POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Dealing with non-stationarity in Constrained Markov Decision Processes

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Anna Lunghi, 991932

Advisor:
Prof. Nicola Gatti

Co-advisors:
Francesco Emanuele Stradi
Prof. Matteo Castiglioni
Prof. Alberto Marchesi

Academic year:
2023-2024

Abstract: In *constrained Markov decision processes* (CMDPs) with *adversarial* rewards and constraints, a well-known impossibility result prevents any algorithm from attaining both sublinear regret and sublinear constraint violation, when competing against a best-in-hindsight policy that satisfies constraints on average. In this thesis, we show that this negative result can be eased in CMDPs with *non-stationary* rewards and constraints, by providing algorithms whose performances smoothly degrade as non-stationarity increases. Specifically, we propose algorithms attaining $\tilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *positive* constraint violation under *bandit* feedback, where C is a corruption value measuring the environment non-stationarity. This can be $\Theta(T)$ 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 show how to obtain the same results by embedding such an algorithm in a general *meta-procedure*. This is of independent interest, as it can be applied to *any* non-stationary constrained online learning setting. Finally we design an algorithm that under the same condition, with *known* C, attains $\tilde{\mathcal{O}}(\sqrt{T})$ regret and $\tilde{\mathcal{O}}(\sqrt{T} + C)$ positive constraint violation.

Key-words: Online Learning, Markov Decision Processes, Constrained Markov Decision Processes, Adversarial Online Learning, Non-stationarity

1. Introduction

1.1. Reinforcement Learning, MDPs, CMDPs and applications

Artificial intelligence and machine learning have become increasingly prevalent in recent years. In particular, reinforcement learning has seen significant growth. As stated by [31], reinforcement learning problems involve learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The agent's actions influence the environment, and the learner is not told which actions to take, as in many forms of machine learning, but instead must discover which actions yield the most reward by trying them out. These three characteristics—being closed-loop in an essential way, not having direct instructions on which actions to take, and having consequences of actions, including reward signals, play out over extended time periods—are the most important distinguishing features of reinforcement learning problems.

This definition of reinforcement learning encompasses a large part of the human experience. With the advent of robotics, automatic systems, and similar technologies, it is becoming increasingly predominant in technological advancements. The applicability of reinforcement learning is vast, ranging from daily use in recommendation systems to life-saving medical procedures, from the everyday use of autonomous car driving to exceptional scientific endeavors like rovers on Mars.

Markov Decision Processes (MDPs) ([24]) are a useful representation of sequential decision-making by an agent. This model includes several important assumptions, such as the feedback on the agent's utility (it can be a loss or a reward) being immediately available, although not necessarily complete. Additionally, what happened in past episodes and the history of the states visited before the current state during the same episode do not influence the agent's utility. The goal of the agent is to maximize (or minimize) their cumulative reward (or losses). Taking, for example, the definition of [31], the MDP framework encompasses all cases of goal-directed decision making that can be reduced to three signals passing back and forth between the agent and the environment: one signal is actions and represents the choices of the agent, the second signal is states and represents the basis on which the choice is made, and the third signal is rewards that define the agent's goal.
This framework, however, does not always contain enough information to represent the reality of the decision-making problem. For example, a notable component this type of framework leaves out is the limitations the agent has to respect when making choices. More often than not, it would be helpful in real applications to introduce a fourth signal that represents the limitations within which the agent makes their decisions. Constrained Markov Decision Processes (CMDPs) do exactly that.

Constrained Markov Decision Processes (CMDPs) [2] are an extension of classical MDPs, where the agent, aside from maximizing their utility, must keep the incurred cost under a certain threshold. This cost depends on the agent's decisions similarly to losses or rewards. This problem setting allows for the inclusion of safety constraints in the decision-making environment, which can be crucial in real-world applications. Examples of fields where CMDPs are valuable include autonomous driving (*e.g.* [16, 35]), where it is imperative to avoid collisions, recommendation systems that must not suggest offensive materials (*e.g.* [28]), bidding agents in auctions, where they must not deplete their budget (*e.g.* [15, 36]), robots with limited battery capacity, medical procedures, and many others. In many of these cases, the environment is hardly static but can be better represented by introducing non-stationarity. This thesis will deal with non-stationary CMDPs.

1.2. Related works

Online Learning in MDPs The literature on online learning problems [10] in MDPs is wide (see [3, 13, 23] 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. [4] 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 [3]. Finally, online learning in MDPs has been studied form the configurator perspective, namely, the agent whose actions are different possible transition functions of the underlying environment (see, [21]).

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, [26] 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 [27] 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 [17] by providing an algorithm that achieves in the same setting a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$. Furthermore, [6] study adversarial MDPs where the bandit feedback depends on a different agent. Related to the non-stationarity of the transitions , [32] 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, [18] 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 [20], who show that it is impossible to suffer from sublinear regret and sublinear constraint violation when an adversary chooses losses and constraints.

[19] 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. [8, 9] 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, [37] 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. [34] 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. [7] 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. [12] 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. [25] 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. [29] 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. [14] 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, [11] and [33] 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 [11] and [33] 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 [33] hold only for large T. Finally, [5] study a generalization of stochastic CMDPs with partial observability on the constraints.

1.3. Goal and Original contributions

The goal of this thesis is to develop an algorithm that performs on non-stationary CMDPs attaining performance metrics - cumulative regret and positive cumulative constraints violation (for the formal definition see Section 2.2)- which are optimal in the stationary case (both $\tilde{\mathcal{O}}(\sqrt{T})$) and that scales smoothly with the non-stationarity, *i.e.* $\tilde{\mathcal{O}}(\sqrt{T} + C)$, with C measure of non-stationaity. Notice that this type of bound is in the worst case linear, which is perfectly coherent with a famous impossibility result of [20] that states that it is not possible to achieve both regret and constraints violation that are sub-linear in T when both rewards and costs are adversarial. In this thesis we propose first an algorithm designed to operate in environments where the extent of adversarial corruption (measure of non-stationarity) is known, then we extend the results to an algorithm that provably achieves similar results in the more challenging and broadly applicable scenario where the adversarial non-stationarity is completely unknown *a priori*. Finally, we present a modified version of the algorithm for uncostrained MDPs presented by [17] adapted to our constrained MDPs setting, tailored for situations with high non-stationarity in the reward vectors, which would render the initial algorithm ineffective. This algorithm, given prior knowledge of the corruption on the cost constraints, achieves sublinear regret and $\tilde{\mathcal{O}}(\sqrt{T} + C_G)$ constraint violations.
A preliminary version of the thesis can be found in the paper "Learning Constrained Markov Decision Processes With Non-stationary Rewards and Constraints" ([30]). In addition to the results presented in the paper, this thesis includes a minor result not covered in that work, which can be found in Section 5.

1.4. Thesis Structure

- Section 2 introduce the formal definition of the CMDPs, the notation and the performance metrics used

- Section 3 present an algorithm that deals with non-stationary CMDPs when the non-stationary C is known, with its theoretical results.
- Section 4 present an algorithm that deals with non-stationary CMDPs when the non-stationary C is *not* known, with its theoretical results.
- Section 5 present an algorithm that consider the rewards as purely adversarial when the non-stationary on the costs C_G is known with its theoretical results.
- Section 6 contains all the results and future work.

2. Preliminaries

We study *episodic constrained* MDPs [2] (CMDPs), in which a learner interacts with an unknown environment over T episodes, with goal of maximizing long-term rewards subject to some constraints. The CMDPs are defined as tuples $M = \left(X, A, P, \{\mathcal{R}_t\}_{t=1}^T, \{\mathcal{G}_t\}_{t=1}^T, \alpha\right)$. X is a finite set of states of the environment, A is a finite set of actions available to the learner in each state, $P : X \times A \times X \to [0, 1]$ is the transition function, with $P(x'|x, a)$ denoting the probability of going from state $x \in X$ to $x' \in X$ by taking action $a \in A$.Without loss of generality this thesis considers *loop-free* CMDPs, *i.e.* X is partitioned into L layers $X_0, \dots, X_L$ such that the first and the last layers are singletons ($X_0 = \{x_0\}$ 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 $k \in [0 \dots L]$. At each episode $t \in [T]$,[1] 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 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. 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 follows:

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

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 [18]. 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:

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

We let $g^\circ \in [0, 1]^{|X \times A|}$ be the constraint cost vector that attains the minimum in the definition of C_G. Finally, we introduce the total adversarial corruption C, which is defined as $C \coloneqq \max\{C_G, C_r\}$.In this thesis we derive all results in terms of the total adversarial corruption C.

[1] In this thesis, $[a \dots b]$ denotes 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.

