide a regret upper bound of the order of $\tilde{\mathcal{O}}(T^{3/4})$, which is improved by Jin et al. [2020] by providing an algorithm that achieves in the same setting a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$. Related to the non-stationarity of the transitions, Wei et al. [2022] study MDPs with adversarial corruption on transition functions and rewards, reaching a regret upper bound of order $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ (where C is the amount of adversarial corruption) with respect to the optimal policy of the non-corrupted MDP . Finally, Jin et al. [2024] is the first to study completely adversarial MDPs with changing transition functions, providing a $\tilde{\mathcal{O}}(\sqrt{T} + C)$ regret bounds, where C is a corruption measure of the adversarially changing transition functions.

Online learning with constraints A central result is provided by Mannor et al. [2009], who show that it is impossible to suffer from sublinear regret and sublinear constraint violation when an adversary chooses losses and constraints. Liakopoulos et al. [2019] try to overcome such an impossibility result by defining a new notion of regret. They study a class of online learning problems with long-term budget constraints that can be chosen by an adversary. The learner's regret metric is modified by introducing the notion of a *K-benchmark*, *i.e.*, a comparator that meets the problem's allotted budget over any window of length K. Castiglioni et al. [2022a,b] deal with the problem

of online learning with stochastic and adversarial losses, providing the first *best-of-both-worlds* algorithm for online learning problems with long-term constraints.

Online learning in CMDPs Online Learning In MDPs with constraints is generally studied when the constraints are selected stochastically. Precisely, Zheng and Ratliff [2020] deal with episodic CMDPs with stochastic losses and constraints, where the transition probabilities are known and the feedback is bandit. The regret upper bound of their algorithm is of the order of $\tilde{\mathcal{O}}(T^{3/4})$, while the cumulative constraint violation is guaranteed to be below a threshold with a given probability. Wei et al. [2018] deal with adversarial losses and stochastic constraints, assuming the transition probabilities are known and the feedback is full information. The authors present an algorithm that guarantees an upper bound of the order of $\tilde{\mathcal{O}}(\sqrt{T})$ on both regret and constraint violation. Bai et al. [2020] provide the first algorithm that achieves sublinear regret when the transition probabilities are unknown, 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 dilemma in episodic CMDPs. These approaches guarantee sublinear regret and constraint violation when transition probabilities, rewards, and constraints are unknown and stochastic, while the feedback is bandit. Stradi et al. [2025b] are the first to attain sublinear positive violation in stochastic CMDPs employing a primal-dual method. Qiu et al. [2020] provide a primal-dual approach based on *optimism in the face of uncertainty*. This work shows the effectiveness of such an approach when dealing with episodic CMDPs with adversarial losses and stochastic constraints, achieving both sublinear regret and constraint violation with full-information feedback. Stradi et al. [2025a] is the first work to tackle CMDPs with adversarial losses and bandit feedback. They propose an algorithm which achieves sublinear regret and sublinear positive constraints violations, assuming that the constraints are stochastic. Stradi et al. [2024b] are the first to study CMDPs with adversarial constraints. Given the well-known impossibility result to learn with adversarial constraints, they propose an algorithm that attains sublinear violation (with cancellations allowed) and a fraction of the optimal reward when the feedback is full. Finally, Ding and Lavaei [2023] and Wei et al. [2023] consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded, as in our work. Nevertheless, our settings differ in multiple aspects. First of all, we consider positive constraints violations, while the aforementioned works allow the cancellations in their definition. We consider a static regret adversarial baseline, while Ding and Lavaei [2023] and Wei et al. [2023] consider the stronger baseline of dynamic regret. Nevertheless, our bounds are not comparable, since we achieve linear regret and violations only in the worst case scenario in which $C = T$, while a sublinear corruption would lead to linear dynamic regret in their work. Finally, we do not make any assumption on the number of episodes, while both the regret and violations bounds presented in Wei et al. [2023] hold only for large T.

B Events dictionary

In the following, we introduce the main events which are related to estimation of the unknown stochastic parameters of the environment.

- **Event** $\mathcal{E}_P$: for all $t \in [T], P \in \mathcal{P}_t$. $\mathcal{E}_P$ holds with probability at least $1 - 4\delta$ by Lemma 18. The event is related to the estimation of the unknown transition function.
- **Event** $\mathcal{E}_G$: for all $t \in [T], i \in [m], (x, a) \in X \times A$:

$$\left| \widehat{g}_{t,i}(x,a) - \frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[g_{\tau,i}(x,a)] \right| \leq \xi_t(x,a).$$

  Similarly,

$$\left| \widehat{g}_{t,i}(x,a) - g_i^\circ(x,a) \right| \leq \xi_t(x,a),$$

  where $g_i^\circ \in [0,1]^{|X \times A|} \coloneqq [G^\circ]_i$.

  $\mathcal{E}_G$ holds with probability at least $1 - \delta$ by Corollary 2. The event is related to the estimation of the unknown constraint functions.

- **Event** $\mathcal{E}_r$: for all $t \in [T], (x,a) \in X \times A$:

$$\left| \widehat{r}_t(x,a) - \frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[r_\tau(x,a)] \right| \leq \phi_t(x,a).$$

Similarly,

$$\left| \widehat{r}_t(x,a) - r^\circ(x,a) \right| \leq \phi_t(x,a).$$

$\mathcal{E}_r$ holds with probability at least $1-\delta$ by Corollary 4. The event is related to the estimation of the unknown reward function.

- **Event** $\mathcal{E}_{\widehat{q}}$: for any $P_t^x \in \mathcal{P}_t$:

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

$\mathcal{E}_{\widehat{q}}$ holds with probability at least $1 - 6\delta$ by Lemma 19. The event is related to the convergence to the true unknown occupancy measure. Notice that $\mathbb{P}\left[\mathcal{E}_{\widehat{q}} \cap \mathcal{E}_P\right] \geq 1 - 6\delta$ by construction.

C Confidence intervals

In this section we will provide the preliminary results related to the high probability confidence sets for the estimation of the cost constraints matrices and the reward vectors.

We start bounding the distance between the *non-corrupted* costs and rewards with respect to the mean of the adversarial distributions.

Lemma 2. *For all $i \in [m]$, fixing $(x,a) \in X \times A$, it holds:*

$$\left| g_i^\circ(x,a) - \frac{1}{T} \sum_{t \in [T]} \mathbb{E}[g_{t,i}(x,a)] \right| \leq \frac{C_G}{T}.$$

Similarly, fixing $(x,a) \in X \times A$, it holds:

$$\left| r^\circ(x,a) - \frac{1}{T} \sum_{t \in [T]} \mathbb{E}[r_t(x,a)] \right| \leq \frac{C_r}{T}.$$

Proof. By triangle inequality and from the definition of C_G, it holds:

$$\begin{aligned}
\left| g_i^\circ(x,a) - \frac{1}{T} \sum_{t \in [T]} \mathbb{E}[g_{t,i}(x,a)] \right| &= \left| \frac{1}{T} \sum_{t \in [T]} (g_i^\circ(x,a) - \mathbb{E}[g_{t,i}(x,a)]) \right| \\
&\leq \frac{1}{T} \sum_{t \in [T]} \left| g_i^\circ(x,a) - \mathbb{E}[g_{t,i}(x,a)] \right| \\
&\leq \frac{C_G}{T}.
\end{aligned}$$

Notice that the proof holds for all $i \in [m]$ since C_G is defined employing the maximum over $i \in [m]$. Following the same steps, it holds:

$$\begin{aligned}
\left| r^\circ(x,a) - \frac{1}{T} \sum_{t \in [T]} \mathbb{E}[r_t(x,a)] \right| &= \left| \frac{1}{T} \sum_{t \in [T]} (r^\circ(x,a) - \mathbb{E}[r_t(x,a)]) \right| \\
&\leq \frac{1}{T} \sum_{t \in [T]} \left| r^\circ(x,a) - \mathbb{E}[r_t(x,a)] \right| \\
&\leq \frac{C_r}{T},
\end{aligned}$$

which concludes the proof. □

In the following lemma, we bound the distance between the empirical mean of the constraints function and the true *non-corrupted* value.

Lemma 3. *Fixing $i \in [m]$, $(x,a) \in X \times A$, $t \in [T]$, for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$:*

$$\left|\widehat{g}_{t,i}(x,a) - g_i^{\circ}(x,a)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}}.$$

Proof. We start bounding the quantity of interest as follows:

$$\begin{aligned}
\left|\widehat{g}_{t,i}(x,a) - g_i^{\circ}(x,a)\right| &= \left|\left(\frac{\sum_{\tau\in[t]} \mathbb{I}_\tau(x,a) g_{\tau,i}(x,a)}{\max\{N_t(x,a),1\}}\right) - g_i^{\circ}(x,a)\right| \\
&\leq \left|\frac{1}{\max\{N_t(x,a),1\}} \sum_{\tau\in[t]} \mathbb{I}_\tau(x,a)\left(g_{\tau,i}(x,a) - \mathbb{E}[g_{\tau,i}(x,a)]\right)\right| \\
&\quad + \left|\frac{1}{\max\{N_t(x,a),1\}} \sum_{\tau\in[t]} \mathbb{I}_\tau(x,a)[\mathbb{E}[g_{\tau,i}(x,a)] - g_i^{\circ}(x,a)]\right|, \qquad (7)
\end{aligned}$$

where we employed the triangle inequality and the definition of $\widehat{g}_{t,i}(x,a)$.

We bound the two terms in Equation (7) separately. For the first term, by Hoeffding's inequality and noticing that constraints values are bounded in $[0,1]$, it holds that:

$$\mathbb{P}\left[\mathcal{A} \geq \frac{c}{\max\{N_t(x,a),1\}}\right] \leq 2\exp\left(-\frac{2c^2}{\max\{N_t(x,a),1\}}\right),$$

where,

$$\mathcal{A} = \left|\left(\frac{\sum_{\tau\in[t]} \mathbb{I}_\tau(x,a) g_{\tau,i}(x,a)}{\max\{N_t(x,a),1\}}\right) - \left(\frac{\sum_{\tau\in[t]} \mathbb{I}_\tau(x,a)\mathbb{E}[g_{\tau,i}(x,a)]}{\max\{N_t(x,a),1\}}\right)\right|,$$

Setting $\delta = 2\exp\left(-\frac{2c^2}{\max\{N_t(x,a),1\}}\right)$ and solving to find a proper value of c we get that with probability at least $1-\delta$:

$$\left|\frac{1}{\max\{N_t(x,a),1\}} \sum_{\tau\in[t]} \mathbb{I}_\tau(x,a)\left(g_{\tau,i}(x,a) - \mathbb{E}[g_{\tau,i}(x,a)]\right)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2}{\delta}\right)}.$$

Finally, we focus on the second term. Thus, employing the triangle inequality and the definition of C_G, it holds:

$$\begin{aligned}
&\left|\frac{1}{\max\{N_t(x,a),1\}} \sum_{\tau\in[t]} \mathbb{I}_\tau(x,a)\left[\mathbb{E}[g_{\tau,i}(x,a)] - g_i^{\circ}(x,a)\right]\right| \\
&\qquad \leq \frac{1}{\max\{N_t(x,a),1\}} \sum_{\tau\in[t]} \mathbb{I}_\tau(x,a)\left|\mathbb{E}[g_{\tau,i}(x,a)] - g_i^{\circ}(x,a)\right| \\
&\qquad \leq \frac{1}{\max\{N_t(x,a),1\}} \sum_{\tau\in[T]} \left|\mathbb{E}[g_{\tau,i}(x,a)] - g_i^{\circ}(x,a)\right| \\
&\qquad \leq \frac{C_G}{\max\{N_t(x,a),1\}},
\end{aligned}$$

which concludes the proof. □

We now prove a similar result for the rewards function.

Lemma 4. *Fixing $(x,a) \in X \times A$, $t \in [T]$, for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$:*

$$\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2}{\delta}\right)} + \frac{C_r}{\max\{N_t(x,a),1\}}.$$

Proof. The proof is analogous to the one of Lemma 3. □

We now generalize the previous results as follows.

Lemma 5. *Given any $\delta \in (0,1)$, for any $(x,a) \in X \times A, t \in [T]$, and $i \in [m]$, it holds with probability at least $1-\delta$:*

$$\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}}.$$

Proof. First let's define $\zeta_t(x,a)$ as:

$$\zeta_t(x,a) := \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}}.$$

From Lemma 3, given $\delta' \in (0,1)$, we have, fixed any $i \in [m]$, $t \in [T]$ and $(x,a) \in X \times A$:

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

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

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

Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,i,t} \left\{\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)\right| \leq \zeta_t(x,a)\right\}\right] \\
&\qquad = 1 - \mathbb{P}\left[\bigcup_{x,a,i,t} \left\{\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)\right| \leq \zeta_t(x,a)\right\}^c\right] \\
&\qquad \geq 1 - \sum_{x,a,i,t} \mathbb{P}\left[\left\{\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)\right| \leq \zeta_t(x,a)\right\}^c\right] \qquad (8) \\
&\qquad \geq 1 - |X||A|mT\delta',
\end{aligned}$$

where Inequality (8) 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 $N_t(x,a)$, we have, with probability at least $1-\delta$:

$$\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}},$$

which concludes the proof. □

We provide a similar result for the rewards function.

