POLITECNICO
MILANO 1863

POLITECNICO DI MILANO
DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA
DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

REGRET MINIMIZATION IN CONSTRAINED MARKOV DECISION PROCESSES

Doctoral Dissertation of:
Francesco Emanuele Stradi

Supervisor:
Prof. Nicola Gatti

Tutor:
Prof. Marco Domenico Santambrogio

The Chair of the Doctoral Program:
Prof. Luigi Piroddi

2025 – Cycle XXXVIII

Ringraziamenti

A mio nonno.
A mia madre.
Alla mia ragazza.
A tutta la mia famiglia.
Ai miei mentori Matteo, Nicola e Alberto.
Ai miei piu fidati colleghi Francesco, Gianmarco, Davide e Pierriccardo.
A tutti i miei coautori.

Abstract

Over the last years, *reinforcement learning* (RL) has attracted increasing interest from the AI community due to its capability of dealing with complex real-world environments in a sequential way. Standard online RL is usually modeled through the Markov decision process (MDP) framework, where an agent interacts sequentially with a multi-state environment and observes the rewards associated with its trajectory in the MDP. While online learning in classical (unconstrained) MDPs has received considerable attention in recent years, the setting of *constrained Markov decision processes* (CMDPs) is still largely unexplored. This is surprising, as in many real-world applications—such as autonomous driving, automated bidding, and recommender systems—there are often additional constraints and requirements that an agent needs to satisfy during the learning process.

In this dissertation, we study *online learning* in constrained MDPs. In online CMDPs, the learner (that is, the agent) selects a policy—a distribution over actions for any state—at each step of a given time horizon. As a result, the learner traverses the CMDP and observes the rewards and the costs associated with the constraints for the resulting trajectory. We focus on designing algorithms tailored to simultaneously *minimize the regret*—that is, the difference in performance between the optimal safe policy and the ones selected by the algorithm during the learning process—and *satisfy the constraints*, according to some measures of constraints violation. Constrained MDPs have been recently studied in the simplest case, where both rewards and constraints are assumed to be stochastically sampled from fixed distributions. Nonetheless, the gap between standard MDPs and constrained settings is still far from being fully addressed.

In this dissertation, we push forward the theoretical understanding of online learning in CMDPs. Specifically, we study different scenarios where rewards and costs may be either *stochastic*—that is, sampled from fixed distributions—or *adversarial*—that is, without any statistical assumptions. Moreover, we consider different notions of constraints violation. In the first part, we focus on stochastic CMDPs. We close one of the main open problems in the field by developing the first primal-dual algorithm—without relying on the linear programming formulation of CMDPs—able to achieve the optimal rate for both strong regret and strong constraints violation. These metrics do not allow negative regret (resp. violation) caused by selecting unsafe (resp. safe) policies. In the second part, we study CMDPs with adversarial rewards and stochastic constraints. In such a setting, we develop

algorithms capable of attaining different (and optimal) violation rates under the necessary assumptions. In the third part, we provide algorithms capable of simultaneously handling stochastic and adversarial constraints. Specifically, we design *best-of-both-worlds* algorithms that attain optimal regret and violation bounds when the constraints can be either stochastic or adversarial. Finally, in the fourth and last part of this dissertation, we consider settings where rewards and constraints are sampled from non-stationary distributions, assuming that the level of non-stationarity over time is bounded. In this case, we show how specific challenges arising in adversarial settings can be overcome when the non-stationarity is sufficiently limited.

Sommario

Negli ultimi anni, l'*apprendimento per rinforzo* (reinforcement learning, RL) ha suscitato un crescente interesse nella comunità dell'intelligenza artificiale grazie alla sua capacità di affrontare ambienti reali complessi in modo sequenziale. Il RL è solitamente modellato attraverso il framework dei processi decisionali di Markov (Markov Decision Process, MDP), in cui un agente interagisce in modo sequenziale con un ambiente a più stati e osserva le ricompense associate alla sua traiettoria nel MDP. Sebbene l'apprendimento online in MDP classici (non vincolati) abbia ricevuto notevole attenzione negli ultimi anni, l'ambito dei *processi decisionali di Markov vincolati* (Constrained Markov Decision Processes, CMDP) rimane in gran parte inesplorato. Questo è sorprendente, poiché in molte applicazioni reali—come la guida autonoma, le aste automatizzate e i sistemi di raccomandazione—spesso vi sono vincoli aggiuntivi e requisiti che l'agente deve rispettare durante il processo di apprendimento.