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 $x_{k+1} \sim P(\cdot|x_k, a_k)$
end for

2.1. Occupancy measures

We introduce here *occupancy measures*, following the notation by [27]. Given a transition function P and a policy π, the occupancy measure $q^{P,\pi} \in [0,1]^{|X \times A \times X|}$ induced by P and π is such that, for every $x \in X_k$, $a \in A$, and $x' \in X_{k+1}$ with $k \in [0 \dots L-1]$:

$$q^{P,\pi}(x,a,x') \coloneqq \mathbb{P}[x_k = x, a_k = a, x_{k+1} = x' | P, \pi], \tag{3}$$

which represents the probability that, under P and π, the learner reaches state x, plays action a, and gets to the next state x'. Moreover, we also define the following quantities:

$$q^{P,\pi}(x,a) \coloneqq \sum_{x' \in X_{k+1}} q^{P,\pi}(x,a,x') \quad \text{and} \quad q^{P,\pi}(x) \coloneqq \sum_{a \in A} q^{P,\pi}(x,a). \tag{4}$$

The following lemma characterizes when a vector $q \in [0,1]^{|X \times A \times X|}$ is a *valid* occupancy measure.

Lemma 1 ([26]). *A vector $q \in [0,1]^{|X \times A \times X|}$ is a valid occupancy measure of an episodic loop-free CMDP if and only if it satisfies the following conditions:*

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

where P is the transition function of the CMDP and P^q is the one induced by q (see Equation (5)).

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

$$P^q(x'|x,a) = \frac{q(x,a,x')}{q(x,a)} \quad \text{and} \quad \pi^q(a|x) = \frac{q(x,a)}{q(x)}. \tag{5}$$

2.2. Performance metrics to evaluate learning algorithms

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 (6)), 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 $\alpha \in [0, L]^m$.

$$\mathrm{OPT}_{r,G,\alpha} \coloneqq \begin{cases} \max_{q \in \Delta(P)} & r^\top q \quad \text{s.t.} \\ & G^\top q \leq \alpha, \end{cases} \tag{6}$$

where $q \in [0,1]^{|X \times A|}$ is a vector encoding an occupancy measure, whose values are defined for state-action pairs according to Equation (4), and $\Delta(P)$ is the set of all valid occupancy measures given the transition function P (this set can be encoded by linear constraints thanks to Lemma 1).

We also introduce a problem-specific *feasibility parameter* related to Program (6), denoted by $\rho \in [0, L]$ and formally defined as $\rho \coloneqq \sup_{q \in \Delta(P)} \min_{i \in [m]} \left[\alpha - G^\top q\right]_i$.[2] Intuitively, ρ represents by how much feasible solutions to Program (6) strictly satisfy the constraints.

We say that an instance of Program (6) satisfies *Slater's condition* if the following holds.

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

Condition 2.1 (Slater). *There exists an occupancy measure $q^\circ \in \Delta(P)$ such that $G^\top q^\circ < \alpha$.*

Notice that Slater Condition (Condition 2.1) is equivalent to say that ρ, the *feasibility parameter*, is strictly greater than 0.

We are now ready to introduce the notion of *(cumulative) regret* and *positive (cumulative) constraints violation*, which are the performance metrics that we use to evaluate our learning algorithm. In particular the cumulative regret measures the cumulative difference in terms of rewards from the optimal static solution, while positive cumulative constraints violation measures the the violation on the costs constraints. We define the cumulative regret over T episodes as follows:

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

where $\overline{r} := \frac{1}{T}\sum_{t\in[T]} \mathbb{E}[r_t]$ and $\overline{G} := \frac{1}{T}\sum_{t\in[T]} \mathbb{E}[G_t]$. In the following, we denote by q^* an occupancy measure solving Program (6) instantiated with $\overline{r}, \overline{G}, \alpha$, while its corresponding policy (computed by Equation (5)) is π^*. Thus, $\mathrm{OPT}_{\overline{r},\overline{G},\alpha} = \overline{r}^\top q^*$ and the regret reduces to $R_T := \sum_{t\in[T]} \mathbb{E}[r_t]^\top (q^* - q^{P,\pi_t})$. Furthermore, we define the positive cumulative constraint violation over T episodes as follows:

$$V_T := \max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[G_t]^\top q^{P,\pi_t} - \alpha\right]_i^+,$$

where $[\cdot]^+ := \max\{0, \cdot\}$. Notice that, in our definition of V_T, constraint violations are *not* allowed to cancel out across episodes. This is a much more demanding performance metric than those employed by [11, 33], which instead allow for cancellations. For ease of notation, we compactly refer to q^{P,π_t} as q_t, thus omitting the dependency on P and π.

2.3. Stochastic, adversarial and corruption robust, a mixed framework

Our setting encompasses both purely stochastic and purely adversarial scenarios. It is interesting to study how the definitions of cumulative regret and positive cumulative constraint violation in our case relate to those defined in the literature for stochastic and adversarial cases.

First, it is obvious that when $C = 0$, the rewards and costs are completely stochastic, and R_T and V_T collapse to the definitions of regret and violation for the stochastic case. However, in general, our definition is more similar to the adversarial case since it is computed with respect to the best policy in hindsight. When the adversary chooses a singleton distribution in each episode, our definitions of regret and positive constraint violation become identical to those commonly used for the adversarial setting (*e.g.* [14]).

In MDP literature, aside from stochastic and adversarial settings, there is another setting known as the corruption-robust one. In the corruption-robust field, a stochastic MDP is subjected to adversarial corruption, making it non-stationary. Note that our definition of non-stationarity through adversarial corruption is inspired by this research field. However, in the corruption-robust literature, performance is computed with respect to the uncorrupted MDP, which leads to weaker results than those achieved in this thesis.

Finally, it is interesting to underline that the optimal occupancy measure in hindsight q^* has to satisfy the constraints only on average and not in each episode, which is a stronger benchmark to confront.

3. CMDPs with *known* corruption

Section overview *In this section we will introduce the pseudo-code for the* non-stationary safe optimistic policy search *(`NS-SOPS` for short) algorithm, and we will prove that such algorithm, given previous knowledge of the adversarial corruption achieves both regret and positive constraints violation $\tilde{\mathcal{O}}(\sqrt{T} + C)$. Finally we will analyze what happens to the performance of the algorithm when the adversarial corruption with which it is initialized is only a guess on the true value of the corruption and it is not precise.*

Algorithm *non-stationary safe optimistic policy search* (`NS-SOPS`), of which the pseudo-code is reported below, is based on the idea of performing a simple linear optimization employing enough optimism on the estimator to compensate for the uncertainty of the environment. In employing more optimism than the one usually applied for the cost matrices in literature allows us to state that with high probability the optimization space contains the optimal solution q^*.

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 [17]. 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 $x' \in X$, and $N_t(x, a)$ the total number of

visits to the state-action $(x, a) \in X \times A$ up to episode t (excluded), the estimated transition probability at t for (x, a, x') is:

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

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

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

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

Instead, for rewards and constraint costs, the algorithm adopts novel *enlarged* confidence bounds, which are suitably designed to tackle non-stationarity.
At each episode $t \in [T]$, for any state-action pair $(x, a) \in X \times A$ and constraint $i \in [m]$, these are defined as follows:

$$\widehat{r}_t(x, a) \coloneqq \frac{\sum_{\tau \in [t]} \mathbb{I}_\tau(x, a) r_\tau(x, a)}{\max\{N_t(x, a), 1\}} \quad \text{and} \quad \widehat{g}_{t,i}(x, a) \coloneqq \frac{\sum_{\tau \in [t]} \mathbb{I}_\tau(x, a) g_{\tau,i}(x, a)}{\max\{N_t(x, a), 1\}},$$

where $\mathbb{I}_\tau(x, a) = 1$ if and only if (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 $\widehat{g}_{t,i}(x, a)$. Given $\delta \in (0, 1)$, the confidence bound for the reward $r_t(x, a)$ is:

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

while the confidence bound for the constraint costs $g_{t,i}(x, a)$ is defined as:

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

Algorithm 2 Non-stationary safe optimistic policy search (`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 $q \in \Delta(\mathcal{P}_t)$
  end if
  $\pi_{t+1} \leftarrow \pi^{\widehat{q}_{t+1}}$
end for

Algorithm 2 applies an UCB-like approach including *optimism* in both rewards and constraints satisfaction, following an approach similar to that employed by [12]. 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:

$$\texttt{OPT-CB}_{\Delta(\mathcal{P}_t),\overline{r}_t,\underline{G}_t,\alpha} \coloneqq \begin{cases} \arg\max_{q \in \Delta(\mathcal{P}_t)} & \overline{r}_t^\top q \quad \text{s.t.} \\ & \underline{G}_t^\top q \leq \alpha, \end{cases} \tag{7}$$

3.1. Validity of confidence sets

In this section we show some properties of the confidence interval used in Algorithm `NS-SOPS`.
First, for what relates to the transition probabilities we used the confidence sets first proposed by [17] for which the results of Lemma 19 and Lemma 20 hold. We refer to Appendix B for the complete statement. Here what is of most important to notice is that, defining the event $\mathcal{E}_P$ as the intersection for all episode $t \in [T]$ of the events $P \in \mathcal{P}_t$, then $\mathcal{E}_P$, given $\delta \in (0,1)$ use to build $\mathcal{P}_t$, is characterized by a probability at least $1 - 4\delta$.