Lemma 6. *Given any $\delta \in (0,1)$, for any $(x,a) \in X \times A, t \in [T]$, it holds with probability at least $1-\delta$:*

$$\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2T|X||A|}{\delta}\right)} + \frac{C_r}{\max\{N_t(x,a),1\}}.$$

Proof. First let's define $\psi_t(x,a)$ as:

$$\psi_t(x,a) := \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2}{\delta}\right)} + \frac{C_r}{\max\{N_t(x,a),1\}}.$$

From Lemma 4, given $\delta' \in (0,1)$, we have fixed any $t \in [T]$ and $(x,a) \in X \times A$:

$$\mathbb{P}\left[\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \psi_t(x,a)\right] \geq 1-\delta'.$$

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

$$\mathbb{P}\left[\bigcap_{x,a,t}\left\{\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \psi_t(x,a)\right\}\right].$$

Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,t}\left\{\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \psi_t(x,a)\right\}\right] \\
&\qquad = 1 - \mathbb{P}\left[\bigcup_{x,a,t}\left\{\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \psi_t(x,a)\right\}^c\right] \\
&\qquad \geq 1 - \sum_{x,a,t}\mathbb{P}\left[\left\{\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \psi_t(x,a)\right\}^c\right] \qquad (9) \\
&\qquad \geq 1 - |X||A|T\delta',
\end{aligned}$$

where Inequality (9) holds by Union Bound. Noticing that $r_t(x,a) \leq 1$, substituting δ' with $\delta := \delta'/|X||A|T$ in $\psi_t(x,a)$ with an additional Union Bound over the possible values of $N_t(x,a)$, we have, with probability at least $1-\delta$:

$$\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2T|X||A|}{\delta}\right)} + \frac{C_r}{\max\{N_t(x,a),1\}},$$

which concludes the proof. □

In the following, we bound the distance between the empirical estimation of the constraints and the empirical mean of the mean values of the constraints distribution during the learning dynamic.

Lemma 7. *Given $\delta \in (0,1)$, for all episodes $t \in [T]$, state-action pairs $(x,a) \in X \times A$ and constraint $i \in [m]$,it holds, with probability at least $1-\delta$:*

$$\left|\widehat{g}_{t,i}(x,a) - \frac{1}{T}\sum_{\tau\in[T]}\mathbb{E}[g_{\tau,i}(x,a)]\right| \leq \xi_t(x,a),$$

where,

$$\xi_t(x,a) := \min\left\{1, \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}} + \frac{C_G}{T}\right\}.$$

Proof. We first notice that if $\xi_t(x,a) = 1$, the results is derived trivially by definition on the cost function. We prove now the non trivial case $\sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}} + \frac{C_G}{T} \leq 1$. Employing Lemma 2 and Lemma 5, with probability $1-\delta$ for all $(x,a) \in X \times A$, for all $t \in [T]$ and for all $i \in [m]$, it holds that:

$$\left|\widehat{g}_{t,i}(x,a) - \frac{1}{T}\sum_{\tau\in[T]}\mathbb{E}[g_{\tau,i}(x,a)]\right|$$

$$\leq \left| \widehat{g}_{t,i}(x,a) - g_i^\circ(x,a) \right| + \left| g_i^\circ(x,a) - \frac{1}{T} \sum_{t \in [T]} \mathbb{E}[g_{t,i}(x,a)] \right|$$

$$\leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}} + \frac{C_G}{T},$$

where the first inequality follows from the triangle inequality. This concludes the proof. □

For the sake of simplicity, we analyze our algorithm with respect to the total corruption of the environment, defined as the maximum between the reward and the constraints corruption. In the following, we show that this choice does not prevent the confidence set events from holding.

Corollary 1. *Given a corruption guess $\widehat{C} \geq C_G$ and $\delta \in (0,1)$, for all episodes $t \in [T]$, state-action pairs $(x,a) \in X \times A$ and constraint $i \in [m]$, with probability at least $1-\delta$, it holds:*

$$\left| \widehat{g}_{t,i}(x,a) - \frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[g_{\tau,i}(x,a)] \right| \leq \xi_t(x,a),$$

where,

$$\xi_t(x,a) = \min\left\{1, \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{\widehat{C}}{\max\{N_t(x,a),1\}} + \frac{\widehat{C}}{T}\right\}.$$

Proof. Following the same analysis of Lemma 7 for $\widehat{C} \geq C_G$, it holds

$$\left| \widehat{g}_{t,i}(x,a) - \frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[g_{\tau,i}(x,a)] \right|$$

$$\leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}} + \frac{C_G}{T}$$

$$\leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{\widehat{C}}{\max\{N_t(x,a),1\}} + \frac{\widehat{C}}{T},$$

which concludes the proof. □

Corollary 2. *Taking the definition of ξ_t employed in Lemma 7 and defining $\mathcal{E}_G$ as the intersection event:*

$$\mathcal{E}_G := \left\{ \left| \widehat{g}_{t,i}(x,a) - g_i^\circ(x,a) \right| \leq \xi_t(x,a),\ \forall (x,a) \in X \times A, \forall t \in [T], \forall i \in [m] \right\} \bigcap$$

$$\left\{ \left| \widehat{g}_{t,i}(x,a) - \frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[g_{\tau,i}(x,a)] \right| \leq \xi_t(x,a),\ \forall (x,a) \in X \times A, \forall t \in [T], \forall i \in [m] \right\},$$

it holds that $\mathbb{P}[\mathcal{E}_G] \geq 1-\delta$.

Notice that by Corollary 1, $\mathcal{E}_G$ includes all the analogous events where ξ_t is built employing an arbitrary adversarial corruption $\widehat{C}$ such that $\widehat{C} \geq C_G$.

In the following, we provide similar results for the reward function.

Lemma 8. *Given $\delta \in (0,1)$, for all episodes $t \in [T]$ and for all state-action pairs $(x,a) \in X \times A$, with probability at least $1-\delta$, it holds:*

$$\left| \widehat{r}_t(x,a) - \frac{1}{T} \sum_{\tau \in [T]} \mathbb{E}[r_\tau(x,a)] \right| \leq \phi_t(x,a),$$

where,

$$\phi_t(x,a) := \min\left\{1, \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2T|X||A|}{\delta}\right)} + \frac{C_r}{\max\{N_t(x,a),1\}} + \frac{C_r}{T}\right\}.$$

Proof. Employing Lemma 2 and Lemma 6, with probability at least $1-\delta$, for all $(x,a) \in X \times A$ and for all $t \in [T]$, it holds:

$$\begin{aligned}
&\left|\widehat{r}_t(x,a) - \frac{1}{T}\sum_{\tau\in[T]} \mathbb{E}[r_\tau(x,a)]\right| \\
&\qquad \leq \left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| + \left|r^\circ(x,a) - \frac{1}{T}\sum_{t\in[T]} \mathbb{E}[r_t(x,a)]\right| \\
&\qquad \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2T|X||A|}{\delta}\right)} + \frac{C_r}{\max\{N_t(x,a),1\}} + \frac{C_r}{T},
\end{aligned}$$

where the first inequality follows from the triangle inequality. Noticing that, by construction,

$$\left|\widehat{r}_t(x,a) - \frac{1}{T}\sum_{\tau\in[T]} \mathbb{E}[r_\tau(x,a)]\right| \leq 1,$$

for all episodes $t \in [T]$ and $(x,a) \in X \times A$ concludes the proof. □

We conclude the section, showing the overestimating the reward corruption does not invalidate the confidence set estimation.

Corollary 3. *Given a corruption guess $\widehat{C} \geq C_r$ and $\delta \in (0,1)$, for all episodes $t \in [T]$ and for all state-action pairs $(x,a) \in X \times A$, with probability at least $1-\delta$, it holds:*

$$\left|\widehat{r}_t(x,a) - \frac{1}{T}\sum_{\tau\in[T]} \mathbb{E}[r_\tau(x,a)]\right| \leq \phi_t(x,a),$$

where,

$$\phi_t(x,a) := \min\left\{1, \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2T|X||A|}{\delta}\right)} + \frac{\widehat{C}}{\max\{N_t(x,a),1\}} + \frac{\widehat{C}}{T}\right\}.$$

Proof. The proof is analogous to the one of Corollary 1. □

Corollary 4. *Taking the definition of ϕ_t employed in Lemma 8 and defining $\mathcal{E}_r$ as the intersection event:*

$$\begin{aligned}
\mathcal{E}_r := &\left\{\left|\widehat{r}_t(x,a) - r^\circ(x,a)\right| \leq \phi_t(x,a),\ \forall (x,a) \in X \times A, \forall t \in [T]\right\} \quad \bigcap \\
&\left\{\left|\widehat{r}_t(x,a) - \frac{1}{T}\sum_{\tau\in[T]} \mathbb{E}[r_\tau(x,a)]\right| \leq \phi_t(x,a),\ \forall (x,a) \in X \times A, \forall t \in [T]\right\},
\end{aligned}$$

it holds that $\mathbb{P}[\mathcal{E}_r] \geq 1-\delta$.

Notice that by Corollary 3, $\mathcal{E}_r$ includes all the analogous events where ϕ_t is built employing an arbitrary adversarial corruption $\widehat{C}$ such that $\widehat{C} \geq C_r$.

D Omitted proofs when the corruption is known

In the following, we provide the main results attained by Algorithm 2 in term of regret and constraints violations. The following results hold when the corruption of the environment is known to the learner.

We start providing a preliminary result, which shows that the linear program solved by Algorithm 2 at each $t \in [T]$ admits a feasible solution, with high probability.

Lemma 9. *For any $\delta \in (0,1)$, for all episodes $t \in [T]$, with probability at least $1-5\delta$, the space defined by linear constraints $\left\{q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \leq \alpha\right\}$ admits a feasible solution and it holds:*

$$\left\{q \in \Delta(P) : \overline{G}^\top q \leq \alpha\right\} \subseteq \left\{q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \leq \alpha\right\}.$$

Proof. Under the event $\mathcal{E}_P$, we have that $\Delta(P) \subseteq \Delta(\mathcal{P}_t)$, for all episodes $t \in [T]$. Similarly, under the event $\mathcal{E}_G$, it holds that $\left\{ q : \frac{1}{T} \sum_{t \in [T]} \mathbb{E}[G_t]^\top q \leq \alpha \right\} \subseteq \left\{ q : \underline{G}_t^\top q \leq \alpha \right\}$. This implies that any feasible solution of the offline problem, is included in the optimistic safe set $\left\{ q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \leq \alpha \right\}$. Taking the intersection event $\mathcal{E}_P \cap \mathcal{E}_G$ concludes the proof. □

We are now ready to provide the violation bound attained by Algorithm 2.

Theorem 2. *Given any $\delta \in (0,1)$, with probability at least $1 - 8\delta$, Algorithm 2 attains positive violation $V_T = \mathcal{O}\left(L|X| \sqrt{|A|T \ln\left(mT|X||A|/\delta\right)} + \ln(T)|X||A|C \right)$.*

Proof. In the following, we will refer as $\mathcal{E}_{\widehat{q}}$ to the event described in Lemma 19, which holds with probability at least $1 - 6\delta$. Thus, under $\mathcal{E}_G \cap \mathcal{E}_{\widehat{q}}$, the linear program solved by Algorithm 2 has a feasible solution (see Lemma 9) and it holds:

$$
\begin{aligned}
V_T &= \max_{i \in [m]} \sum_{t \in [T]} \left[\mathbb{E}[G_t]^\top q_t - \alpha \right]_i^+ \\
&= \max_{i \in [m]} \sum_{t \in [T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ \right)^\top q_t + g_i^{\circ\top} q_t - \alpha_i \right]^+ \\
&\leq \max_{i \in [m]} \sum_{t \in [T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ \right)^\top q_t + \left(\underline{g}_{t-1,i} + 2\xi_{t-1} \right)^\top q_t - \alpha_i \right]^+ && \text{(10a)} \\
&= \max_{i \in [m]} \sum_{t \in [T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ \right)^\top q_t + \underline{g}_{t-1,i}^\top \left(q_t - \widehat{q}_t \right) + \underline{g}_{t-1,i}^\top \widehat{q}_t + 2\xi_{t-1}^\top q_t - \alpha_i \right]^+ \\
&\leq \max_{i \in [m]} \sum_{t \in [T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ \right)^\top q_t + \underline{g}_{t-1,i}^\top \left(q_t - \widehat{q}_t \right) + 2\xi_{t-1}^\top q_t \right]^+ && \text{(10b)} \\
&\leq \max_{i \in [m]} \sum_{t \in [T]} \left| \left(\mathbb{E}[g_{t,i}] - g_i^\circ \right)^\top q_t \right| + 2 \max_{i \in [m]} \sum_{t \in [T]} \left| \xi_{t-1}^\top q_t \right| + \max_{i \in [m]} \sum_{t \in [T]} \left| \underline{g}_{t-1,i}^\top \left(q_t - \widehat{q}_t \right) \right| && \text{(10c)} \\
&\leq \max_{i \in [m]} \sum_{t \in [T]} \left\| \mathbb{E}[g_{t,i}] - g_i^\circ \right\|_1 + 2 \max_{i \in [m]} \sum_{t \in [T]} \xi_{t-1}^\top q_t + \max_{i \in [m]} \sum_{t \in [T]} \left\| q_t - \widehat{q}_t \right\|_1 && \text{(10d)} \\
&\leq C_G + 2 \max_{i \in [m]} \sum_{t \in [T]} \xi_{t-1}^\top q_t + \sum_{t \in [T]} \left\| q_t - \widehat{q}_t \right\|_1, && \text{(10e)}
\end{aligned}
$$

where Inequality (10a) follows from Corollary 2, Inequality (10b) holds since Algorithm 2 ensures, for all $t \in [T]$ and for all $i \in [m]$, that $\underline{g}_{t,i}^\top \widehat{q}_t \leq \alpha_i$, Inequality (10c) holds since $[a+b]^+ \leq |a| + |b|$, for all $a, b \in \mathbb{R}$, Inequality (10d) follows from Hölder inequality since $||\underline{g}_{t,i}(x,a)||_\infty \leq 1$ and $||q_t(x,a)||_\infty \leq 1$, and finally Equation (10e) holds for the definition of C_G.