Now we introduce the result on the confidence sets on costs and rewards that are original contributions of this work. 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:*

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

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:

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

Notice that the proof holds for all $i \in [m]$ since g° and C_G are 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 $1 - \delta$:*

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

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

$$\begin{aligned}\left|\widehat{g}_{t,i}(x,a) - g^{\circ}(x,a)\right| &= \left|\left(\frac{\sum_{\tau\in[t]}\mathbb{I}_{\tau}(x,a)g_{\tau,i}(x,a)}{\max\{N_t(x,a),1\}}\right) - g^{\circ}(x,a)\right| \\ &\leq \left|\frac{1}{\max\{N_t(x,a),1\}}\sum_{\tau\in[t]}\mathbb{I}_{\tau}(x,a)\left(g_{\tau,i}(x,a) - \mathbb{E}[g_{\tau,i}(x,a)]\right)\right| \\ &\quad + \left|\frac{1}{\max\{N_t(x,a),1\}}\sum_{\tau\in[t]}\mathbb{I}_{\tau}(x,a)[\mathbb{E}[g_{\tau,i}(x,a)] - g^{\circ}(x,a)]\right|, \end{aligned} \tag{8}$$

where we employed the triangle inequality and the definition of $\widehat{g}_{t,i}(x,a)$.
We bound the two terms in Equation (8) 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:

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

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 $1-\delta$:*

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

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 $(x,a) \in X \times A$:

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

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

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

Thus, we have:

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

where Inequality (9) 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 $1-\delta$:

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

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:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,t}\left\{\left|\widehat{r}_t(x,a)-r^\circ(x,a)\right|\le\psi_t(x,a)\right\}\right]\\
&\qquad=1-\mathbb{P}\left[\bigcup_{x,a,t}\left\{\left|\widehat{r}_t(x,a)-r^\circ(x,a)\right|\le\psi_t(x,a)\right\}^c\right]\\
&\qquad\ge 1-\sum_{x,a,t}\mathbb{P}\left[\left\{\left|\widehat{r}_t(x,a)-r^\circ(x,a)\right|\le\psi_t(x,a)\right\}^c\right] \qquad (10)\\
&\qquad\ge 1-|X||A|T\delta',
\end{aligned}$$

where Inequality (10) holds by Union Bound. Noticing that $r_t(x,a)\le 1$, 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 $1-\delta$:

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

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 $1-\delta$:*

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

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

$$\begin{aligned}
&\left|\widehat{g}_{t,i}(x,a)-\frac{1}{T}\sum_{\tau\in[T]}\mathbb{E}[g_{\tau,i}(x,a)]\right|\\
&\qquad\le\left|\widehat{g}_{t,i}(x,a)-g^\circ(x,a)\right|+\left|g^\circ(x,a)-\frac{1}{T}\sum_{t\in[T]}\mathbb{E}[g_{t,i}(x,a)]\right|\\
&\qquad\le\sqrt{\frac{1}{2\max\{N_t(x,a),1\}}\ln\left(\frac{2mT|X||A|}{\delta}\right)}+\frac{C_G}{\max\{N_t(x,a),1\}}+\frac{C_G}{T},
\end{aligned}$$

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 $\widehat{C}\ge C_G$ 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:*

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

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

$$\begin{aligned}
&\left|\widehat{g}_{t,i}(x,a) - \frac{1}{T}\sum_{\tau\in[T]}\mathbb{E}[g_{\tau,i}(x,a)]\right| \\
&\qquad \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}}\ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_t(x,a),1\}} + \frac{C_G}{T} \\
&\qquad \leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}}\ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{\widehat{C}}{\max\{N_t(x,a),1\}} + \frac{\widehat{C}}{T},
\end{aligned}$$

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:*

$$\begin{aligned}
\mathcal{E}_G := &\left\{\left|\widehat{g}_{t,i}(x,a) - g^\circ(x,a)\right| \leq \xi_t(x,a),\ \forall (x,a)\in X\times A, \forall t\in[T], \forall i\in[m]\right\} \quad \bigcap \\
&\left\{\left|\widehat{g}_{t,i}(x,a) - \frac{1}{T}\sum_{\tau\in[T]}\mathbb{E}[g_{\tau,i}(x,a)]\right| \leq \xi_t(x,a),\ \forall (x,a)\in X\times A, \forall t\in[T], \forall i\in[m]\right\},
\end{aligned}$$

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:*

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

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

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 $\widehat{C} \ge C_r$.

3.2. Theoretical guarantees of NS-SOPS

To prove the theoretical guarantees of Algorithm NS-SOPS we decompose the proof in two main part: first prove that the algorithm is optimistic enough in its confidence set to include in the optimization-space the optimal occupancy-measure and second prove that the optimism is limited enough such that the error linked to the excess of optimism concentrates to a reasonable bound.
First we state and prove that the optimization spaces chosen are large enough.

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 $\left\{q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \le \alpha\right\}$ admits a feasible solution and it holds:*

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

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

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

Theorem 3.1. *Given any $\delta \in (0,1)$, with probability at least $1-9\delta$, Algorithm 2 attains:*

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

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 $t \in [T]$:

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

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

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

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

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

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

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

By Lemma 19 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:

$$\sum_{t\in[T]} \phi_{t-1}^\top q_t \leq \sum_{t\in[T]} \sum_{x,a} \phi_{t-1}(x,a)\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} \tag{11a}$$

$$\leq \sum_{x,a}\sum_{t\in[T]} \mathbb{I}_t(x,a)\left(\sqrt{\frac{1}{2\max\{N_{t-1}(x,a),1\}}\ln\left(\frac{2T|X||A|}{\delta}\right)} + \right.$$
$$\left. + \frac{C_r}{\max\{N_{t-1}(x,a),1\}} + \frac{C_r}{T} \right) + L\sqrt{2T\ln\frac{1}{\delta}} \tag{11b}$$

$$\leq \sqrt{\frac{1}{2}\ln\left(\frac{2T|X||A|}{\delta}\right)} \sum_{x,a}\sum_{t\in[T]} \mathbb{I}_t(x,a)\sqrt{\frac{1}{\max\{N_{t-1}(x,a),1\}}} +$$
$$+ C_r \sum_{x,a}\sum_{t\in[T]} \left(\frac{\mathbb{I}_t(x,a)}{\max\{N_{t-1}(x,a),1\}} + \frac{1}{T} \right) + L\sqrt{2T\ln\frac{1}{\delta}}$$

$$\leq 3|X||A|\sqrt{\frac{1}{2}T\ln\left(\frac{2T|X||A|}{\delta}\right)} + |X||A|(2+\ln(T))C_r + |X||A|C_r + L\sqrt{2T\ln\frac{1}{\delta}} \tag{11c}$$

$$\leq 3|X||A|\sqrt{\frac{1}{2}T\ln\left(\frac{2T|X||A|}{\delta}\right)} + (3+\ln(T))|X||A|C_r + L\sqrt{2T\ln\frac{1}{\delta}}$$

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

where Inequality (11a) follows from Azuma-Hoeffding inequality, Equality (11b) holds for the definition of ϕ_t, and Inequality (11c) holds since $1 + \sum_{t=1}^{N_T(x,a)} \sqrt{\frac{1}{t}} \leq 1 + 2\sqrt{N_T(x,a)} \leq 3\sqrt{N_T(x,a)}$ and $1 + \sum_{t=1}^{N_T(x,a)} \frac{1}{t} \leq 2 + \ln(T)$. Thus, we observe that with probability at least $1 - 9\delta$ it holds:

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

Employing Theorem 3 and the definition of C, which is at least equal to C_r, concludes the proof. □

We study now the positive constraints violation guarantees of Algorithm 2.

Theorem 3.2. *Given any $\delta \in (0,1)$, with probability at least $1 - 8\delta$, Algorithm 2 attains:*

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

Proof. In the following, we will refer as $\mathcal{E}_{\widehat{q}}$ to the event described in Lemma 20, which holds with probability at least $1-6\delta$. Thus, under $\mathcal{E}_G \cap \mathcal{E}_{\widehat{q}}$, it holds:

$$
\begin{aligned}
V_T &= \max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[G_t]^\top q_t - \alpha\right]_i^+ \\
&= \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t + g^{\circ\top} q_t - \alpha_i\right]^+ \\
&\le \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t + \left(\underline{g}_{t-1,i} + 2\xi_{t-1}\right)^\top q_t - \alpha_i\right]^+ && \text{(12a)} \\
&= \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + \underline{g}_{t-1,i}^\top \widehat{q}_t + 2\xi_{t-1}^\top q_t - \alpha_i\right]^+ \\
&\le \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + 2\xi_{t-1}^\top q_t\right]^+ && \text{(12b)} \\
&\le \max_{i\in[m]} \sum_{t\in[T]} \left|\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t\right| + 2\max_{i\in[m]} \sum_{t\in[T]} \left|\xi_{t-1}^\top q_t\right| + \max_{i\in[m]} \sum_{t\in[T]} \left|\underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t)\right| && \text{(12c)} \\
&\le \max_{i\in[m]} \sum_{t\in[T]} \left\|\mathbb{E}[g_{t,i}] - g^\circ\right\|_1 + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^\top q_t + \max_{i\in[m]} \sum_{t\in[T]} \left\|q_t - \widehat{q}_t\right\|_1 && \text{(12d)} \\
&\le C_G + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^\top q_t + \sum_{t\in[T]} \left\|q_t - \widehat{q}_t\right\|_1, && \text{(12e)}
\end{aligned}
$$

where Inequality (12a) follows from Corollary 2, Inequality (12b) holds since Algorithm 2 ensures, for all $t \in [T]$ and for all $i \in [m]$, that $\underline{g}_{t,i}^\top \widehat{q}_t \le \alpha_i$, Inequality (12c) holds since $[a+b]^+ \le |a|+|b|$, for all $a, b \in \mathbb{R}$, Inequality (12d) follows from Hölder inequality since $||\underline{g}_{t,i}(x,a)||_\infty \le 1$ and $||q_t(x,a)||_\infty \le 1$, and finally Equation (12e) holds for the definition of C_G.
To bound the last term of Equation (12e), we notice that, under $\mathcal{E}_{\widehat{q}}$, by Lemma 20, 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 (12e) we proceed as follows. Under $\mathcal{E}_{\widehat{q}}$,with probability at least $1-\delta$, it holds:

$$
\begin{aligned}
&\sum_{t\in[T]} \xi_{t-1}^\top q_t \le \sum_{t\in[T]} \sum_{x,a} \xi_{t-1}(x,a)\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} && \text{(13a)} \\
&\le \sum_{x,a}\sum_{t\in[T]} \mathbb{I}_t(x,a)\left(\sqrt{\frac{1}{2\max\{N_{t-1}(x,a),1\}}\ln\left(\frac{2mT|X||A|}{\delta}\right)} + \right. \\
&\qquad\qquad \left. + \frac{C_G}{\max\{N_{t-1}(x,a),1\}} + \frac{C_G}{T}\right) + L\sqrt{2T\ln\frac{1}{\delta}} && \text{(13b)} \\
&\le \sqrt{\frac{1}{2}\ln\left(\frac{2mT|X||A|}{\delta}\right)}\sum_{x,a}\sum_{t\in[T]} \mathbb{I}_t(x,a)\sqrt{\frac{1}{\max\{N_{t-1}(x,a),1\}}} + \\
&\qquad\qquad + C_G\sum_{x,a}\sum_{t\in[T]}\left(\frac{\mathbb{I}_t(x,a)}{\max\{N_{t-1}(x,a),1\}} + \frac{1}{T}\right) + L\sqrt{2T\ln\frac{1}{\delta}} \\
&\le 3|X||A|\sqrt{\frac{1}{2}T\ln\left(\frac{2mT|X||A|}{\delta}\right)} + |X||A|(2+\ln(T))C_G + |X||A|C_G + L\sqrt{2T\ln\frac{1}{\delta}} && \text{(13c)} \\
&\le 3|X||A|\sqrt{\frac{1}{2}T\ln\left(\frac{2mT|X||A|}{\delta}\right)} + (3+\ln(T))|X||A|C_G + L\sqrt{2T\ln\frac{1}{\delta}} \\
&= \mathcal{O}\left(|X||A|\sqrt{T\ln\left(\frac{mT|X||A|}{\delta}\right)} + \ln(T)|X||A|C_G\right),
\end{aligned}
$$

where Inequality (13a) follows from the Azuma-Hoeffding inequality and noticing that $\sum_{x,a} \xi_{t-1}(x,a) q_t(x,a) \leq L$, Equality (13b) follows from the definition of ξ_t and finally, Inequality (13c) holds since $1 + \sum_{t=1}^{N_T(x,a)} \sqrt{\frac{1}{t}} \leq 1 + 2\sqrt{N_T(x,a)} \leq 3\sqrt{N_T(x,a)}$ and $1 + \sum_{t=1}^{N_T(x,a)} \frac{1}{t} \leq 2 + \ln(T)$. 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. □

3.3. Theoretical result when the corruption is a guess

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:*

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

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:

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

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 $i \in [m]$:

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

which can be rewritten as:

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

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 3.3. *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 $1-8\delta$:*

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

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 3.1 concludes the proof. □

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

Theorem 3.4. *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 $1-9\delta$:*

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

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

$$\begin{aligned}
V_T &= \max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[G_t]^\top q_t - \alpha\right]_i^+ && \text{(14a)}\\
&= \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t + g^{\circ\top} q_t - \alpha_i\right]^+ \\
&\leq \max_{i\in[m]} \sum_{t\in[T]} \Bigg[\left(\mathbb{E}[g_{t,i}] - g^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + \underline{g}_{t-1,i}^\top \widehat{q}_t + 2\xi_{t-1}^\top q_t + \\
&\qquad + \sum_{x,a}\left(\frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} + \frac{\epsilon}{T}\right) q_t(x,a) - \alpha_i\Bigg]^+ && \text{(14b)}\\
&\leq C_G + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^\top q_t + \sum_{t\in[T]} \|q_t - \widehat{q}_t\|_1 + \\
&\qquad + \sum_{t\in[T]}\sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a) + \epsilon L, && \text{(14c)}
\end{aligned}$$

where Inequality (14b) follows from Lemma 10 and Inequality (14c) is derived as in the proof of Theorem 3.2, 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 $\sum_{t\in[T]}\sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a)$ as follows:

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

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 3.2 to bound the first 3 elements of Inequality (14c) 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 $1-9\delta$,

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

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 3.5. *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 $1-8\delta$:*

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

Proof. The proof follows by employing Corollary 1 to the proof of Theorem 3.2. □

4. CMDPs with *unknown* Corruption

Section overview *In this section we propose the pseudo-code for the meta-algorithm* `Lag-FTRL` *that, running different instances of* `NS-SOPS` *initiated with different guesses on the amount of adversarial corruption attains both regret and constraints violation* $\tilde{\mathcal{O}}(\sqrt{T}+C)$ *whenever the amount of adversarial corruption* C *is* not *known a priori. The algorithm proposed use a Lagrangian function on an FTRL mechanism. We will then state and prove the theoretical guarantees.*

At a high level, the *Lagrangified follow-the-regularized-leader* (`Lag-FTRL` for short) algorithm works by instantiating several different instances of Algorithm 2, 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 $\pi_t^{j_t}$ prescribed by $\texttt{Alg}^{j_t}$ and provides the observed feedback to instance $\texttt{Alg}^{j_t}$ only. This approach has also the advantage of lessen the computational burden of updating all the sub-algorithm at each episode. In fact `Lag-FTRL` updates after each episode only the `NS-SOPS` Algorithm which proposed policy has been selected at that episode.

Algorithm 3 Lagrangified follow-the-regularized-leader (`Lag-FTRL`)

Require: $\delta \in (0,1)$
1: $\Lambda \leftarrow \frac{Lm+1}{\rho}$, $M \leftarrow \lceil \log_2 T \rceil$
2: $\gamma \leftarrow \sqrt{\ln(M/\delta)/TM}$, $\eta \leftarrow \frac{1}{2Lm\Lambda}\sqrt{\ln(T)/T}$
3: **for** $j \in [M]$ **do**
4: $\quad \texttt{Alg}^j \leftarrow$ Algorithm 2 with $C = 2^j$
5: **end for**
6: $w_{1,j} \leftarrow 1/M$ for all $j \in [M]$
7: **for** $t \in [T]$ **do**
8: $\quad$ Sample index $j_t \sim w_t$
9: $\quad \pi_t^{j_t} \leftarrow$ policy that $\texttt{Alg}^{j_t}$ would choose
10: $\quad$ Choose policy $\pi_t^{j_t}$ in Algorithm 1 and observe feedback from interaction
11: $\quad$ Let $\texttt{Alg}^{j_t}$ observe received feedback
12: $\quad$ **for** $j \in [M]$ **do**
13: $\quad\quad$ Build $\ell_{t,j}$ as in Equation (15)
14: $\quad$ **end for**
15: $\quad w_{t+1} \leftarrow \arg\min_{w \in \Delta_M} w^\top \sum_{\tau \in [t]} \ell_t + \frac{1}{\eta} \sum_{j \in [M]} \frac{1}{w_j}$
16: **end for**

The pseudocode of the `Lag-FTRL` algorithm is provided in Algorithm 3. At Line 4, it instantiates $M \coloneqq \log_2 T$ instances of Algorithm 2, with each instance $\texttt{Alg}^j$, for $j \in [M]$, receiving as input a "guess" on the adversarial corruption $C = 2^j$. 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 $\frac{1}{M}$ (Line 6). 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 8), and it employs the policy $\pi_t^{j_t}$ prescribed by $\texttt{Alg}^{j_t}$ (Line 9). 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 11). Then, at Line 13, the algorithm builds an *optimistic* loss estimator for each instance $\texttt{Alg}^j$. In particular, at episode $t \in [T]$, for every $j \in [M]$ the optimistic loss estimator is defined as:

$$\ell_{t,j} \coloneqq \frac{\mathbb{I}(j_t = j)}{w_{t,j} + \gamma} \left(\sum_{k \in [0 \dots L-1]} \left(1 - r_t(x_k^t, a_k^t)\right) + \Lambda \sum_{i \in [m]} \left[\left(\widehat{g}_{t,i}^j\right)^\top \widehat{q}_t^j - \alpha_i \right]^+ \right), \tag{15}$$

where γ is a suitably-defined implicit exploration factor, $\{(x_k^t, a_k^t)\}_{k \in [0 \dots L-1]}$ is the sequence of state-action pairs visited at episode t, Λ is a suitably-defined upper bound on the optimal values of Lagrangian multipliers,[3] $\widehat{g}_{t,i}^j$

[3] Notice that, in the definition of Λ, ρ is the feasibility parameter of Program (6) for the reward vector $\overline{r}$, the constraint

is the empirical constraint cost built by 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 episode t. Finally, the algorithm updates the weight vector according to an FTRL update with a suitable log-barrier regularization (see Line 15 for the definition of the update).