To bound the last term of Equation (10e), we notice that, under $\mathcal{E}_{\widehat{q}}$, by Lemma 19, it holds:

$$
\sum_{t \in [T]} \left\| q_t - \widehat{q}_t \right\|_1 = \mathcal{O}\left(L|X| \sqrt{|A|T \ln\left(\frac{T|X||A|}{\delta} \right)} \right).
$$

To bound the second term of Equation (10e) we proceed as follows. Under $\mathcal{E}_{\widehat{q}}$,with probability at least $1 - \delta$, it holds:

$$
\begin{aligned}
&\sum_{t \in [T]} \xi_{t-1}^\top q_t \leq \sum_{t \in [T]} \sum_{x,a} \xi_{t-1}(x,a) \mathbb{I}_t(x,a) + L \sqrt{2T \ln \frac{1}{\delta}} && \text{(11a)} \\
&\leq \sum_{x,a} \sum_{t \in [T]} \mathbb{I}_t(x,a) \left(\sqrt{\frac{1}{2 \max\{N_{t-1}(x,a), 1\}} \ln\left(\frac{2mT|X||A|}{\delta} \right)} + \right.
\end{aligned}
$$

$$+\frac{C_G}{\max\{N_{t-1}(x,a),1\}}+\frac{C_G}{T}\Bigg)+L\sqrt{2T\ln\frac{1}{\delta}} \tag{11b}$$

$$\leq\sqrt{\frac{1}{2}\ln\left(\frac{2mT|X||A|}{\delta}\right)}\sum_{x,a}\sum_{t\in[T]}\mathbb{I}_t(x,a)\sqrt{\frac{1}{\max\{N_{t-1}(x,a),1\}}}+$$
$$+C_G\sum_{x,a}\sum_{t\in[T]}\left(\frac{\mathbb{I}_t(x,a)}{\max\{N_{t-1}(x,a),1\}}+\frac{1}{T}\right)+L\sqrt{2T\ln\frac{1}{\delta}}$$
$$\leq 3\sqrt{\frac{1}{2}|X||A|LT\ln\left(\frac{2mT|X||A|}{\delta}\right)}+|X||A|(2+\ln(T))C_G+|X||A|C_G+L\sqrt{2T\ln\frac{1}{\delta}} \tag{11c}$$
$$\leq 3\sqrt{\frac{1}{2}|X||A|LT\ln\left(\frac{2mT|X||A|}{\delta}\right)}+(3+\ln(T))|X||A|C_G+L\sqrt{2T\ln\frac{1}{\delta}}$$
$$=\mathcal{O}\left(\sqrt{|X||A|LT\ln\left(\frac{mT|X||A|}{\delta}\right)}+\ln(T)|X||A|C_G\right),$$

where Inequality (11a) follows from the Azuma-Hoeffding inequality and noticing that $\sum_{x,a}\xi_{t-1}(x,a)q_t(x,a)\leq L$, Equality (11b) follows from the definition of ξ_t and finally, Inequality (11c) holds since $1+\sum_{t=1}^{N_T(x,a)}\sqrt{\frac{1}{t}}\leq 1+2\sqrt{N_T(x,a)}\leq 3\sqrt{N_T(x,a)}$, since $1+\sum_{t=1}^{N_T(x,a)}\frac{1}{t}\leq 2+\ln(T)$ and by Cauchy-Schwarz inequality. Finally, we notice that the intersection event $\mathcal{E}_G\cap\mathcal{E}_{\widehat{q}}\cap\mathcal{E}_{\text{Azuma}}$ holds with the following probability,

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_G\cap\mathcal{E}_{\widehat{q}}\cap\mathcal{E}_{\text{Azuma}}\right]&=1-\mathbb{P}\left[\mathcal{E}_G^C\cup\mathcal{E}_{\widehat{q}}^C\cup\mathcal{E}_{\text{Azuma}}^C\right]\\&\geq 1-\left(\mathbb{P}\left[\mathcal{E}_G^C\right]+\mathbb{P}\left[\mathcal{E}_{\widehat{q}}^C\right]+\mathbb{P}\left[\mathcal{E}_{\text{Azuma}}^C\right]\right)\\&\geq 1-8\delta.\end{aligned}$$

Noticing that, by Corollary 1, what holds for a ξ_t built with corruption value C_G, still holds for a higher corruption (by definition, $C\geq C_G$) concludes the proof. □

We conclude the section providing the regret bound attained by Algorithm 2.

Theorem 3. *Given any $\delta\in(0,1)$, with probability at least $1-9\delta$, Algorithm 2 attains regret $R_T=\mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(T|X||A|/\delta\right)}+\ln(T)|X||A|C\right)$.*

Proof. First, we notice that under the event $\mathcal{E}_r$ it holds that, for all $(x,a)\in X\times A$ and for all $t\in[T]$:

$$\overline{r}_t(x,a)-2\phi_t(x,a)\leq\frac{1}{T}\sum_{t\in[T]}\mathbb{E}[r_t(x,a)].$$

Let's observe that, by Lemma 9, under the event $\mathcal{E}_G\cap\mathcal{E}_P$, $\widehat{q}_t$ is optimal solution for $\overline{r}_{t-1}$ in $\left\{q\in\Delta(\mathcal{P}_t):\underline{G}_t^\top q\leq\alpha\right\}$. Thus, under $\mathcal{E}_G\cap\mathcal{E}_P$ the optimal feasible solution q^* is such that:

$$\overline{r}_{t-1}^\top\widehat{q}_t\geq\overline{r}_{t-1}^\top q^*.$$

Thus under the event $\mathcal{E}_r$, it holds:

$$\begin{aligned}\frac{1}{T}\sum_{t\in[T]}\mathbb{E}[r_t]^\top q^*&\leq\overline{r}_{t-1}^\top q^*\\&\leq\overline{r}_{t-1}^\top\widehat{q}_t\\&\leq\left(\frac{1}{T}\sum_{t\in[T]}\mathbb{E}[r_t]+2\phi_{t-1}\right)^\top\widehat{q}_t.\end{aligned}$$

Thus, we can rewrite the regret (under the event $\mathcal{E}_G \cap \mathcal{E}_r \cap \mathcal{E}_P$) as,

$$
\begin{aligned}
R_T &= \sum_{t\in[T]} \mathbb{E}[r_t]^\top (q^* - q_t) \\
&= \sum_{t\in[T]} \frac{1}{T} \sum_{\tau\in[T]} \mathbb{E}[r_\tau]^\top (q^* - q_t) + \sum_{t\in[T]} \left(\mathbb{E}[r_t] - \overline{r}\right)^\top (q^* - q_t) \\
&= \sum_{t\in[T]} \frac{1}{T} \sum_{\tau\in[T]} \mathbb{E}[r_\tau]^\top (q^* - \widehat{q}_t + \widehat{q}_t - q_t) + \sum_{t\in[T]} \left(\mathbb{E}[r_t] - r^\circ + r^\circ - \overline{r}\right)^\top (q^* - q_t) \\
&\leq \sum_{t\in[T]} \left[\frac{1}{T} \sum_{\tau\in[T]} \mathbb{E}\left[r_\tau\right]^\top (q^* - \widehat{q}_t)\right] + \sum_{t\in[T]} \|\widehat{q}_t - q_t\|_1 + \sum_{t\in[T]} \|\mathbb{E}[r_t] - r^\circ\|_1 + \sum_{t\in[T]} \|r^\circ - \overline{r}\|_1 \\
&\leq \sum_{t\in[T]} 2\phi_{t-1}^\top q_t + \sum_{t\in[T]} \|\widehat{q}_t - q_t\|_1 + 2C_r.
\end{aligned}
$$

By Lemma 18 with probability at least $1 - 6\delta$ under event $\mathcal{E}_{\widehat{q}}$ we can bound $\sum_{t\in[T]} \|\widehat{q}_t - q_t\|_1$ as:

$$
\sum_{t\in[T]} \|\widehat{q}_t - q_t\|_1 = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right).
$$

Finally with probability at least $1 - \delta$ it holds:

$$
\begin{aligned}
&\sum_{t\in[T]} \phi_{t-1}^\top q_t \leq \sum_{t\in[T]} \sum_{x,a} \phi_{t-1}(x,a)\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} && \text{(12a)} \\
&\leq \sum_{x,a}\sum_{t\in[T]} \mathbb{I}_t(x,a)\left(\sqrt{\frac{1}{2\max\{N_{t-1}(x,a),1\}}\ln\left(\frac{2T|X||A|}{\delta}\right)} + \right. \\
&\qquad\qquad \left. + \frac{C_r}{\max\{N_{t-1}(x,a),1\}} + \frac{C_r}{T}\right) + L\sqrt{2T\ln\frac{1}{\delta}} && \text{(12b)} \\
&\leq \sqrt{\frac{1}{2}\ln\left(\frac{2T|X||A|}{\delta}\right)}\sum_{x,a}\sum_{t\in[T]} \mathbb{I}_t(x,a)\sqrt{\frac{1}{\max\{N_{t-1}(x,a),1\}}} + \\
&\qquad\qquad + C_r\sum_{x,a}\sum_{t\in[T]}\left(\frac{\mathbb{I}_t(x,a)}{\max\{N_{t-1}(x,a),1\}} + \frac{1}{T}\right) + L\sqrt{2T\ln\frac{1}{\delta}} \\
&\leq 3\sqrt{\frac{1}{2}|X||A|LT\ln\left(\frac{2T|X||A|}{\delta}\right)} + |X||A|(2+\ln(T))C_r + |X||A|C_r + L\sqrt{2T\ln\frac{1}{\delta}} && \text{(12c)} \\
&\leq 3\sqrt{\frac{1}{2}|X||A|LT\ln\left(\frac{2T|X||A|}{\delta}\right)} + (3+\ln(T))|X||A|C_r + L\sqrt{2T\ln\frac{1}{\delta}} \\
&= \mathcal{O}\left(\sqrt{|X||A|LT\ln\left(\frac{T|X||A|}{\delta}\right)} + \ln(T)|X||A|C_r\right),
\end{aligned}
$$

where Inequality (12a) follows from Azuma-Hoeffding inequality, Equality (12b) holds for the definition of ϕ_t, and Inequality (12c) holds since $1 + \sum_{t=1}^{N_T(x,a)} \sqrt{\frac{1}{t}} \leq 1 + 2\sqrt{N_T(x,a)} \leq 3\sqrt{N_T(x,a)}$, $1 + \sum_{t=1}^{N_T(x,a)} \frac{1}{t} \leq 2 + \ln(T)$ and by Cauchy-Schwarz inequality. Thus, we observe that with probability at least $1 - 9\delta$ it holds:

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

Employing Corollary 3 and the definition of C, which is at least equal to C_r, concludes the proof. $\square$

E Omitted proofs when the knowledge of C is not precise

In this section, we focus on the performances of Algorithm 2 when a guess on the corruption value is given as input. These preliminary results are "the first step" to relax the assumption on the knowledge about the corruption.

First, we present some preliminary results on the confidence set.

Lemma 10. *Given the corruption guess $\widehat{C}_G$, where $C_G = \widehat{C}_G + \epsilon$, with $\epsilon > 0$, and confidence ξ_t as defined in Algorithm 2 using $\widehat{C}_G$ as corruption value, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all episodes $t \in [T]$, state-action pair $(x,a) \in X \times A$ and constraint $i \in [m]$, the following result holds:*

$$g_i^\circ(x,a) \leq \widehat{g}_{t,i}(x,a) + \xi_t(x,a) + \left(\frac{\epsilon}{\max\{N_t(x,a),1\}} + \frac{\epsilon}{T}\right).$$

Similarly, recalling the definition of $\underline{G}_t$, for all episodes $t \in [T]$, state-action pairs $(x,a) \in X \times A$ and constraints $i \in [m]$, it holds:

$$g_i^\circ(x,a) \leq \underline{g}_{t,i}(x,a) + 2\xi_t(x,a) + \left(\frac{\epsilon}{\max\{N_t(x,a),1\}} + \frac{\epsilon}{T}\right).$$

Proof. To prove the result, we recall that, by Corollary 2, with probability at least $1-\delta$, the following holds, for all episodes $t \in [T]$, state-action pairs $(x,a) \in X \times A$ and constraints $i \in [m]$:

$$\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)]\right| \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}} + \frac{C_G}{T},$$

which can be rewritten as:

$$\left|\widehat{g}_{t,i}(x,a) - g_i^\circ(x,a)]\right| \leq \xi_t(x,a) + \frac{\epsilon}{\max\{N_t(x,a),1\}} + \frac{\epsilon}{T},$$

where,

$$\xi_t(x,a) := \min\left\{1, \sqrt{\frac{1}{2\max\{N_t(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{\widehat{C}_G}{\max\{N_t(x,a),1\}} + \frac{\widehat{C}_G}{T}\right\},$$

and $C_G = \widehat{C}_G + \epsilon$, which concludes the proof. □

We are now ready study the regret bound attained by the algorithm when the guess on the corruption is an overestimate.

Theorem 7. *For any $\delta \in (0,1)$, Algorithm 2, when instantiated with corruption value $\widehat{C}$ which is an overestimate of the true value of C, i.e. $\widehat{C} > C_G$ and $\widehat{C} > C_r$, attains with probability at least $1-8\delta$:*

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

Proof. By Corollary 1, it holds that the decision space of the linear program performed by Algorithm 2 contains with high probability the optimal solution that respects to the constraints. Furthermore, employing Corollary 3 and following the proof of Theorem 3 concludes the proof. □

We proceed bounding the violation attained by our algorithm when an underestimate of the corruption is given as input.