4.1. Theoretical guarantees of Lag-FTRL

Next, we prove the theoretical guarantees attained by Algorithm 3 . As a first preliminary step, we extend the well-known strong duality result for CMDPs [2] to the case of bounded Lagrangian multipliers.

Lemma 11. *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 (6) satisfies Slater's condition (Condition 2.1), then the following holds:*

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

where $\lambda \in \mathbb{R}^m_{\geq 0}$ is a vector of Lagrangian multipliers and ρ is the feasibility parameter of Program (6).

Proof. The proof follows the one of Theorem 3.3 in [9]. 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:

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

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:

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

Thus,

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

where the second inequality holds by the *max-min* inequality and the last step holds by the well-known strong duality result in CMDPs [2]. This concludes the proof. □

Intuitively, Lemma 11 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 11 to the case of a Lagrangian function suitably-modified to encompass *positive* violations. We call it *positive Lagrangian* of Program (6), defined as follows.

Definition 1 (Positive Lagrangian)**.** *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 $\alpha \in [0,L]^m$, the* positive Lagrangian *of Program (6) is defied as a function $\mathcal{L} : \mathbb{R}_+ \times \Delta(P) \to \mathbb{R}$ such that it holds $\mathcal{L}(\beta, q) := r^\top q - \beta \sum_{i\in[m]} \left[G^\top q - \alpha\right]_i^+$ for every $\beta \geq 0$ and $q \in \Delta(P)$.*

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.

Notice that the offline optimization problem defining the positive Lagrangian does no admit Slater's condition, since, by definition of the $[\cdot]^+$ operator, it does not exist an occupancy measure such that q° s.t. $\left[G^\top q^\circ - \alpha\right]_i^+ \leq 0$ for any $i \in [m]$. Nevertheless, we show that a strong duality-kind of result still holds for $\mathcal{L}(\frac{L}{\rho}, q)$, when Slater's condition holds for the optimization problem when the $[\cdot]^+$ operator is not applied to the constraints, namely, Program (6). This is done in the following result.

Theorem 4.1. *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 (6) satisfies Slater's condition (Condition 2.1), then the following holds:*

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

where ρ is the feasibility parameter of Program (6).

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

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

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 11. 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 $\max_{q\in\Delta(P)} \mathcal{L}(1/\rho, q) \geq \mathrm{OPT}_{r,G,\alpha}$, concludes the proof. □

Theorem 4.1 intuitively shows that a L/ρ multiplicative factor on the positive constraint violation is enough to compensate the large rewards 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.

4.1.1 Preliminary results

In the following sections we will refer as:

$$\widehat{V}_T := \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j} + \gamma} \sum_{i\in[m]}\left[\widehat{g}_{t,i}^{\,j\top}\widehat{q}_t^{\,j} - \alpha_i\right]^+, \tag{16}$$

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

$$\widehat{V}_{T,j^*} := \sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*} + \gamma} \sum_{i\in[m]}\left[\widehat{g}_{t,i}^{\,j^*\top}\widehat{q}_t^{\,j^*} - \alpha_i\right]^+, \tag{17}$$

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:

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

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 $\mathbb{I}_t(x,a)\mathbb{I}(j_t = j) = 1$.
We start providing some preliminary results 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:*

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

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:

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

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 13. *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 [22] 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:

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

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 14. *For any $\delta\in(0,1)$ and properly setting the learning rate η such that $\eta\leq\frac{\rho}{2Lm(Lm+1)}$, Algorithm 3 attains, with probability at least $1-2\delta$:*

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

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

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

Since 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 $\eta\leq\frac{\rho}{2(L\rho+L^2m^2+Lm)}\leq\frac{\rho}{2(L^2m^2+Lm)}$, if $\eta\leq\frac{\rho}{2Lm(Lm+1)}$, Algorithm 3 attains, by Lemma 17 :

$$\begin{aligned}
&\sum_{t\in[T]}\left[\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))-\frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))\right]+\frac{Lm+1}{\rho}\widehat{V}_T\\
&\qquad\leq\frac{M\ln T}{\eta}+\eta\frac{TM(L\rho+L^2m^2+Lm)^2}{\rho^2}+\frac{Lm+1}{\rho}\widehat{V}_{T,j^*},
\end{aligned}\tag{18}$$

where we used that $\left(\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))+\frac{Lm+1}{\rho}\sum_{i\in[m]}\left[\widehat{g}_{t,i}^{\ j\top}\widehat{q}_t^{\ j}-\alpha_i\right]^+\right)\leq\frac{(L\rho+L^2m^2+Lm)}{\rho}$ and $\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}^2\ell_{t,j}^2\leq\frac{TM(L\rho+L^2m^2+Lm)^2}{\rho^2}$. Thus, with probability at least $1-2\delta$, it holds:

$$\begin{aligned}
&\sum_{t\in[T]}\mathbb{E}[r_t]^\top q_t^{j^*}-\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}\mathbb{E}[r_t]^\top q_t^j+\frac{Lm+1}{\rho}\widehat{V}_T\\
&=\sum_{t\in[T]}\sum_{j\in[M]}w_{t,j}\left(L-\mathbb{E}[r_t]^\top q_t^j\right)-\sum_{t\in[T]}\left(L-\mathbb{E}[r_t]^\top q_t^{j^*}\right)+\frac{Lm+1}{\rho}\widehat{V}_T\qquad(19)\\
&\leq\mathcal{O}\left(\frac{M\ln T}{\eta}+\eta\ m^4L^4TM+\gamma TLM+L\sqrt{T\ln\left(\frac{1}{\delta}\right)}+\frac{L}{\gamma}\ln\left(\frac{1}{\delta}\right)\right)+\frac{Lm+1}{\rho}\widehat{V}_{T,j^*},\qquad(20)
\end{aligned}$$

where Equation (19) holds since $\sum_{j\in[M]}w_{t,j}=1,\ \forall t\in[T]$, and Inequality (20) holds, with probability at least $1-2\delta$, by Lemma 12, Lemma 13 and Equation (18). 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 14. Thus, we bound the distance between the incurred violation and the estimated one.

Lemma 15. *For any $\gamma \in \mathbb{R}_{\geq 0}$, given $\delta \in (0,1)$, with probability at least $1-10\delta$, it holds:*

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

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 3.2 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,

$$\widehat{\xi}_{t,j}(x,a) := \min\left\{1, \sqrt{\frac{1}{2\max\{N_t^j(x,a),1\}}\ln\left(\frac{2mMT|X||A|}{\delta}\right)} + \frac{C}{\max\{N_t^j(x,a),1\}} + \frac{C}{T}\right\}, \tag{21}$$

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:

$$\widehat{g}_{t,i}^j(x,a) + \widehat{\xi}_{t,j}(x,a) \geq g(x,a). \tag{22}$$

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

$$\begin{aligned}
\widehat{V}_T &= \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\widehat{g}_{t,i}^{j\top} \widehat{q}_t^j - \alpha_i\right]^+ \\
&= \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[(\widehat{g}_{t,i}^{j\top} q_t^j + \widehat{\xi}_{t,j}^\top q_t^j - \alpha_i) - \widehat{\xi}_{t,j}^\top q_t^j - \widehat{g}_{t,i}^{j\top}(q_t^j - \widehat{q}_t^j)\right]^+ \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left(\left[(\widehat{g}_{t,i}^j + \widehat{\xi}_{t,j})^\top q_t^j - \alpha_i\right]^+ - \widehat{\xi}_{t,j}^\top q_t^j - \widehat{g}_{t,i}^{j\top}|q_t^j - \widehat{q}_t^j|\right) && \text{(23a)} \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left(\left[g^{\circ\top} q_t^j - \alpha_i\right]^+ - \widehat{\xi}_{t,j}^\top q_t^j - \|q_t^j - \widehat{q}_t^j\|_1\right) && \text{(23b)} \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left(\left[\mathbb{E}[g_{t,i}]^\top q_t^j - \alpha_i\right]^+ - \widehat{\xi}_{t,j}^\top q_t^j\right) - \sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \\
&\qquad \cdot \sum_{i\in[m]} \left[(g^\circ - \mathbb{E}[g_{t,i}])^\top q_t^j\right]^+ - \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), && \text{(23c)}
\end{aligned}$$

where Inequality (23a) holds since $[a-b]^+ \geq [a]^+ - b,\ a \in \mathbb{R}, b \in \mathbb{R}_{\geq 0}$, Inequality (23b) follows from Inequality (22) 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 (23c) holds under event $\mathcal{E}_{\widehat{q}}$ by Lemma 20 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:

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

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:

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

where we employed the Azuma-Hoeffding inequality and where the last step holds following the proof of Theorem 3.2.