Theorem 8. *For any $\delta \in (0,1)$, Algorithm 2, when instantiated with corruption value $\widehat{C}$ which is an underestimate of the true value of C_G, i.e. $\widehat{C} < C_G$, attains with probability at least $1-9\delta$:*

$$V_T = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{mT|X||A|}{\delta}\right)} + \ln(T)|X||A|C_G\right).$$

Proof. First, let's define $\epsilon \in \mathbb{R}^+$ such that $\epsilon := C_G - \widehat{C}$. Then, with probability at least $1-\delta$:

$$V_T = \max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[G_t]^\top q_t - \alpha\right]_i^+ \tag{13a}$$

$$= \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + g_i^{\circ\top} q_t - \alpha_i\right]^+$$

$$\leq \max_{i\in[m]} \sum_{t\in[T]} \Bigg[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + \underline{g}_{t-1,i}^\top \widehat{q}_t + 2\xi_{t-1}^\top q_t + \\ + \sum_{x,a} \left(\frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} + \frac{\epsilon}{T}\right) q_t(x,a) - \alpha_i\Bigg]^+ \tag{13b}$$

$$\leq C_G + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^\top q_t + \sum_{t\in[T]} \|q_t - \widehat{q}_t\|_1 + \\ + \sum_{t\in[T]} \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a) + \epsilon L, \tag{13c}$$

where Inequality (13b) follows from Lemma 10 and Inequality (13c) is derived as in the proof of Theorem 2, and considering that $\|q_t\|_1 = L,\ \forall t \in [T]$. Now, employing the Azuma-Hoeffding inequality, we can bound, with probability at least $1-\delta$ the term $\sum_{t=1}^T \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a)$ as follows:

$$\sum_{t\in[T]} \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a) \leq L\sqrt{2T\ln\frac{1}{\delta}} + \sum_{t\in[T]} \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} \mathbb{I}_t(x,a)$$

$$\leq L\sqrt{2T\ln\frac{1}{\delta}} + \epsilon|X||A|(1+\ln(T)),$$

where we applied Azume Hoeffding inequality and the fact that $\sum_{t\in[N_T(x,a)]} \frac{1}{t} \leq 1 + \ln(T)$. Finally, following the steps of the proof of Theorem 2 to bound the first 3 elements of Inequality (13c) under $\mathcal{E}_{\widehat{q}}$ with probability at least $1-\delta$, and considering that $\epsilon \leq C_G$ and $\widehat{C} \leq C_G$, it holds, with probability at least $1-9\delta$,

$$V_T = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)} + \ln(T)|X||A|C_G\right),$$

which concludes the proof. □

Finally, we provide the violation bound attained by Algorithm 2 when an overestimate of the corruption value is given as input.

Theorem 9. *For any $\delta \in (0,1)$, Algorithm 2, when instantiated with corruption value $\widehat{C}$ which is an overestimate of the true value of C_G, i.e. $\widehat{C} > C_G$, attains with probability at least $1-8\delta$:*

$$V_T = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)} + \ln(T)|X||A|\widehat{C}\right).$$

Proof. The proof follows by employing Corollary 1 to the proof of Theorem 2. □

F Omitted proofs when the corruption is *not* known

In the following section we provide the omitted proofs of the theoretical guarantees attained by Algorithm 3. The algorithm is designed to work when the corruption value is *not* known.

F.1 Lagrangian formulation of the constrained optimization problem

Since Algorithm 3 is based on a Lagrangian formulation of the constrained problem, it is necessary to show that this approach is well characterized. Precisely, we show that a *strong duality-like* result holds even when the Lagrangian function is defined taking the positive violations.

First, we show that strong duality holds with respect to the standard Lagrangian function, even considering a subset of the Lagrangian multiplier space.

Lemma 1. *Given a CMDP with a transition function P, for every reward vector $r \in [0,1]^{|X\times A|}$, constraint cost matrix $G \in [0,1]^{|X\times A|\times m}$, and threshold vector $\alpha \in [0,L]^m$, if Program* (3) *satisfies Slater's condition (Condition 1):*

$$\min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{q\in\Delta(P)} r^\top q - \sum_{i\in[m]} \lambda_i \left[G^\top q - \alpha\right]_i = \max_{q\in\Delta(P)} \min_{\|\lambda\|_1 \in [0, L/\rho]} r^\top q - \sum_{i\in[m]} \lambda_i \left[G^\top q - \alpha\right]_i$$
$$= \text{OPT}_{r,G,\alpha},$$

where $\lambda \in \mathbb{R}^m_{\geq 0}$ is a vector of Lagrangian multipliers and ρ is the feasibility parameter of Program (3)*.*

Proof. The proof follows the one of Theorem 3.3 in [Castiglioni et al., 2022b]. First we prove that, given the Lagrangian function $\mathcal{Q}(\lambda, q) := r^\top q - \sum_{i\in[m]} \lambda_i \left(G_i^\top q - \alpha_i\right)$, it holds:

$$\min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) = \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q),$$

with $\lambda \in \mathbb{R}^m_{\geq 0}$. In fact notice that for all $\lambda \in \mathbb{R}^m_{\geq 0}$ such that $\|\lambda\|_1 > L/\rho$:

$$\max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) \geq \mathcal{Q}(\lambda, q^\circ) \geq -\sum_{i\in[m]} \lambda_i \left(G_i^\top q^\circ - \alpha_i\right) \geq \|\lambda\|_1 \rho > L,$$

where q° is defined as $q^\circ := \arg\max_{q\in\Delta(P)} \min_{i\in[m]} \left[\alpha_i - G_i^\top q\right]$. Moreover since

$$\min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) \leq \max_{q\in\Delta(P)} \mathcal{Q}(\underline{0}, q) = \max_{q\in\Delta(P)} r^\top q \leq L,$$

it holds:

$$\min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) = \min\left\{ \min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q), \min_{\|\lambda\|_1 \geq L/\rho} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) \right\}$$
$$= \min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q).$$

Thus,

$$\begin{aligned}
\text{OPT}_{r,G,\alpha} &= \max_{q\in\Delta(P)} \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \mathcal{Q}(\lambda, q) \\
&\leq \max_{q\in\Delta(P)} \min_{\|\lambda\|_1 \geq L/\rho} \mathcal{Q}(\lambda, q) \\
&\leq \min_{\|\lambda\|_1 \geq L/\rho} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) \\
&= \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{q\in\Delta(P)} \mathcal{Q}(\lambda, q) \\
&= \text{OPT}_{r,G,\alpha},
\end{aligned}$$

where the second inequality holds by the *max-min* inequality and the last step holds by the well-known strong duality result in CMDPs [Altman, 1999]. This concludes the proof. □

In the following, we extend the previous result for the Lagrangian function which encompasses the positive violations.

Theorem 4. *Given a CMDP with a transition function P, for every reward vector $r \in [0,1]^{|X\times A|}$, constraint cost matrix $G \in [0,1]^{|X\times A|\times m}$, and threshold vector $\alpha \in [0,L]^m$, if Program* (3) *satisfies Slater's condition (Condition 1), then the following holds:*

$$\max_{q\in\Delta(P)} \mathcal{L}(L/\rho, q) = \max_{q\in\Delta(P)} r^\top q - \frac{L}{\rho} \sum_{i\in[m]} \left[G^\top q - \alpha\right]_i^+ = \text{OPT}_{r,G,\alpha},$$

where ρ is the feasibility parameter of Program (3)*.*

Proof. Following the definition of Lagrangian function, we have:

$$\begin{aligned}
\max_{q\in\Delta(P)} \mathcal{L}(L/\rho, q) &= \max_{q\in\Delta(P)} r^\top q - \frac{L}{\rho}\sum_{i\in[m]}\left[G_i^\top q - \alpha_i\right]^+ \\
&\leq \max_{q\in\Delta(P)} \min_{\|\lambda\|_1\in[0,L/\rho]} r^\top q - \sum_{i\in[m]}\lambda_i[G_i^\top q - \alpha_i]^+ \\
&\leq \min_{\|\lambda\|_1\in[0,L/\rho]} \max_{q\in\Delta(P)} r^\top q - \sum_{i\in[m]}\lambda_i[G_i^\top q - \alpha_i]^+ \\
&\leq \min_{\|\lambda\|_1\in[0,L/\rho]} \max_{q\in\Delta(P)} r^\top q - \sum_{i\in[m]}\lambda_i\left(G_i^\top q - \alpha_i\right) \\
&= \text{OPT}_{r,G,\alpha}
\end{aligned}$$

where $\lambda \in \mathbb{R}^m_{\geq 0}$ is the Lagrangian vector, the second inequality holds by the *max-min inequality* and the last step follows from Lemma 1. Noticing that for all q belonging to $\left\{q \in \Delta(P) : G^\top q \leq \alpha\right\}$, we have $\mathcal{L}(1/\rho, q) = r^\top q$, which implies that $\max_{q\in\Delta(P)} \mathcal{L}(1/\rho, q) \geq \text{OPT}_{r,G,\alpha}$, concludes the proof. □

F.2 Preliminary results

In the following sections we will refer as:

$$\widehat{V}_T := \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j}+\gamma} \sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j\top}\widehat{q}_t^{\,j} - \alpha_i\right]^+, \tag{14}$$

to the estimated violation attained by the instances of Algorithm 3. Furthermore, we will refer as:

$$\widehat{V}_{T,j^*} := \sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j^*\top}\widehat{q}_t^{\,j^*} - \alpha_i\right]^+, \tag{15}$$

to the estimated violation attained by the optimal instance j^*, namely, the integer in $[M]$ such that the true corruption $C \in [2^{j^*-1}, 2^{j^*}]$.

Furthermore, we will refer as q_t^j to the occupancy measure induced by the policy proposed by $\texttt{Alg}^j$ at episode t, with $j \in [M], t \in [T]$, and we will refer as:

$$\widetilde{g}_{t,i}^j(x,a) := \frac{\sum_{\tau\in[t]}\mathbb{I}_\tau(x,a)\mathbb{I}(j_\tau = j)g_{\tau,i}(x,a)}{\max\{N_t^j(x,a),1\}},$$

to the estimate of the cost computed for j-th algorithm, where $N_t^j(x,a)$ is a counter initialize to 0 in $t = 0$, and which increases by one from episode t to episode $t+1$ whenever $\mathbb{I}_t(x,a)\mathbb{I}(j_t = j) = 1$.

F.2.1 Stability parameters

In the following sections, we will employ the stability parameters β defined as follows:

- $\beta_1 = \mathcal{O}\left(L^2|X|^2|A|\ln\left(\frac{T|X||A|}{\delta}\right)\right)$
- $\beta_2 = \mathcal{O}\left(|X|^2|A|^2\log(T)\log\left(\log(T)/\delta\right)\right)$
- $\beta_3 = \mathcal{O}\left(\ln(T)^2|X||A|\right)$
- $\beta_4 = \mathcal{O}\left(L^2|X|^2|A|\ln\left(\frac{mT|X||A|}{\delta}\right)\right)$
- $\beta_5 = \mathcal{O}\left(|X|^2|A|^2\log(T)\log\left(\log(T)/\delta\right)\right)$
- $\beta_6 = \mathcal{O}\left(\ln(T)^2|X||A|\right)$

F.2.2 Omitted proofs and lemmas

We start providing some preliminary results on the optimistic estimator employed by Algorithm 3.

Lemma 11. *For any $\delta \in (0,1)$, given $\gamma \in \mathbb{R}_{\geq 0}$, with probability at least $1-\delta$, it holds:*

$$\widehat{R}_T \leq \mathcal{O}\left(\gamma TLM + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}\right),$$

where $\widehat{R}_T = \sum_{t\in[T]}\sum_{j\in[M]}\left(w_{t,j}\left(L - \mathbb{E}[r_t]^\top q_t^j\right) - \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}\left(1-r_t\left(x_k^t,a_k^t\right)\right)\right)$.

Proof. We first observe that by construction:

$$\mathbb{E}\left[\sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}\left(1-r_t\left(x_k^t,a_k^t\right)\right)\right] = \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}^2}{w_{t,j}+\gamma}\left(L-\mathbb{E}[r_t]^\top q_t^j\right).$$

Moreover, still by construction, for all episodes $t\in[T]$, it holds:

$$\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}\left(1-r_t\left(x_k^t,a_k^t\right)\right) \leq \sum_{j\in[M]}\mathbb{I}(j_t=j)\sum_{(x_k^t,a_k^t)}\left(1-r_t\left(x_k^t,a_k^t\right)\right) \leq L.$$

Thus, employing Azuma-Hoeffding inequality, with probability at least $1-\delta$, it holds:

$$\sum_{t\in[T]}\sum_{j\in[M]}\left(\frac{w_{t,j}^2}{w_{t,j}+\gamma}(L-\mathbb{E}[r_t]^\top q_t^j) - \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))\right) \leq L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}.$$

Finally we notice that:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}\left(L-\mathbb{E}[r_t]^\top q_t^j\right) - \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}^2}{w_{t,j}+\gamma}\left(L-\mathbb{E}[r_t]^\top q_t^j\right)\\
&\qquad = \sum_{t\in[T]}\sum_{j\in[M]}\left(\frac{w_{t,j}}{w_{t,j}+\gamma}\right)\gamma\left(L-\mathbb{E}[r_t]^\top q_t^j\right)\\
&\qquad \leq \gamma TLM.
\end{aligned}$$

Adding and subtracting $\mathbb{E}\left[\sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}\left(1-r_t\left(x_k^t,a_k^t\right)\right)\right]$ to the quantity of interest and employing the previous bound concludes the proof. □

We provide an additional result on the optimistic estimator employed by Algorithm 3.

Lemma 12. *For any $\delta\in(0,1)$, given $\gamma\in\mathbb{R}_{\geq 0}$, with probability at least $1-\delta$, it holds:*

$$\sum_{t\in[T]}\frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma}\sum_{(x_k^t,a_k^t)}\left(1-r_t\left(x_k^t,a_k^t\right)\right) - \sum_{t\in[T]}\left(L-\mathbb{E}[r_t]^\top q_t^{j^*}\right) = \mathcal{O}\left(\frac{L}{\gamma}\ln\left(\frac{1}{\delta}\right)\right)$$

Proof. The proof closely follows the idea of Corollary 5. We define the loss $\bar{\ell}_t = \sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))$, the optimistic loss estimator $\widehat{\ell}_t := \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))$ and the unbiased estimator $\widetilde{\ell}_t := \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))$.

Employing the same argument as Neu [2015] it holds:

$$\widehat{\ell}_t = \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma}\bar{\ell}_t \leq \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma\bar{\ell}_t/L}\bar{\ell}_t \leq \frac{L}{2\gamma}\frac{2\gamma\bar{\ell}_t/w_{t,j^*}L}{1+\gamma\bar{\ell}_t/w_{t,j^*}L}\mathbb{I}(j_t=j^*) \leq \frac{L}{2\gamma}\ln\left(1+\frac{2\gamma}{L}\widetilde{\ell}_t\right),$$

since $\frac{z}{1+z/2}\leq\ln(1+z), z\in\mathbb{R}^+$. Employing the previous inequality, it holds:

$$\mathbb{E}\left[\exp\left(\frac{2\gamma}{L}\widehat{\ell}_t\right)\middle|\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\exp\left(\frac{2\gamma}{L}\frac{L}{2\gamma}\ln\left(1+\frac{2\gamma}{L}\widetilde{\ell}_t\right)\right)\middle|\mathcal{F}_{t-1}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[1 + \frac{2\gamma}{L}\widetilde{\ell}_t \middle| \mathcal{F}_{t-1}\right] \\
&= 1 + \frac{2\gamma}{L}\mathbb{E}\left[\frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}} \sum_{(x_k^t, a_k^t)} (1 - r_t(x_k^t, a_k^t)) \middle| \mathcal{F}_{t-1}\right] \\
&\leq 1 + \frac{2\gamma}{L}\left(L - \mathbb{E}[r_t]^\top q_t^{j^*}\right) \\
&\leq \exp\left(\frac{2\gamma}{L}\left(L - \mathbb{E}[r_t]^\top q_t^{j^*}\right)\right),
\end{aligned}
$$

where $\mathcal{F}_{t-1}$ is the filtration up to episode t. We conclude the proof employing the Markov inequality as follows:

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{t\in[T]} \frac{2\gamma}{L}\left(\widehat{\ell}_t - \left(L - \mathbb{E}[r_t]^\top q_t^{j^*}\right)\right) \geq \epsilon\right) \\
&\qquad \leq \mathbb{E}\left[\exp\left(\sum_{t\in[T]} \frac{2\gamma}{L}\left(\widehat{\ell}_t - \left(L - \mathbb{E}[r_t]^\top q_t^{j^*}\right)\right)\right)\right]\exp(-\epsilon) \\
&\qquad \leq \exp(-\epsilon).
\end{aligned}
$$

Solving $\delta = \exp(-\epsilon)$ for ϵ we obtain:

$$
\mathbb{P}\left(\sum_{t\in[T]} \left(\widehat{\ell}_t - \left(L - \mathbb{E}[r_t]^\top q_t^{j^*}\right)\right) \geq \frac{L}{2\gamma}\ln\left(\frac{1}{\delta}\right)\right) \leq \delta.
$$

This concludes the proof. □

We are now ready to prove the regret bound attained by FTRL with respect to the Lagrangian underlying problem.

Lemma 13. *For any $\delta \in (0,1)$ and properly setting the learning rate η such that $\eta \leq \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T} + \beta_2 + \beta_5 + \sqrt{\beta_4 T}\right)}$, Algorithm 3 attains, with probability at least $1 - 2\delta$:*

$$
\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{j^*} - \sum_{t\in[T]}\sum_{j\in[M]} w_{t,j}\mathbb{E}[r_t]^\top q_t^j + \frac{Lm+1}{\rho}\widehat{V}_T - \frac{Lm+1}{\rho}\widehat{V}_{T,j^*} \\
&\qquad + \left(\frac{m(mL+1)}{\rho}\beta_5 + \beta_2\right)\nu_{T,j^*} + \left(\sqrt{\beta_1} + \left(\frac{m(Lm+1)}{\rho}\right)\sqrt{\beta_4}\right)\sqrt{T}\nu_{T,j*} \\
&\leq \mathcal{O}\Bigg(\frac{M\ln T}{\eta} + \eta\, m^4 L^4 T M + \eta\, M\ln(T) m^4 L^2 \beta_5^2 + \eta\, M \ln(T)\beta_2^2 \\
&\qquad\qquad + \eta T(\beta_1 + L^2 m^4 \beta_4) M\log(T) + \gamma T L M + L\sqrt{T\ln(1/\delta)} + \frac{L}{\gamma}\ln(1/\delta)\Bigg).
\end{aligned}
$$

Proof. First, we define $\ell_{t,j}$, for all $t \in [T]$, for all $j \in [M]$ as:

$$
\ell_{t,j} := \frac{\mathbb{I}(j_t = j)}{w_{t,j} + \gamma}\left(\sum_{(x_k^t, a_k^t)} (1 - r_t(x_k^t, a_k^t)) + \frac{Lm+1}{\rho}\sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j\top}\widehat{q}_t^j - \alpha_i\right]^+\right),
$$

and $b_{t,j}$ for all $t \in [T]$, for all $j \in [M]$ as:

$$
b_{t,j} := \left(\left(\frac{m(mL+1)}{\rho}\beta_5 + \beta_2\right) + \left(\sqrt{\beta_1} + \frac{m(Lm+1)}{\rho}\sqrt{\beta_4}\right)\sqrt{T}\right)(\nu_{t,j} - \nu_{t-1,j}),
$$

with $\nu_{t,j} = \max_{\tau\in[t]} \frac{1}{w_{\tau,j}}$.

First we prove that $\eta w_{t,j}|\ell_{t,j} - b_{t,j}| \leq 1/2$ for all $t \in [T], j \in [M]$, to apply Lemma 16. It

holds that $\eta w_{t,j}|\ell_{t,j}| \leq \frac{\eta(L\rho+L^2m^2+Lm)}{\rho} \leq \frac{1}{2}$ for all $j \in [M]$, for all $t \in [T]$ as long as $\eta \leq \frac{\rho}{2(L\rho+L^2m^2+Lm)} \leq \frac{\rho}{2(L^2m^2+Lm)}$, which is true if $\eta \leq \frac{\rho}{2Lm(Lm+1)}$. It also holds that

$$\begin{aligned}
\eta w_{t,j}|b_{t,j}| &= \eta w_{t,j}\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)(\nu_{t,j}-\nu_{t-1,j}) \\
&\leq \eta\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)\left(1-\frac{\nu_{t-1,j}}{\nu_{t,j}}\right) \\
&\leq \eta\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right) \\
&\leq \frac{1}{2},
\end{aligned}$$

if $\eta \leq \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T}+\beta_2+\beta_5+\sqrt{\beta_4 T}\right)}$, where we used the fact that $\nu_{t,j} \neq \nu_{t-1,j} \iff {}^1\!/_{w_{t,j}} = \nu_{t,j}$. Thus, if the previous conditions on η hold, and notice that the second condition implies the first, Algorithm 3 attains, by Lemma 16 :

$$\begin{aligned}
&\sum_{t\in[T]}\left[\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))-\frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))\right]+\frac{Lm+1}{\rho}\widehat{V}_T \\
&\leq \frac{M\ln T}{\eta}+2\eta\frac{TM(L\rho+L^2m^2+Lm)^2}{\rho^2} \\
&\quad+2\eta\left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2 M\ln(T)+2T\left(\sqrt{\beta_1}+\left(\frac{m(Lm+1)}{\rho}\right)\sqrt{\beta_4}\right)^2 M\ln(T)\right) \\
&\qquad+\frac{Lm+1}{\rho}\widehat{V}_{T,j^*}+\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}b_{t,j}-\sum_{t\in[T]}b_{t,j^*}, \qquad (16)
\end{aligned}$$

where we used the following inequalities:

- First inequality:

$$\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}^2(\ell_{t,j}-b_{t,j})^2 \leq 2\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}^2\ell_{t,j}^2+2\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}^2b_{t,j}^2,$$

- Second inequality:

$$\left(\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))+\frac{Lm+1}{\rho}\sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j\top}\widehat{q}_t^{\,j}-\alpha_i\right]^+\right) \leq \frac{(L\rho+L^2m^2+Lm)}{\rho},$$

- Third inequality:

$$\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}^2\ell_{t,j}^2 \leq \frac{TM(L\rho+L^2m^2+Lm)^2}{\rho^2},$$

and that, it holds:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}^2b_{t,j}^2 \\
&=\sum_{t\in[T]}\sum_{j\in[M]}(w_{t,j}b_{t,j})^2 \\
&\leq\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)^2\sum_{j\in[M]}\sum_{t\in[T]}\left(\frac{1}{\nu_{t,j}}(\nu_{t,j}-\nu_{t-1,j})\right)^2
\end{aligned} \qquad (17a)$$

$$\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right)\sum_{j\in[M]}\sum_{t\in[T]}\left(1-\frac{\nu_{t-1,j}}{\nu_{t,j}}\right)^2$$

$$\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right)\sum_{j\in[M]}\sum_{t\in[T]}\left(1-\frac{\nu_{t-1,j}}{\nu_{t,j}}\right)$$

$$\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right)\sum_{j\in[M]}\sum_{t\in[T]}\ln\left(\frac{\nu_{t,j}}{\nu_{t-1,j}}\right) \tag{17b}$$

$$\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right)\sum_{j\in[M]}\ln\left(\prod_{t\in[T]}\frac{\nu_{t,j}}{\nu_{t-1,j}}\right)$$

$$\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right)\sum_{j\in[M]}\ln\left(\frac{\nu_{T,j}}{\nu_{0,j}}\right)$$

$$\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right)M\ln\left(T\right), \tag{17c}$$

where Inequality (17a) is true since $\nu_{t,j}-\nu_{t-1,j}\neq 0$ only when $w_{t,j}={}^1\!/_{\nu_{t,j}}$ by definition, Inequality (17b) holds since $1-a\leq -\ln a$, and Inequality (17c) holds since by definition $\nu_{T,j}\leq T$ and $\nu_{0,j}=M$. Notice also that, following a similar reasoning, it holds:

$$\begin{aligned}
&\sum_{t\in[T]}w_{t,j}b_{t,j}-\sum_{t\in[T]}b_{t,j^*}\\
&=\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)\sum_{t\in[T]}\sum_{j\in[M]}\left(1-\frac{\nu_{t-1,i}}{\nu_{t,i}}\right)\\
&\quad-\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)\sum_{t\in[T]}(\nu_{t,j^*}-\nu_{t-1,j^*})\\
&\leq\mathcal{O}\left(m^2L\beta_5M\log(T)+\beta_2M\log(T)+(\sqrt{\beta_1}+Lm^2\sqrt{\beta_4})\sqrt{T}M\log(T)\right)\\
&\quad-\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)\nu_{T,j^*}
\end{aligned}$$

Thus, with probability at least $1-2\delta$, it holds:

$$\begin{aligned}
&\sum_{t\in[T]}\mathbb{E}[r_t]^\top q_t^{j^*}-\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}\mathbb{E}[r_t]^\top q_t^j+\frac{Lm+1}{\rho}\widehat{V}_T\\
&=\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}\left(L-\mathbb{E}[r_t]^\top q_t^j\right)-\sum_{t\in[T]}\left(L-\mathbb{E}[r_t]^\top q_t^{j^*}\right)+\frac{Lm+1}{\rho}\widehat{V}_T \qquad (18)\\
&\leq\mathcal{O}\Bigg(\frac{M\ln T}{\eta}+\eta\,m^4L^4TM+\eta\,M\ln(T)m^4L^2\beta_5^2+\eta\,M\ln(T)\beta_2^2\\
&\qquad+\eta T(\beta_1+L^2m^4\beta_4)M\log(T)+\gamma TLM+L\sqrt{T\ln({}^1\!/_\delta)}+\frac{L}{\gamma}\ln({}^1\!/_\delta)\Bigg)+\frac{Lm+1}{\rho}\widehat{V}_{T,j^*}\\
&\qquad-\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)\nu_{T,j^*}-\left(\sqrt{\beta_1}+\left(\frac{m(Lm+1)}{\rho}\right)\sqrt{\beta_4}\right)\sqrt{T}\nu_{T,j*}, \qquad (19)
\end{aligned}$$

where Equation (18) holds since $\sum_{j\in[M]}w_{t,j}=1,\ \forall t\in[T]$, and Inequality (19) holds, with probability at least $1-2\delta$, by Lemma 11, Lemma 12 and Equation (16). This concludes the proof. □

In order to provide the desired bound R_T and V_T for Algorithm 3, it is necessary to study the relation between the aforementioned performance measures and the terms appearing from the FTRL analysis in Lemma 13.

Thus, we bound the distance between the incurred violation and the estimated one.