Upper-bound to $\sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[(g^\circ-\mathbb{E}[g_{t,i}])^\top q_t^j\right]^+$.
We simply bound the quantity of interest as follows:

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

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

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

Thus, plugging the previous bounds in Equation (23c), we obtain, with probability at least $1-10\delta$:

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

This concludes the proof. □

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

Lemma 16. *For any $\delta \in (0,1)$, with probability at least $1-14\delta$, it holds:*

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

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:

$$
\begin{aligned}
&\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[\widehat{g}_{t,i}^{j^*\top} \widehat{q}_t^{j^*} - \alpha_i\right]^+ \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left(\left[\widehat{g}_{t,i}^{j^*\top}(\widehat{q}_t^{j^*} - q_t^{j^*}) + \widehat{g}_{t,i}^{j^*\top} q_t^{j^*} - \widehat{\xi}_{t,j^*}^\top q_t^{j^*} - \alpha_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*}\right) && \text{(24a)} \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \Bigg(\left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*} + \\
&\qquad\qquad + \left[g^{\circ\top} q_t^{j^*} - \mathbb{E}[g_{t,i}]^\top q_t^{j^*}\right]^+ + \|\widehat{q}_t^{j^*} - q_t^{j^*}\|_1\Bigg) && \text{(24b)} \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left(\left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*} + \left[(g^\circ - \mathbb{E}[g_{t,i}])^\top q_t^{j^*}\right]^+\right) \\
&\qquad\qquad + \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), && \text{(24c)}
\end{aligned}
$$

where Inequality (24a) holds since $[a+b]^+ \leq [a]^+ + [b]^+$, $\forall a,b \in \mathbb{R}$ and by the definition of $\widehat{\xi}_{t,j^*}$ (see Equation (21)) which implies that all its elements are positive, Inequality (24b) 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 (24c) holds with probability at least $1-6\delta$ by Lemma 20.

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

It is immediate to bound the quantity of interest employing the definition of corruption C and by Lemma 18. Indeed, with probability at least $1-\delta$:

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

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 $1 - 10\delta$, it holds:

$$\begin{aligned}
&\sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*} + \gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ \\
&\quad\leq \sum_{t\in[T]} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \alpha_i\right]^+ + L\frac{\ln\left(\frac{M}{\delta}\right)}{2\gamma} && \text{(25a)}\\
&\quad\leq mV_{T,j^*} + L\frac{\ln(\frac{M}{\delta})}{2\gamma} && \text{(25b)}\\
&\quad= \mathcal{O}\left(m|X||A|L\sqrt{T\ln\left(\frac{mT|X||A|}{\delta}\right)} + m\ln(T)|X||A|C + L\frac{\ln(\frac{M}{\delta})}{2\gamma}\right), && \text{(25c)}
\end{aligned}$$

where Inequality (25a) holds by Corollary 5 with probability at least $1-\delta$, and where V_{T,j^*} in Inequality (25b) is the positive cumulative violation of costs constraints that Algorithm 2 would attain when instantiated with value of corruption $C = 2^{j^*}$, if it were to run on its own, and thus Equality (25c) holds by Theorem 3.2 and Corollary 3.5 with probability at least $1 - 9\delta$ considering that, by definition of j^* , 2^{j^*} is at most $2C$.

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 18. Now we observe that, with probability at least $1-\delta$, it holds:

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

where employed the same steps as in the proof of Theorem 3.2, 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. □

4.1.2 Main Results

Theorem 4.2. *Given any $\delta \in (0,1)$, with probability at least $1 - 26\delta$, Algorithm 3 attains:*

$$R_T = \mathcal{O}\left(m^2L^2|X||A|\sqrt{T\ln\left(mT|X||A|/\delta\right)}\ln(T) + m^2L\ln(T)|X||A|C\right).$$

Proof. We first decompose the regret as follows:

$$\begin{aligned}
R_T &= \sum_{t\in[T]} \overline{r}^\top(q^* - q_t)\\
&= \sum_{t\in[T]} (\mathbb{E}[r_t] - \overline{r})^\top(q_t - q^*) + \sum_{t\in[T]} \mathbb{E}[r_t]^\top q^* - \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t\\
&\leq \sum_{t\in[T]} \|\mathbb{E}[r_t] - \overline{r}\|_1 + \sum_{t\in[T]} \left(\mathbb{E}[r_t]^\top q^* - \mathbb{E}[r_t]^\top q_t\right) && \text{(26a)}\\
&\leq 2C + \sum_{t\in[T]} \left(\mathbb{E}[r_t]^\top q^* - \mathbb{E}[r_t]^\top q_t\right). && \text{(26b)}
\end{aligned}$$

where Inequality (26a) holds since $|q_t(x,a) - q^*(x,a)| \leq 1,\ \forall (x,a) \in X \times A$, and where Inequality (26b) holds by definition of C. Moreover, by Lemma 14, with probability at least $1-2\delta$, it holds:

$$\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}[r_t]^\top q^* - \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t \\
&\quad = \sum_{t\in[T]} \mathbb{E}[r_t]^\top \left(q^* - q_t^{j^*}\right) + \sum_{t\in[T]} \mathbb{E}[r_t]^\top \left(q_t^{j^*} - q_t^{j_t}\right) \\
&\quad \leq R_T^{j^*} + \sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} \mathbb{E}[r_t]^\top q_t^j - \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{j_t} - \frac{Lm+1}{\rho}\widehat{V}_T + \frac{Lm+1}{\rho}\widehat{V}_{T,j^*} + \\
&\qquad\qquad + \mathcal{O}\left(\frac{M\ln T}{\eta} + \eta\, m^4 M L^4 T + \gamma T L M + \frac{L}{\gamma}\ln\left(\frac{1}{\delta}\right)\right), \qquad (27)
\end{aligned}$$

where Equation (27) holds by Lemma 14, with probability at least $1-2\delta$, and where $R_T^{j^*}$ is the regret $\texttt{Alg}^{j^*}$ would incur in. Now, employing the Azuma-Hoeffding inequality to $\sum_{t\in[T]}\sum_{j\in[M]} w_{t,j}\mathbb{E}[r_t]^\top q_t^j - \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{i_t}$, Theorem 3.1 to bound $R_T^{j^*}$ with high probability and Lemma 16 to bound $\widehat{V}_{T,j^*}$ with high probability, we observe that with probability at least $1-24\delta$, it holds:

$$\begin{aligned}
&R_T^{j^*} + \sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} \mathbb{E}[r_t]^\top q_t^j - \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{j_t} + \frac{Lm+1}{\rho}\widehat{V}_{T,j^*} \\
&\quad \leq \mathcal{O}\left(m^2L^2|X||A|\sqrt{T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + m^2 L\ln(T)|X||A|C + L^2 m\frac{\ln(\frac{M}{\delta})}{\gamma}\right). \qquad (28)
\end{aligned}$$

Finally, combining the previous results and by Union Bound, with probability at least $1-26\delta$, it holds:

$$\begin{aligned}
R_T + \frac{Lm+1}{\rho}\widehat{V}_T \leq \mathcal{O}\Bigg(&m^2L^2|X||A|\sqrt{T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + m^2 L\ln(T)|X||A|C + \\
&+ L^2 m\frac{\ln(\frac{M}{\delta})}{\gamma} + \frac{M\ln T}{\eta} + \eta m^4 M L^4 T + \gamma T L M\Bigg), \qquad (29)
\end{aligned}$$

which concludes the proof after observing that $\widehat{V}_T \geq 0$, by definition, and setting $\gamma = \sqrt{\frac{\ln(M/\delta)}{TM}}$, $\eta = \sqrt{\frac{\ln(T)}{T}}\frac{\rho}{2Lm(Lm+1)}$. □

Theorem 4.3. *Given any $\delta \in (0,1)$, with probability at least $1-30\delta$, Algorithm 3 attains:*

$$V_T = \mathcal{O}\left(m^2L^2|X||A|\sqrt{T\ln\left(\frac{mT|X||A|}{\delta}\right)}\ln(T) + m^2 L\ln(T)|X||A|C\right).$$

Proof. By Equation (29), which holds with probability at least $1-26\delta$, we obtain:

$$\begin{aligned}
R_T + \frac{Lm+1}{\rho}\widehat{V}_T \leq \mathcal{O}\Bigg(&m^2L^2|X||A|\sqrt{T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + m^2 L\ln(T)|X||A|C + \\
&+ L^2 m\frac{\ln(\frac{M}{\delta})}{\gamma} + \frac{M\ln T}{\eta} + \eta\, m^4 M L^4 T + \gamma T L M\Bigg).
\end{aligned}$$