Lemma 14. *For any $\gamma \in \mathbb{R}_{\geq 0}$, given $\delta \in (0,1)$, with probability at least $1-10\delta$, it holds:*

$$V_T - \widehat{V}_T = \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{mT|X||A|}{\delta}\right)} + m\ln(T)|X||A|C + \gamma TLM\right).$$

Proof. We start defining the quantity $\widehat{\xi}_{t,j}(x,a)$ – for all episode $t \in [T]$, for all state-action pairs $(x,a) \in X \times A$, for all instance $j \in [M]$ – as in Theorem 2 but using the true value of adversarial corruption C, considering that the counter $N_t^j(x,a)$ increases on one unit from episode t to $t+1$, if and only if $\mathbb{I}(j_t = j)\mathbb{I}_t(x,a) = 1$, and by applying a Union Bound over all instances $j \in [M]$ namely,

$$\widehat{\xi}_{t,j}(x,a) := \min\left\{1, \sqrt{\frac{1}{2\max\{N_t^j(x,a),1\}}\ln\left(\frac{2mMT|X||A|}{\delta}\right)} + \frac{C}{\max\{N_t^j(x,a),1\}} + \frac{C}{T}\right\}, \tag{20}$$

By Corollary 2, and applying a Union Bound on instances $j \in [M]$ simultaneously $\forall t \in [T], \forall i \in [m], \forall (x,a) \in X \times A, \forall j \in [M]$, with probability at least $1-\delta$, it holds:

$$\widehat{g}_{t,i}^j(x,a) + \widehat{\xi}_{t,j}(x,a) \geq g_i^\circ(x,a). \tag{21}$$

Resorting to the definition of $\widehat{V}_T$, we obtain that, with probability at least $1-\delta$, under $\mathcal{E}_{\widehat{q}}$:

$$\begin{aligned}
\widehat{V}_T &= \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[{\widehat{g}_{t,i}^j}^\top \widehat{q}_t^j - \alpha_i\right]^+ \\
&= \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[({\widehat{g}_{t,i}^j}^\top q_t^j + \widehat{\xi}_{t,j}^\top q_t^j - \alpha_i) - \widehat{\xi}_{t,j}^\top q_t^j - {\widehat{g}_{t,i}^j}^\top (q_t^j - \widehat{q}_t^j)\right]^+ \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left(\left[(\widehat{g}_{t,i}^j + \widehat{\xi}_{t,j})^\top q_t^j - \alpha_i\right]^+ - \widehat{\xi}_{t,j}^\top q_t^j - {\widehat{g}_{t,i}^j}^\top |q_t^j - \widehat{q}_t^j|\right) \qquad (22a) \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left(\left[{g_i^\circ}^\top q_t^j - \alpha_i\right]^+ - \widehat{\xi}_{t,j}^\top q_t^j - \|q_t^j - \widehat{q}_t^j\|_1\right) \qquad (22b) \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left(\left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+ - \widehat{\xi}_{t,j}^\top q_t^j\right) - \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \\
&\qquad \cdot \sum_{i\in[m]}\left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^j\right]^+ - \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \qquad (22c)
\end{aligned}$$

where Inequality (22a) holds since $[a-b]^+ \geq [a]^+ - b,\ a \in \mathbb{R}, b \in \mathbb{R}_{\geq 0}$, Inequality (22b) follows from Inequality (21) and since, by definition, $\widehat{g}_{t,i}^j(x,a) \leq 1, \forall (x,a) \in X \times A, \forall i \in [m], \forall t \in [T], \forall j \in [M]$ and, finally, Inequality (22c) holds under event $\mathcal{E}_{\widehat{q}}$ by Lemma 19 after noticing that $\sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\|q_t^j - \widehat{q}_t^j\|_1 \leq \sum_{t\in[T]}\sum_{j\in[M]}\mathbb{I}(j_t=j)\left(\frac{w_{t,j}}{w_{t,j}+\gamma}\right)\sum_{i\in[m]}\|q_t^j - \widehat{q}_t^j\|_1 \leq m\sum_{t\in[T]}\|q_t^{j_t} - \widehat{q}_t^{j_t}\|_1$.

We will bound the previous terms separately.

Lower-bound to $\sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+$.

We bound the term by the Azuma-Hoeffding inequality. Indeed, with probability at least $1-\delta$, it holds:

$$\sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+$$

$$\geq \left(\sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}^2}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ \right) - mL\sqrt{2T \ln\left(\frac{1}{\delta}\right)},$$

where we used the following upper-bound to the martingale sequence:

$$\begin{aligned}
\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ &\leq \sum_{j\in[M]} \mathbb{I}(j_t=j) \left(\frac{w_{t,j}}{w_{t,j}+\gamma} \right) \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j \right]^+ \\
&\leq \sum_{j\in[M]} \mathbb{I}(j_t=j) \sum_{i\in[m]} \|q_t^j\|_1 \\
&\leq m\|q_t^{j_t}\|_1 \\
&\leq mL.
\end{aligned}$$

Moreover, we observe the following bounds:

$$\begin{aligned}
\sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ - \sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}^2}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ \\
\leq \gamma TLm,
\end{aligned}$$

and,

$$\sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ \geq \sum_{j\in[M]} \max_{i\in[m]} \sum_{t\in[T]} w_{t,j} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+.$$

Combining the previous results, we obtain, with probability at least $1-\delta$:

$$\begin{aligned}
&\sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ \\
&\qquad \geq \sum_{j\in[M]} \max_{i\in[m]} \sum_{t\in[T]} w_{t,j} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i \right]^+ - \left(\gamma TLm + Lm\sqrt{2T \ln\left(\frac{1}{\delta}\right)} \right).
\end{aligned}$$

Upper-bound to $\sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \widehat{\xi}_{t,j}^\top q_t^j$.

We bound the term noticing that, with probability at least $1-\delta$, it holds:

$$\begin{aligned}
&\sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \widehat{\xi}_{t,j}^\top q_t^j \\
&\qquad \leq \sum_{j\in[M]} m \max_{i\in[m]} \sum_{t\in[T]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \widehat{\xi}_{t,j}^\top q_t^j \\
&\qquad \leq \sum_{j\in[M]} m \max_{i\in[m]} \sum_{t\in[T]} \sum_{x,a} \mathbb{I}(j_t=j)\mathbb{I}_t(x,a)\widehat{\xi}_{t,j}(x,a) + L\sqrt{2T \ln \frac{1}{\delta}} \\
&\qquad = \mathcal{O}\left(m\sqrt{|X||A|LT \ln\left(\frac{mMT|X||A|}{\delta}\right)} + m \ln T|X||A|C + L\sqrt{T \ln \frac{1}{\delta}} \right),
\end{aligned}$$

where we employed the Azuma-Hoeffding inequality and where the last step holds following the proof of Theorem 2.

Upper-bound to $\sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^j \right]^+$.

We simply bound the quantity of interest as follows:

$$\sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^j \right]^+$$

$$\leq m \max_{i\in[m]} \sum_{t\in[T]} \sum_{j\in[M]} \mathbb{I}(j_t = j) \|g_i^\circ - \mathbb{E}[g_{t,i}]\|_1$$
$$\leq mC.$$

Final result. To conclude we employ the Azuma-Hoeffding inequality on the violation definition, obtaining, with probability at least $1-\delta$:

$$V_T = \sum_{j\in[M]} \max_{i\in[m]} \sum_{t\in[T]} \mathbb{I}(j_t = j) \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+$$
$$\leq \sum_{j\in[M]} \max_{i\in[m]} \sum_{t\in[T]} w_{t,j} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+ + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}.$$

Thus, plugging the previous bounds in Equation (22c), we obtain, with probability at least $1-10\delta$:

$$\begin{aligned}
&V_T - \widehat{V}_T \\
&\leq \sum_{j\in[M]} \max_{i\in[m]} \sum_{t\in[T]} \mathbb{I}(j_t = j) \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+ - \sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\widehat{g}_{t,i}^{j\top} \widehat{q}_t^j - \alpha_i\right]^+ \\
&\leq m \sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j}+\gamma} \widehat{\xi}_{t,j}^\top q_t^j + \sum_{t\in[T]} \sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\frac{1}{T}\sum_{\tau\in[T]} (\mathbb{E}[g_{\tau,i}] - \mathbb{E}[g_{t,i}])^\top q_t^j\right]^+ \\
&\qquad + \gamma TLm + 2Lm\sqrt{2T\left(\frac{1}{\delta}\right)} + \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right) \\
&= \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + m\ln(T)|X||A|C + \gamma TLM\right)
\end{aligned}$$

This concludes the proof. □

We proceed bounding the estimated violation attained by the optimal instance j^*.

Lemma 15. *For any $\delta \in (0,1)$, with probability at least $1-16\delta$, it holds:*

$$\widehat{V}_{T,j^*} \leq \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + m\beta_6 C + m\ln(T)|X||A|C + Lm\frac{\ln\left(\frac{M}{\delta}\right)}{2\gamma}\right)$$
$$+ m\sqrt{\beta_4 T}\nu_{T,j^*} + m\beta_5\nu_{T,j^*}.$$

Proof. We start by observing that with, probability at least $1-\delta$ under $\mathcal{E}_{\widehat{q}}$, the quantity of interest is bounded as follows:

$$\begin{aligned}
&\sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[\widehat{g}_{t,i}^{j^*\top} \widehat{q}_t^{j^*} - \alpha_i\right]^+ \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left(\left[\widehat{g}_{t,i}^{j^*\top}(\widehat{q}_t^{j^*} - q_t^{j^*}) + \widehat{g}_{t,i}^{j^*\top} q_t^{j^*} - \widehat{\xi}_{t,j^*}^\top q_t^{j^*} - \alpha_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*}\right) \qquad (23a) \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left(\left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*} + \right. \\
&\qquad \left. + \left[g_i^{\circ\top} q_t^{j^*} - \mathbb{E}[g_{t,i}]^\top q_t^{j^*}\right]^+ + \|\widehat{q}_t^{j^*} - q_t^{j^*}\|_1\right) \qquad (23b) \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left(\left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*} + \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^{j^*}\right]^+\right)
\end{aligned}$$

$$+ \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \quad (23c)$$

where Inequality (23a) holds since $[a+b]^+ \leq [a]^+ + [b]^+,\ \forall a, b \in \mathbb{R}$ and by the definition of $\widehat{\xi}_{t,j^*}$ (see Equation (20)) which implies that all its elements are positive, Inequality (23b) holds with probability at least $1-\delta$ by Corollary 2 and by union bound over M, and since that $\|\widehat{g}_{t,i}\|_\infty \leq 1$ and Inequality (23c) holds with probability at least $1-6\delta$ by Lemma 19.

Upper-bound to $\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^{j^*}\right]^+$.

It is immediate to bound the quantity of interest employing the definition of corruption C and by Lemma 17. Indeed, with probability at least $1-\delta$:

$$\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^{j^*}\right]^+ \leq Lm\sqrt{2T\ln\left(\frac{1}{\delta}\right)} + mC.$$

Upper-bound to $\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+$.

We bound the quantity of interest as follows. With probability at least $1-11\delta$, it holds:

$$\begin{aligned}
&\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ \\
&\qquad \leq m\sqrt{\beta_4 T}\nu_{T,j^*} + m\beta_5\nu_{T,j^*} + 2m\beta_6 C + Lm\frac{\ln\left(\frac{M}{\delta}\right)}{2\gamma}, \qquad (24a)
\end{aligned}$$

thank to Corollary 5 and Corollary 6 .

Upper-bound to $\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \widehat{\xi}_{t,j^*}^\top q_t^{j^*}$.

First, notice that, with probability at least $1-\delta$, it holds:

$$\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \widehat{\xi}_{t,j^*}^\top q_t^{j^*} - m\sum_{t\in[T]} \mathbb{I}(j_t=j^*)\widehat{\xi}_{t,j^*}^\top q_t^{j^*} \leq L\sqrt{2T\ln\left(\frac{1}{\delta}\right)},$$

where we employed Lemma 17. Now we observe that, with probability at least $1-\delta$, it holds:

$$\begin{aligned}
\sum_{t=1}^T \widehat{\xi}_{t-1,j^*}^\top q_t \mathbb{I}(j_t=j^*) &= \sum_{t=1}^T \sum_{x,a} \widehat{\xi}_{t-1,j^*}(x,a) q_t^{j^*}(x,a)\mathbb{I}(j_t=j^*) \\
&\leq \sum_{t=1}^T \sum_{x,a} \widehat{\xi}_{t-1,j^*}(x,a)\mathbb{I}_t(x,a)\mathbb{I}(j_t=j^*) + L\sqrt{2T\ln\frac{1}{\delta}} \\
&= \mathcal{O}\left(\sqrt{|X||A|LT\ln\left(\frac{mMT|X||A|}{\delta}\right)} + \ln(T)|X||A|C + L\sqrt{T\ln\frac{1}{\delta}}\right),
\end{aligned}$$

where employed the same steps as in the proof of Theorem 2, considering that the counter increases if and only if $\mathbb{I}_t(x,a)\mathbb{I}(j_t=j^*)=1$.

Combining the previous bounds concludes the proof. □

F.3 Main results

In the following, we provide the main results attained by Algorithm 3 in terms of regret and violations. We start providing the regret bound and the related proof.

Theorem 6. *If Program* (3) *instantiated with $\overline{r}$, $\overline{G}$ and α satisfies Slater's condition (Condition 1), then, given any $\delta \in (0,1)$, with probability at least $1-30\delta$, Algorithm 3 attains regret $R_T = \mathcal{O}(m^2L^2|X|\sqrt{|A|T\log\left(mT|X||A|/\delta\right)}\log(T)^2 + m^2L|X|^2|A|^2\log(T)^3\log\left(\log(T)/\delta\right) + m^2L\log(T)^2|X||A|C)$.*

Proof. Employing algorithm 3, with probability at least $1 - 14\delta$, it holds:

$$\begin{aligned}
R_T &= \sum_{t\in[T]} \overline{r}^\top q^* - \sum_{t\in[T]} \overline{r}^\top q_t \\
&= \sum_{t\in[T]} \overline{r}^\top (q^* - q_t^{j^*}) + \sum_{t\in[T]} \overline{r}^\top (q_t^{j^*} - q_t) \\
&= \sqrt{\beta_1 T}\nu_{T,j^*} + \beta_2 \nu_{T,j^*} + 2\beta_3 C + \sum_{t\in[T]} \overline{r}^\top (q_t^{j^*} - q_t) \qquad (25a) \\
&\leq \sqrt{\beta_1 T}\nu_{T,j^*} + \beta_2 \nu_{T,j^*} + 2\beta_3 C + 2C - \frac{Lm+1}{\rho}\widehat{V}_T + \frac{Lm+1}{\rho}\widehat{V}_{T,j^*} \\
&\quad - (\sqrt{\beta_1} + \frac{m(Lm+1)}{\rho}\sqrt{\beta_4})\sqrt{T}\nu_{T,j*} - \left(\beta_2 + \frac{m(mL+1)}{\rho}\beta_5\right)\nu_{T,j^*} \\
&\quad + \mathcal{O}\Big(\frac{M\ln T}{\eta} + \eta\, m^4 L^4 T M + \eta\, M \ln(T) m^4 L^2 \left(\beta_2^2 + \beta_5^2\right) \\
&\quad + \eta T(\beta_1 + L^2 m^4 \beta_4) M \log(T) + \gamma T L M + L\sqrt{T\ln(1/\delta)} + \frac{Lm}{\gamma}\ln(1/\delta) \Big). \qquad (25b)
\end{aligned}$$

where Inequality (25a) hold with probability at least $1 - 11\delta$ by Corollary 7,Inequality (25b) holds with probability at least $1 - 3\delta$ thanks to Lemma 13 and to the following reasoning, which holds with probability at least $1 - \delta$:

$$\begin{aligned}
\sum_{t\in[T]} \overline{r}^\top (q_t^{j*} - q_t) &= \sum_{t\in[T]} (\overline{r} - \mathbb{E}[r_t])^\top (q_t^{j*} - q_t) + \sum_{t\in[T]} \mathbb{E}[r_t]^\top (q_t^{j*} - q_t) \\
&\leq \sum_{t\in[T]} \|\overline{r} - \mathbb{E}[r_t]\|_1 + \sum_{t\in[T]} \mathbb{E}[r_t]^\top \left(q_t^{j*} - q_t\right) \qquad (26a) \\
&\leq 2C + \sum_{t\in[T]} \mathbb{E}[r_t]^\top \left(q_t^{j*} - q_t\right) \qquad (26b) \\
&\leq 2C + \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{j*} - \sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} \mathbb{E}[r_t]^\top q_t^j + L\sqrt{2T\ln(1/\delta)} \qquad (26c)
\end{aligned}$$

where Inequality (26a) holds since $|q_t(x,a) - q_t^{j^*}(x,a)| \leq 1,\ \forall (x,a) \in X \times A$, where Inequality (26b) holds by definition of C, and where Inequality (26c) use Azuma-Hoeffding inequality.