In order to obtain the final violations bound, it is necessary to find an upper bound for $-R_T$. We proceed as follows,

$$\begin{aligned}
\overline{r}^\top q^* &= \mathrm{OPT}_{\overline{r},\overline{G},\alpha} && (30a)\\
&= \max_{q\in\Delta(P)} \left(\overline{r}^\top q - \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q - \alpha_i\right]^+\right) && (30b)\\
&\geq \overline{r}^\top q_t - \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t - \alpha_i\right]^+,
\end{aligned}$$

where Equality (30a) holds since q^* is the feasible occupancy that maximizes the reward vector $\overline{r}$ and Equality (30b) holds by Theorem 4.1 . 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:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t - \alpha_i\right]^+ \\
&\quad\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t - \alpha_i\right]^+ + \sum_{i\in[m]}\left[(\overline{G}_i - \mathbb{E}[g_{t,i}])^\top q_t\right]^+\right) && \text{(31a)}\\
&\quad\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t - \alpha_i\right]^+ + \sum_{i\in[m]}\left\|\overline{G}_i - \mathbb{E}[g_{t,i}]\right\|_1\right) && \text{(31b)}\\
&\quad\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t - \alpha_i\right]^+ + \sum_{i\in[m]}\left(\left\|\overline{G}_i - g^\circ\right\|_1 + \left\|g^\circ - \mathbb{E}[g_{t,i}]\right\|_1\right)\right)\\
&\quad\leq mV_T + 2mC, && \text{(31c)}
\end{aligned}$$

where Inequality (31a) holds since $[a+b]^+ \leq [a]^+ + [b]^+, a\in\mathbb{R}, b\in\mathbb{R}$, Inequality (31b) holds since $q_t(x,a)\leq 1 \forall t\in[T], \forall(x,a)\in X\times A$, and finally Inequality (31c) 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:

$$\begin{aligned}
R_T + \frac{Lm+1}{\rho}V_T &= \sum_{t\in[T]}\mathbb{E}[r_t]^\top(q^* - q_t) + \frac{Lm+1}{\rho}V_T\\
&= \sum_{t\in[T]}\left(\mathbb{E}[r_t] - \overline{r}\right)^\top(q^* - q_t) + \sum_{t\in[T]}\overline{r}^\top(q^* - q_t) + \frac{Lm+1}{\rho}V_T\\
&\geq -\sum_{t\in[T]}\left\|\mathbb{E}[r_t] - \overline{r}\right\|_1 + \sum_{t\in[T]}\overline{r}^\top(q^* - q_t) + \frac{Lm+1}{\rho}V_T && \text{(32a)}\\
&\geq -2C - \frac{L}{\rho}(mV_T + 2mC) + \frac{Lm+1}{\rho}V_T && \text{(32b)}\\
&= -2C - \frac{2LmC}{\rho} + V_T\left(\frac{Lm+1}{\rho} - \frac{Lm}{\rho}\right)\\
&= \frac{1}{\rho}V_T - \left(2C + \frac{2LmC}{\rho}\right), && \text{(32c)}
\end{aligned}$$

where Inequality (32a) 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 (32b) 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 (32c), we finally bound the cumulative violation as follows:

$$\begin{aligned}
V_T &\leq 2\rho C + 2LmC + \rho R_T + (Lm+1)V_T\\
&= 2\rho C + 2LmC + (Lm+1)\left(V_T - \widehat{V}_T\right) + \rho\left(R_T + \frac{Lm+1}{\rho}\widehat{V}_T\right)\\
&\leq \mathcal{O}\Bigg(m^2L^2|X||A|\sqrt{T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + m^2L\ln(T)|X||A|C +\\
&\qquad + L^2m\frac{\ln(\frac{M}{\delta})}{\gamma} + \frac{M\ln(T)}{\eta} + \eta m^4ML^4T + \gamma mTL^2M\Bigg),
\end{aligned}$$

where the last inequality holds by Equation (29) and by Lemma 15, with probability at least $1-4\delta$ under $\mathcal{E}_{\widehat{q}}$. Employing a Union Bound, setting $\gamma = \sqrt{\frac{\ln(M/\delta)}{TM}}$ and $\eta = \sqrt{\frac{\ln(T)}{T}}\frac{\rho}{2Lm(Lm+1)}$ concludes the proof. □

5. Adversarial loss and corrupted cost with known corruption

Section Overview *In this section we will consider a peculiar case in which the adversarial corruption on the rewards is expected to be almost linear and the corruption on the cost C_G is known* a priori. *First, we will*

expand on when and how this particular algorithm is useful, then we will state its theoretical guarantees.

The algorithm 3 (`Lag-FTRL`) performs well when the adversarial corruption value is low enough (sublinear, ideally $C = \tilde{\mathcal{O}}(\sqrt{T})$) and very poorly when the adversarial corruption is linear. This result, as stated in the introduction, is perfectly coherent with the impossibility result of [20].
In the field of non-stationary CMDPs, it may be useful to consider an environment where only one of the two adversarial corruptions is potentially linear, while the other remains sublinear. Specifically, we consider the case where the adversarial reward corruption is potentially linear and the adversarial corruption on the cost constraints is known (or at least an overestimation of it is known) and sublinear. In this section, we propose an algorithm that treats the rewards as purely adversarial and achieves a regret of $\tilde{\mathcal{O}}(\sqrt{T})$ and positive cumulative constraint violations of at most $\tilde{\mathcal{O}}(\sqrt{T} + C_G)$.
First, we expand on the setting considered in this section. The environment remains the same as in Section 1, with the only difference lying in how we represent the reward vectors. In the introduction, the rewards are analyzed and considered with respect to a theoretical non-corrupted reward vector r°, fixed in all episodes. In this section, we ignore the non-corrupted reward vector and consider the rewards in each episode as purely adversarial. Note that this is merely a difference in approach and perspective and does not impact the structure of the CMDPs.
In this scenario, we use an approach based on Online Mirror Descent (OMD), inspired by the unconstrained algorithm proposed by [17], which deals with adversarial losses, unknown transition probabilities, and bandit feedback. In this thesis, we consider the CMDPs problem as one of maximizing rewards, which can be easily converted to a problem of minimizing losses by taking the losses as $1 - r(x, a)$.
Our main contribution in this case lies in the choice of the optimization space for the occupancy measures. This space is adapted to the non-stationarity of the costs and, with high probability, contains the optimal occupancy measure at each episode, while still guaranteeing constraint violations of at most $\tilde{\mathcal{O}}(\sqrt{T} + C_G)$.
We define a loss estimator function $f_t(x, a)$ as :

$$f_t(x,a) := \frac{\mathbb{I}_t(x,a)}{u_t(x,a) + \gamma}(1 - r_t(x,a)), \tag{33}$$

where

$$u_t(x,a) := \sup_{P' \in \mathcal{P}_t} q^{\pi_t, P'}(x,a). \tag{34}$$

Algorithm 4 Bounded Violation Optimistic Policy Search for Non-stationary Constraints (known C_G)

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

$\eta, \gamma \leftarrow \sqrt{\frac{L \ln(\frac{L|X||A|}{\delta})}{T|X||A|}}$

Initialize , $\widehat{q}_1(x,a,x') \leftarrow \frac{1}{|X_k||A||X_{k+1}|}$, $\forall (x,a,x') \in X_k \times A \times X_{k+1},\ k \in [0, L-1]$

$\pi_1 \leftarrow \pi^{\widehat{q}_1}$

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

  Choose policy π_t in Algorithm 1 and observe feedback from interaction

  Compute $\mathcal{P}_t$, f_t, and $\underline{G_t}$

  $\widehat{q}_{t+1} \longleftarrow \underset{q \in \Delta(\mathcal{P}_t): \underline{G_t}^\top q \leq \alpha}{\arg\min} \sum_{x,a} f_t(x,a) q(x,a) + D(q||\widehat{q}_t)$

  Update policy $\pi_{t+1} = \pi^{\widehat{q}_{t+1}}$

end for

5.1. Theoretical Results

We start by discussing the positive cumulative constraints violation. Notice that Algorithm 4 has the exact same optimization space as Algorithm 2, so the same results as in Theorem 3.2 holds, thanks to an completly analogous proof..

Theorem 5.1. *For any $\delta \in (0,1)$, Algorithm 4 attains, with probability at least $1 - 8\delta$:*

$$V_T = \mathcal{O}\left(L|X||A|\sqrt{T \ln\left({}^{mT|X||A|}/_{\delta}\right)} + \ln(T)|X||A|C\right).$$

Proof. The proof is completly analogous to the one of Theorem 3.2. □

Looking now at the achieved regret, to prove that algorithm 4 achieves sub linear regret, it is fundamental to prove that the optimal policy q^* is with high probability in the optimization space of the OMD in each episode *i.e.*

$$q^* \in \{q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \leq \underline{\alpha}\}, \quad \forall t \in [T].$$

This result is given by Lemma 9. Now, given that with high probability the optimal solution is contained in the gradually restricting convex optimization spaces at each episodes, the problem of bounding the regret can be re conducted to the problem of binding the regret of an OMD with bandit feedback and unknown transition probabilities in the equivalent unconstrained MDP. Thank to this the result presented in [17] holds.