We can apply Lemma 15 to bound $\widehat{V}_{T,j^*}$ with high probability. In fact we observe that with probability at least $1 - 16\delta$, it holds:

$$\begin{aligned}
&\frac{Lm+1}{\rho}\widehat{V}_{T,j^*} \\
&\leq \mathcal{O}\left(m^2 L^2 |X| \sqrt{|A| T \ln\left(\frac{mMT|X||A|}{\delta}\right)} + m^2 L \beta_6 C + m^2 L \ln(T) |X||A| C + L^2 m^2 \frac{\ln\left(\frac{M}{\delta}\right)}{2\gamma} \right) \\
&\qquad + \frac{(Lm+1)m}{\rho}\beta_5 \nu_{T,j^*} + \frac{m(Lm+1)}{\rho}\sqrt{\beta_4 T}\nu_{T,j^*}.
\end{aligned}$$

Finally, combining the previous results and by Union Bound, with probability at least $1 - 30\delta$, it holds:

$$\begin{aligned}
&R_T + \frac{Lm+1}{\rho}\widehat{V}_T \\
&\leq \mathcal{O}\Big(\frac{M\ln T}{\eta} + \eta\, m^4 L^4 T M + \eta\, M \ln(T) m^4 L^2 (\beta_2^2 + \beta_5^2) + \eta T(\beta_1 + L^2 m^4 \beta_4) M \log(T) \\
&\quad + \gamma T L M + L\sqrt{T\ln(1/\delta)} + \frac{Lm}{\gamma}\ln(1/\delta) \\
&\quad + m^2 L^2 |X| \sqrt{|A| T \ln\left(\frac{mMT|X||A|}{\delta}\right)} + mL\beta_6 C + \beta_3 C + m^2 L |X||A| \ln(T) C \Big) \qquad (27)
\end{aligned}$$

which concludes the proof after observing that $\widehat{V}_T \geq 0$, by definition, and setting $\gamma = \sqrt{\frac{\ln(M/\delta)}{TM}}$, $\eta \leq \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T}+\beta_2+\beta_5+\sqrt{\beta_4 T}\right)}$. □

We conclude the section providing the violations bound and the related proof.

Theorem 5. *If Program* (3) *instantiated with $\overline{r}$, $\overline{G}$ and α satisfies Slater's condition (Condition 1), then, given any $\delta \in (0,1)$, with probability at least $1-34\delta$, Algorithm 3 attains positive constraint violation $V_T = \mathcal{O}(m^2L^2|X|\sqrt{|A|T\log\left(mT|X||A|/\delta\right)}\log(T)^2 + m^2L|X|^2|A|^2\log(T)^3\log\left(\log(T)/\delta\right) + m^2L\log(T)^2|X||A|C)$.*

Proof. Starting from Inequality (27), in order to obtain the final violations bound, it is necessary to find an upper bound for $-R_T$. We proceed as follows,

$$\begin{aligned}
\overline{r}^\top q^* &= \text{OPT}_{\overline{r},\overline{G},\alpha} && \text{(28a)}\\
&= \max_{q\in\Delta(P)} \left(\overline{r}^\top q - \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q - \alpha_i\right]^+\right) && \text{(28b)}\\
&\geq \overline{r}^\top q_t - \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t - \alpha_i\right]^+,
\end{aligned}$$

where Equality (28a) holds since q^* is the feasible occupancy that maximizes the reward vector $\overline{r}$ and Equality (28b) holds by Theorem 4 . This implies $\overline{r}^\top q_t - \overline{r}^\top q^* \leq \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t - \alpha_i\right]^+$. Moreover, it holds:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t - \alpha_i\right]^+\\
&\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t - \alpha_i\right]^+ + \sum_{i\in[m]}\left[(\overline{G}_i - \mathbb{E}[g_{t,i}])^\top q_t\right]^+\right) && \text{(29a)}\\
&\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t - \alpha_i\right]^+ + \sum_{i\in[m]}\left\|\overline{G}_i - \mathbb{E}[g_{t,i}]\right\|_1\right) && \text{(29b)}\\
&\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t - \alpha_i\right]^+ + \sum_{i\in[m]}\left(\left\|\overline{G}_i - g_i^\circ\right\|_1 + \|g_i^\circ - \mathbb{E}[g_{t,i}]\|_1\right)\right)\\
&\leq mV_T + 2mC, && \text{(29c)}
\end{aligned}$$

where Inequality (29a) holds since $[a+b]^+ \leq [a]^+ + [b]^+, a\in\mathbb{R}, b\in\mathbb{R}$, Inequality (29b) holds since $q_t(x,a) \leq 1 \forall t \in [T], \forall (x,a) \in X\times A$, and finally Inequality (29c) holds by definition of C and V_T and noticing that $m\max_{i\in[m]} a_i \geq \sum_{i\in[m]} a_i,\ \forall\{a_i\}_{i\in[m]} \subset \mathbb{R}^m$. Thus, combining the previous bounds we lower bound the quantity of interest as follows:

$$\begin{aligned}
R_T + \frac{Lm+1}{\rho}V_T &= \sum_{t\in[T]}\mathbb{E}[r_t]^\top(q^* - q_t) + \frac{Lm+1}{\rho}V_T\\
&= \sum_{t\in[T]}\left(\mathbb{E}[r_t] - \overline{r}\right)^\top(q^* - q_t) + \sum_{t\in[T]}\overline{r}^\top(q^* - q_t) + \frac{Lm+1}{\rho}V_T\\
&\geq -\sum_{t\in[T]}\|\mathbb{E}[r_t] - \overline{r}\|_1 + \sum_{t\in[T]}\overline{r}^\top(q^* - q_t) + \frac{Lm+1}{\rho}V_T && \text{(30a)}\\
&\geq -2C - \frac{L}{\rho}(mV_T + 2mC) + \frac{Lm+1}{\rho}V_T && \text{(30b)}\\
&= -2C - \frac{2LmC}{\rho} + V_T\left(\frac{Lm+1}{\rho} - \frac{Lm}{\rho}\right)
\end{aligned}$$

$$= \frac{1}{\rho} V_T - \left(2C + \frac{2LmC}{\rho} \right), \tag{30c}$$

where Inequality (30a) holds since $\underline{v}^\top \underline{w} \geq -\|\underline{v}\|_1 \|\underline{w}\|_\infty, \forall \underline{v}, \underline{w} \in \mathbb{R}^p, p \in \mathbb{N}$, and where Inequality (30b) holds since $\overline{r}^\top (q^* - q_t) \geq -\frac{L}{\rho} \sum_{i \in [m]} \left[\overline{G}_i^\top q_t - \alpha_i \right]^+ \geq -(mV_T + 2mC)$ and by definition of C. Thus, rearranging Inequality (30c), we finally bound the cumulative violation as follows:

$$\begin{aligned}
V_T &\leq 2\rho C + 2LmC + \rho R_T + (Lm+1)V_T \\
&= 2\rho C + 2LmC + (Lm+1)\left(V_T - \widehat{V}_T\right) + \rho \left(R_T + \frac{Lm+1}{\rho} \widehat{V}_T \right) \\
&\leq \mathcal{O}\left(m^2 L^2 |X| \sqrt{|A| T \ln\left(\frac{mMT|X||A|}{\delta} \right)} + m^2 L \ln(T) |X||A| C + \gamma m T L^2 M \right) \\
&\qquad + \mathcal{O}\left(R_T + \frac{Lm+1}{\rho} \widehat{V}_T \right),
\end{aligned}$$

where the last inequality holds by Lemma 14, with probability at least $1 - 4\delta$ under $\mathcal{E}_{\widehat{q}}$. Employing Equation (27) and a Union Bound, setting $\gamma = \sqrt{\frac{\ln(M/\delta)}{TM}}$ and $\eta \leq \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T} + \beta_2 + \beta_5 + \sqrt{\beta_4 T}\right)}$ concludes the proof. □

G Auxiliary lemmas from existing works

In the following section, we provide useful lemma from existing works.

G.1 Auxiliary lemmas for the FTRL master algorithm

In the following, we provide the optimization bound attained by the FTRL instance employed by Algorithm 3.

Lemma 16 (Jin et al. [2024])**.** *The FTRL algorithm over a convex subset Ω of the $(M-1)$-dimensional simplex Δ_M :*

$$w_{t+1} = \arg\min_{w \in \Omega} \left\{ \sum_{\tau \in [t]} \ell_\tau^\top w + \frac{1}{\eta} \sum_{j \in [M]} \ln\left(\frac{1}{w_j} \right) \right\},$$

ensures for all $u \in \Omega$:

$$\sum_{t \in [T]} \ell_t^\top (w_t - u) \leq \frac{M \ln T}{\eta} + \eta \sum_{t \in [T]} \sum_{j \in [M]} w_{t,j}^2 \ell_{t,j}^2,$$

as long as $\eta w_{t,j} |\ell_{t,j}| \leq \frac{1}{2}$ for all t, j.

G.2 Auxiliary lemmas for the optimistic loss estimator

In the following, we provide some results related to the optimistic biased estimator of the loss function. Notice that, given any loss vector $\ell_t \in [0,1]^M$, the following results are provided for $\widehat{\ell}_{t,j} := \frac{\mathbb{I}_t(j)}{w_{t,j} + \gamma_t} \ell_{t,j}$, where $j \in [M]$, $\ell_{t,j}$ is the j-th component of the loss vector, $\mathbb{I}_t(j)$ is the indicator functions which is 1 when arm j is played and γ_t is defined as in the following lemmas.

Lemma 17 (Neu [2015])**.** *Let (γ_t) be a fixed non-increasing sequence with $\gamma_t \geq 0$ and let $\alpha_{t,j}$ be nonnegative $\mathcal{F}_{t-1}$-measurable random variables satisfying $\alpha_{t,j} \leq 2\gamma_t$ for all t and j. Then, with probability at least $1 - \delta$,*

$$\sum_{t \in [T]} \sum_{j \in [M]} \alpha_{t,j} \left(\widehat{\ell}_{t,j} - \ell_{t,j} \right) \leq \ln\left(\frac{1}{\delta} \right).$$

Corollary 5 (Neu [2015]). *Let $\gamma_t = \gamma \geq 0$ for all t. With probability at least $1-\delta$,*

$$\sum_{t\in[T]} \left(\widehat{\ell}_{t,j} - \ell_{t,j}\right) \leq \frac{\ln\left(\frac{M}{\delta}\right)}{2\gamma},$$

simultaneously holds for all $j \in [M]$.

G.3 Auxiliary lemmas for the transitions estimation

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:

$$\overline{P}_t\left(x'|x,a\right) = \frac{M_t(x,a,x')}{\max\left\{1, N_t(x,a)\right\}}.$$

Then, the confidence set for P at episode $t \in [T]$ is defined as:

$$\mathcal{P}_t := \Bigg\{ \widehat{P} : \left|\overline{P}_t(x'|x,a) - \widehat{P}(x'|x,a)\right| \leq \epsilon_t(x'|x,a), \\ \forall (x,a,x') \in X_k \times A \times X_{k+1}, k \in [0...L-1] \Bigg\},$$

where $\epsilon_t(x'|x,a)$ is defined as:

$$\epsilon_t(x'|x,a) \coloneqq 2\sqrt{\frac{\overline{P}_t\left(x'|x,a\right)\ln\left(T|X||A|/\delta\right)}{\max\left\{1, N_t(x,a)-1\right\}}} + \frac{14\ln\left(T|X||A|/\delta\right)}{3\max\left\{1, N_t(x,a)-1\right\}},$$

for some confidence $\delta \in (0,1)$.

Given the estimated transition function space $\mathcal{P}_t$, the following result can be proved.

Lemma 18 (Jin et al. [2020]). *With probability at least $1-4\delta$, we have $P \in \mathcal{P}_t$ for all $t \in [T]$.*

Notice that we refer to the event $P \in \mathcal{P}_t$ for all $t \in [T]$ as $\mathcal{E}_P$.

We underline that the estimated occupancy measure space by Algorithm 2 is the following:

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

To conclude, we restate the result which bounds the cumulative distance between the estimated occupancy measure and the real one.