Theorem 5.2. *For any $\delta \in (0,1)$, with probability at least $1-9\delta$, Algorithm 4 attains:*

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

Proof. The proof is completely analogous to the one of [17]. First we decompose the Regret similarly to [17]

$$\begin{aligned}
R_T &= \sum_{t\in[T]} \mathbb{E}[r_t]^\top (q^* - q_t) \\
&= \sum_{t\in[T]} \sum_{x,a} (1 - \mathbb{E}[r_t(x,a)])(q^*(x,a) - q_t(x,a)) \\
&= \underbrace{\sum_{t\in[T]} \sum_{x,a} (1 - \mathbb{E}[r_t(x,a)])(\widehat{q}_t - q_t)}_{ERROR} + \underbrace{\sum_{t\in[T]} \sum_{x,a} \left(f(x,a) - (1 - \mathbb{E}[r_t(x,a)])\right) \widehat{q}_t(x,a)}_{BIAS1} \\
&+ \underbrace{\sum_{t\in[T]} \sum_{x,a} f(x,a)(\widehat{q}_t(x,a) - q^*(x,a))}_{REG} + \underbrace{\sum_{t\in[T]} \sum_{x,a} \left((1 - \mathbb{E}[r_t(x,a)]) - f(x,a)\right) q^*(x,a)}_{BIAS2}.
\end{aligned}$$

Bound On ERROR

$$\begin{aligned}
ERROR &= \sum_{t\in[T]} \sum_{x,a} (1 - \mathbb{E}[r_t(x,a)])(\widehat{q}_t - q_t) \\
&\leq \sum_{t\in[T]} \sum_{x,a} (\widehat{q}_t(x,a) - q_t(x,a)) \\
&\leq \sum_{t\in[T]} \|\widehat{q}_t - q_t\|_1 \\
&= \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right),
\end{aligned}$$

since $1 - \mathbb{E}[r_t(x,a)] \leq 1$ for all (x,a) in $X \times A$, and by applying Lemma 20.

Bound on BIAS1 First we observe that $f_t(x,a)\widehat{q}_t(x,a) \leq \frac{\widehat{q}_t(x,a)}{u_t(x,a)+\gamma}(1 - r_t(x,a)) \leq 1$ under $\mathcal{E}_P$ thanks to the definition of u_t. Thus, by applying Azuma's inequality with probability at least $1-\delta$ it holds:

$$\sum_{t\in[T]} \sum_{x,a} (f(x,a) - \mathbb{E}[f(x,a)]) \leq L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}.$$

Finally we bound $\sum_{t\in[T]} \sum_{x,a} \left((1 - \mathbb{E}[r_t(x,a)]) - \mathbb{E}[f_t(x,a)]\right) \widehat{q}_t(x,a)$ under $\mathcal{E}_P$ as

$$\begin{aligned}
\sum_{t\in[T]} \sum_{x,a} \left((1 - \mathbb{E}[r_t(x,a)]) - \mathbb{E}[f_t(x,a)]\right) \widehat{q}_t(x,a) &= \sum_{t\in[T]} \sum_{x,a} \frac{\widehat{q}_t(x,a)(u_t(x,a) + \gamma - q_t(x,a))}{u_t(x,a) + \gamma} (1 - \mathbb{E}[r_t(x,a)]) \\
&\leq \sum_{t\in[T]} \|u_t - q_t\|_1 + \gamma T|X||A| \\
&= \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)} + \eta T|X||A|\right),
\end{aligned}$$

since by definition under $\mathcal{E}_P$ we have that $u_t(x,a) \geq q_t(x,a)$ and $u_t(x,a) \geq \widehat{q}_t(x,a)$ for all episode $t \in [T]$ and for all couple state-action $(x,a) \in X \times A$, and finally by applying Lemma 20.

Bound on REG Thanks to Lemma 9 we can employ the same exact reasoning as the one employed by [17], from which it holds the following with probability at least $1 - 5\delta$

$$REG \leq \mathcal{O}\left(\frac{L \ln\left(\frac{|X||A|}{\delta}\right)}{\eta} + \eta|X||A|T + \frac{\eta L \ln\left(\frac{L}{\delta}\right)}{\gamma}\right).$$

Bound on BIAS2 With probability at least $1 - \delta$

$$\begin{aligned} BIAS2 &= \sum_{t\in[T]} \sum_{x,a} \left(f_t(x,a) - (1 - \mathbb{E}[r_t(x,a)])\right) q^* \\ &\leq \sum_{t\in[T]} \sum_{x,a} \left(\frac{\mathbb{I}_t(x,a)}{u_t(x,a) + \gamma}(1 - r_t(x,a)) - (1 - \mathbb{E}[r_t(x,a)])\right) \\ &= \mathcal{O}\left(\frac{\ln\left(\frac{|X||A|}{\delta}\right)}{2\gamma}\right), \end{aligned}$$

which holds by applying 5 with f_t as $\widehat{\ell}_t$, M as $|X \times A|$ dimension of the vectors f_t.

Finally substituting η and γ with $\sqrt{\frac{L \ln\left(\frac{L|X||A|}{\delta}\right)}{T|X||A|}}$ we conclude the proof.

□

6. Conclusion

6.1. Results

In conclusion, we presented an innovative algorithm that operates in non-stationary CMDPs and achieves regret and positive constraint violation of order $\tilde{\mathcal{O}}(\sqrt{T} + C)$ when C is known. Additionally, we introduced a meta-procedure that attains the same guarantees when C is not known *a priori*. Finally, we proposed a hybrid algorithm combining the `NS-SOPS` designed in this thesis and `UOB-REPS` from [17], which is effective in scenarios with high corruption on the rewards, extending beyond the intended use cases of the previous algorithms.

6.2. Future work

A significant limitation in achieving better results in non-stationary CMDPs is the impossibility result of [20], which greatly restricts the potential for improvement when dealing with adversariality in CMDPs. Nonetheless, even though sublinear regret and positive violations are unachievable, there is still room for improvement for our algorithm when C is linear. One promising direction could be the use of meta-procedures, such as the one proposed in [1], to bound the worst-case scenario ($C = \mathcal{O}(T)$) by introducing an arm algorithm specifically designed for handling completely adversarial rewards and costs, potentially through the use of primal-dual techniques.
Finally, to apply these new algorithms to practical cases, it is essential to analyze their computational requirements and potentially improve their efficiency. Some steps towards efficiency have already been taken in this work, given that `lag-FTRL` updates only one instance of a sub-algorithm per episode, but there remains significant room for improvement.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] Qinbo Bai, Vaneet Aggarwal, and Ather Gattami. Provably efficient model-free algorithm for mdps with peak constraints. *arXiv preprint arXiv:2003.05555*, 2020.

[8] 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 2022.

[9] 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, 2022.

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

[11] 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.

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

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

[14] Jacopo Germano, Francesco Emanuele Stradi, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. A best-of-both-worlds algorithm for constrained mdps with long-term constraints. *arXiv preprint arXiv:2304.14326*, 2023.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

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

[25] 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.

[26] 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 2019.

[27] 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., 2019.

[28] 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.

[29] Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning adversarial mdps with stochastic hard constraints. *arXiv preprint arXiv:2403.03672*, 2024.

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

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

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

A. Event 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 19. 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,

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

  $\mathcal{E}_G$ 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 20. 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.

B. Auxiliary Lemmas from Existing Works

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

B.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 17 ([18]). *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 $u \in \Omega$:

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

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

B.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 18 ([22]). *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 ([22]). *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 $j \in [M]$.

B.3. Auxiliary Lemmas for the Transitions Estimation

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

Lemma 19 ([17]). *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:

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

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

Lemma 20 ([17]). *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,*

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

Abstract in lingua italiana

Nei *Processi Decisionali di Markov Vincolati* (Constrained Markov Decision Processes or CMDP) con ricompense e vincoli *avversari*, un noto risultato di impossibilità impedisce a qualsiasi algoritmo di ottenere sia un regret sublineare che una violazione dei vincoli sublineare, quando si compete contro una politica ottimale in retrospettiva che soddisfa i vincoli in media.
In questa tesi, mostriamo che questo risultato negativo può essere attenuato nei CMDP con ricompense e vincoli *non stazionari*, fornendo algoritmi le cui prestazioni degradano gradualmente con l'aumentare della non stazionarietà.
In particolare, proponiamo algoritmi che ottengono un regret di $\tilde{\mathcal{O}}(\sqrt{T}+C)$ e una violazione dei vincoli *positiva* con un feedback di tipo *bandit*, dove C è un valore di corruzione che misura la non stazionarietà dell'ambiente. Nel peggiore dei casi, questo valore può essere $\Theta(T)$, in coerenza con il risultato di impossibilità per i CMDP avversari.
Innanzitutto, progettiamo un algoritmo con le garanzie desiderate quando C è noto.
Successivamente, nel caso in cui C sia *sconosciuto*, mostriamo come ottenere gli stessi risultati incorporando tale algoritmo in una *meta-procedura* generale.
Questa è di per sé interessante, poiché può essere applicata a *qualsiasi* contesto di apprendimento online vincolato non stazionario.
Infine, progettiamo un algoritmo che, nelle stesse condizioni, con C *noto*, ottiene un regret di $\tilde{\mathcal{O}}(\sqrt{T})$ e una violazione positiva dei vincoli di $\tilde{\mathcal{O}}(\sqrt{T}+C)$.

Parole chiave: Online Learning, Markov Decision Processes, Markov Decision Processes Vincolati, Adversarial Online Learning, non-stazionarietà