Lemma 19 (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,*

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

H Auxiliary lemmas for stability

In this section we state the results related to the stability of the arm-algorithms when C is not known. The procedure is inspired by Jin et al. [2024] and Agarwal et al. [2017], but adapted to the

case of *Constrained* MDP in high probability. We first give some important definitions. In these definitions we will use C_t as the value of adversarial corruption at episode $t \in [T]$, where C_t is defined as $C_t := \max\{C_t^G, C_t^r\}$, which meets the requirement of upper bounding the adversarial corruption at each considered episode. In addition it holds that $\sum_{t\in[T]} C_t \leq C_r + C_G$ or equivalently $C \leq \sum_{t\in[T]} C_t \leq 2C$, which does not influence the order of the analysis.

Definition 2. *A CMDP algorithm is **corruption-robust** if it takes θ (a guess on the corruption amount) as input, and achieves for any random stopping time $t' \leq T$, whenever $\sum_{t\in[t']} C_t < \theta$:*

$$\sum_{t\in[t']} \overline{r}^\top (q^* - q_t) \leq \sqrt{\beta_1 t'} + (\beta_2 + \beta_3\theta)\,\mathbb{I}(t' \geq 1),$$

and

$$\max_{i\in[m]} \sum_{t\in[t']} \left[g_{t,i}^\top q_t - \alpha_i\right]^+ \leq \sqrt{\beta_4 t'} + (\beta_5 + \beta_6\theta)\,\mathbb{I}(t' \geq 1).$$

Notice that Algorithm 2 is corruption-robust after applying a doubling trick to make it work for any stopping time, with probability at least $1 - 9\delta$ thank to Theorem 7 and Theorem 9 Furthermore, we introduce the notion of α-stability. An algorithm is considered to be α-stable, if its regret under condition imposed by Algorithm 3 is of order $\nu_T^\alpha \cdot \tilde{\mathcal{O}}\left(R_T\right)$, where R_T is the upper bound on the regret attained by the algorithm if it receives feedback at each episode. In particular, we are interested in the 1-stability.

Definition 3. *An algorithm is 1-**stable** if, under the condition imposed by Algorithm 3, it holds:*

$$\sum_{t\in[T]} \overline{r}^\top (q^* - q_t) \leq \sqrt{\beta_1 T}\nu_{j,T} + \beta_2\nu_{j,T} + \beta_3 C,$$

and

$$\max_{i\in[m]} \sum_{t\in[T]} \left[g_{t,i}^\top q_t - \alpha_i\right]^+ \leq \sqrt{\beta_4 T}\nu_{j,T} + \beta_5\nu_{j,T} + \beta_6 C.$$

We can use the procedure defined by Algorithm 4 - and originally proposed by Jin et al. [2024] - to transform a generic corruption robust algorithm to a 1-stable algorithm. Differently from Jin et al. [2024], in our setting, we use the natural symmetry between regret and positive cumulative constraints violation to stabilize both the regret and the positive cumulative constraints violation. We have a different bound for C_t (value of adversarial corruption at episode t): indeed, $C_t \leq \max\{\|\mathbb{E}[r_t] - r^\circ\|_1, \max_{i\in[m]}\|\mathbb{E}[g_{t,i}] - g_i^\circ\|_1\}$ is bounded by $|X||A|$. Finally, we are interested in obtaining results that hold in high probability rather than in expectation. To do so, we focus on 1-stability guarantee rather than $1/2$-stability as in Jin et al. [2024] since removing the expectation prevents us from achieving the result above with lower coefficients. We can state the following result.

Lemma 20. *Given an algorithm which is corruption robust according to Definition 2 with parameters $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ and $\beta_1 \geq \mathcal{O}(L^2\log(T/\delta))$, $\beta_4 \geq \mathcal{O}(L^2\log(T/\delta))$, with probability at least $1 - p$ with $p \in (0,1)$, then, it is possible convert it to an 1-stable algorithm with probability at least $1 - p - 2\delta$ according to Definition 3 with parameters $(\beta_1', \beta_2', \beta_3', \beta_4', \beta_5', \beta_6')$ as $\beta_1' = \mathcal{O}\left(\beta_1\right), \beta_2' = \mathcal{O}\left(\beta_2 + \beta_3|X||A|\log(\log(T)/\delta)\right), \beta_3' = \mathcal{O}\left(\beta_3\log(T)\right), \beta_4' = \mathcal{O}\left(\beta_4\right), \beta_5' = \mathcal{O}\left(\beta_5 + \beta_6|X||A|\log(\log(T)/\delta)\right), \beta_6' = \mathcal{O}\left(\beta_6\log(T)\right)$, employing Algorithm 4.*

Proof. Suppose Algorithm 4 is initialized with the true value of adversarial corruption C. We will first prove the result for the regret. We will start by considering a generic instance algorithm $k \in [M]$. Define the quantity $d_{t,k} = \mathbb{I}(w_t \in (2^{-k-1}, 2^{-k}])$ and $h_{t,k} = \mathbb{I}(\text{Instance } k \text{ receives feedback at episode } t)$. We observe that with probability at least $1 - \left(p + \mathbb{P}\left(\bigcup_{k\in[\log_2(T)]}\{\sum_{t\in[T]} C_t d_{t,k}h_{t,k} > \theta_k\}\right)\right)$ it holds:

$$\sum_{t\in[T]} \overline{r}^\top (q^* - q_t) d_{t,k}h_{t,k} \leq \sqrt{\beta_1 \sum_{t\in[T]} d_{t,k}h_{t,k}} + (\beta_2 + \beta_3\theta)\max_{t\in[T]} d_{t,k},$$

by the corruption-robust property of instance k. We study now the quantity $\mathbb{P}\left(\bigcup_{k\in[M]}\{\sum_{t\in[T]} C_t d_{t,k}h_{t,k} > \theta_k\}\right)$. Notice that $\mathbb{E}[h_{t,k}|d_{t,k}] = 2^{-k-1}d_{t,k}$, and since

$d_{t,k}$ is an indicator function then $\mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} = \mathbb{E}[h_{t,k}|d_{t,k}]$. In addition, since $\sum_{t\in[T]} C_t \leq 2C$, it holds:

$$\sum_{t\in[T]} C_t \mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} = 2^{-k-1}\sum_{t\in[T]} C_t d_{t,k} \leq 2^{-k}C,$$

and with probability at least $1 - \delta/\log_2(T)$ noticing that $M = \log_2(T)$:

$$\begin{aligned}
&\sum_{t\in[T]} C_t d_{t,k} h_{t,k} - \sum_{t\in[T]} C_t \mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} \\
&\leq 2\sqrt{\sum_{t\in[T]} C_t^2 d_{t,k}\mathbb{E}[h_{t,k}|d_{t,k}]\log\left(\frac{\log_2(T)}{\delta}\right)} + |X||A|\log\left(\frac{\log_2(T)}{\delta}\right) && \text{(31a)}\\
&\leq 2\sqrt{|X||A|\sum_{t\in[T]} C_t d_{t,k}\mathbb{E}[h_{t,k}|d_{t,k}]\log\left(\frac{\log_2(T)}{\delta}\right)} + |X||A|\log\left(\frac{\log_2(T)}{\delta}\right) && \text{(31b)}\\
&\leq \sum_{t\in[T]} C_t \mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} + 2|X||A|\log\left(\frac{\log_2(T)}{\delta}\right), && \text{(31c)}
\end{aligned}$$

where Inequality (31a) holds with probability at least $1 - \delta/\log(T)$ by Freedman inequality, Inequality (31b) holds since $C_t \leq |X||A|$, and Inequality (31c) holds by AM-GM inequality. Therefore, it holds simultaneously for all $k \in [M]$:

$$\begin{aligned}
\sum_{t\in[T]} C_t d_{t,k} h_{t,k} &\leq 2\sum_{t\in[T]} C_t \mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} + 2|X||A|\log\left(\frac{\log_2(T)}{\delta}\right)\\
&\leq 2^{-k+1}C + 2|X||A|\log\left(\frac{\log_2(T)}{\delta}\right) = \theta_k,
\end{aligned}$$

with probability at least $1-\delta$, so $\mathbb{P}\left(\bigcup_{k\in[M]}\{\sum_{t\in[T]} C_t d_{t,k}h_{t,k} > \theta_k\}\right) \leq \delta$. Moreover, notice that with probability at least $1-p-2\delta$ thanks to the definition of corruption robust and Azuma-Hoeffding inequality, it holds simultaneously for all k:

$$\begin{aligned}
&\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)d_{t,k}\\
&= \frac{1}{2^{-k-1}}\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)2^{-k-1}d_{t,k}\\
&= \frac{1}{2^{-k-1}}\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)d_{t,k}\mathbb{E}[h_{t,k} \mid d_{t,k}]\\
&= \frac{1}{2^{-k-1}}\left(\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)d_{t,k}\left(\mathbb{E}[h_{t,k} \mid d_{t,k}] - h_{t,k}\right) + \sum_{t\in[T]} \overline{r}^\top (q^* - q_t)d_{t,k}h_{t,k}\right)\\
&\leq \frac{1}{2^{-k-1}}\left(L\sqrt{2\ln\left(\frac{\log_2(T)}{\delta}\right)\sum_{t\in[T]} d_{t,k}} + \sqrt{\beta_1\sum_{t\in[T]} d_{t,k}} + (\beta_2 + \beta_3\theta_k)\max_{t\in[T]} d_{t,k}\right)\\
&\leq \mathcal{O}\left(\frac{1}{2^{-k-1}}\left(\left(\sqrt{\beta_1} + L\sqrt{\log\left(\frac{T}{\delta}\right)}\right)\sqrt{T}\max_{t\in[T]} d_{t,k} + (\beta_2 + \beta_3\theta)\max_{t\in[T]} d_{t,k}\right)\right),
\end{aligned}$$

noticing that $\mathbb{E}\left[d_{t,k}\left(\mathbb{E}\left[h_{t,k}|d_{t,k}\right] - h_{t,k}\right)\right] = \mathbb{E}\left[h_{t,k}|d_{t,k}\right] - \mathbb{E}[h_{t,k}]d_{t,k} = \mathbb{E}\left[h_{t,k}|d_{t,k}\right] - \mathbb{E}\left[h_{t,k}|d_{t,k}\right] = 0$, since the expectation is taken w.r.t. the randomization of Algorithm 4 and the distribution generated given the external probability of receiving feedback w_t.

To conclude with probability at least $1 - p - 2\delta$:

$$\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)\mathbb{I}\left(w_t \geq \frac{1}{T}\right)$$

$$
\begin{aligned}
&\leq \sum_{k\in[M]} \sum_{t\in[T]} \overline{r}^{\top}(q^* - q_t) d_{t,k} \\
&\leq \mathcal{O}\left(\sqrt{\beta_1 T} \max_{t\in[T]} \frac{1}{w_t} + (\beta_2 + \beta_3 |X||A| \log(^{\log(T)}/_{\delta})) \max_{t\in[T]} \frac{1}{w_t} + \beta_3 \log(T) C\right) \\
&\leq \mathcal{O}\left(\left(\sqrt{\beta_1' T} + \beta_2'\right) \nu_T + \beta_3' C\right),
\end{aligned}
$$

with $\sqrt{\beta_1} \geq \mathcal{O}(L\sqrt{\log(^{T}/_{\delta})})$. Notice that the analogous reasoning can be applied to the positive cumulative constraints violation with parameters $\beta_4, \beta_5, \beta_6$. □

Algorithm 4 Adapted `STABILIZE` Jin et al. [2024]

Require: $C, \delta \in (0, 1)$

1: Initialize $M = \log_2(T)$ instance of Algorithm 2, each instance $k \in [M]$ initialized with corruption parameter:

$$
\theta_k := 2^{-k+1} C + 2|X||A| \log\left(\frac{\log_2(T)}{\delta}\right)
$$

2: **for** $t \in [T]$ **do**
3: Observe w_t, probability of receiving feedback.
4: **if** $w_t > \frac{1}{T}$ **then**
5: Let k_t be such that $w_t \in (2^{-k_t - 1}, 2^{-k_t}]$
6: Choose π_t as policy proposed by instance k_t
7: If the algorithm receives feedback send it to instance k_t with probability $\frac{2^{-k_t-1}}{w_t}$
8: **if** $w_t \leq \frac{1}{T}$ **then**
9: Propose random policy π_t

Corollary 6. *Being j^* such that $C \in (2^{j^*-1}, 2^{j^*}]$ then with probability at least $1 - 11\delta$ it holds:*

$$
\max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[g_{t,i}]^{\top} q_t^{j^*} - \alpha_i\right]^+ \leq \sqrt{\beta_4 T} \nu_{T,j^*} + \beta_5 \nu_{T,j^*} + 2\beta_6 C,
$$

with $\sqrt{\beta_4} = \mathcal{O}\left(L|X|\sqrt{|A| \ln(^{mT|X||A|}/_{\delta})}\right)$, $\beta_5 = \mathcal{O}\left(|X|^2|A|^2 \log(T) \log\left(^{\log(T)}/_{\delta}\right)\right)$ and $\beta_6 = \mathcal{O}\left(\ln(T)^2 |X||A|\right)$.

Corollary 7. *Being j^* such that $C \in (2^{j^*-1}, 2^{j^*}]$ then with probability at least $1 - 11\delta$ it holds:*

$$
\sum_{t\in[T]} \overline{r}^{\top}(q^* - q_t^{j^*}) \leq \sqrt{\beta_1 T} \nu_{T,j^*} + \beta_2 \nu_{T,j^*} + 2\beta_3 C,
$$

where $\sqrt{\beta_1} = \mathcal{O}\left(L|X|\sqrt{|A| \ln(^{T|X||A|}/_{\delta})}\right)$, $\beta_2 = \mathcal{O}\left(|X|^2|A|^2 \log(T) \log\left(^{\log(T)}/_{\delta}\right)\right)$ and $\beta_3 = \mathcal{O}\left(\ln(T)^2 |X||A|\right)$.