In questa dissertazione, studiamo l'*apprendimento online* nei MDP vincolati. Nei CMDP online, il decisore (cioè l'agente) seleziona una politica (policy)—una distribuzione di probabilità sulle azioni per ogni stato—a ogni step di un dato orizzonte temporale. Di conseguenza, l'agente attraversa il CMDP e osserva le ricompense e i costi associati ai vincoli lungo la traiettoria risultante. Ci concentriamo sulla progettazione di algoritmi in grado di *minimizzare il regret*—cioè la differenza di performance tra la politica ottimale sicura e quelle selezionate dall'algoritmo durante l'apprendimento—e *rispettare i vincoli*, secondo alcune misure di violazione dei vincoli. I CMDP sono stati recentemente studiati nel caso più semplice, in cui ricompense e vincoli sono assunti come campionati in modo stocastico da distribuzioni fisse. Tuttavia, il divario tra MDP standard e ambienti vincolati è ancora lontano dall'essere completamente colmato.

In questa dissertazione, approfondiamo la comprensione teorica dell'apprendimento online nei CMDP. Analizziamo diversi scenari in cui ricompense e costi possono essere di natura sia *stocastica* — ovvero campionati da distribuzioni fisse — sia *avversaria*, cioè generati in modo arbitrario senza alcuna assunzione probabilistica. Inoltre, esploriamo diverse definizioni di violazione dei vincoli. La prima parte è dedicata allo studio dei CMDP stocastici. In questo contesto, risolviamo uno dei principali problemi aperti del settore sviluppando il primo algoritmo primal-dual — che non si basa sulla formulazione tramite programmazione lineare — capace di raggiungere livelli ottimali sia per la met-

rica di *regret forte* sia per la *violazione forte* dei vincoli. Tali metriche non ammettono compensazioni: non è possibile, ad esempio, annullare il regret causato da una politica insicura con i benefici di una politica sicura. Nella seconda parte, esaminiamo i CMDP con ricompense avversarie e vincoli stocastici. Proponiamo una famiglia di algoritmi in grado di raggiungere diversi, e ottimali, livelli di violazione, a seconda delle ipotesi adottate. La terza parte si concentra su scenari più generali, in cui i vincoli possono essere sia stocastici sia avversari. In questo caso, sviluppiamo algoritmi *best-of-both-worlds*, capaci di garantire prestazioni ottimali in termini di regret e violazione, indipendentemente dalla natura del vincolo. Infine, nella quarta e ultima parte di questa dissertazione, affrontiamo contesti in cui ricompense e vincoli derivano da distribuzioni non stazionarie, ipotizzando che il grado di variazione nel tempo sia limitato. Mostriamo come, in tali ambienti, sia possibile affrontare con successo le sfide tipiche dei modelli avversari, a patto che la non stazionarietà resti entro soglie controllabili.

Contents

List of Algorithms

List of Tables

CHAPTER 1

Introduction

The framework of *Markov decision processes* (MDPs) (Bellman, 1957; Puterman, 2014) has been extensively employed to model sequential decision-making problems. In *reinforcement learning* (RL) (Sutton and Barto, 1998), the goal is to learn an optimal policy for an agent interacting with an environment modeled as an MDP.

A different line of work (Even-Dar et al., 2009; Neu et al., 2010) is concerned with problems in which an agent interacts with an unknown MDP with the goal of guaranteeing that the overall reward achieved during the learning process is as much as possible. This approach is more akin to *online learning* (Auer et al., 2002; Orabona, 2019), and it is far less investigated than classical RL approaches.

In real-world applications, there are usually additional constraints and specifications that an agent has to obey during the learning process, and these cannot be captured by the classical definition of MDP. For instance, autonomous vehicles must avoid crashing while navigating (Wen et al., 2020; Isele et al., 2018), bidding agents in ad auctions are constrained to a given budget (Wu et al., 2018; He et al., 2021), while recommender systems should *not* present offending items to users (Singh et al., 2020). In order to model such features of real-world problems, Altman (1999) introduced *constrained* MDPs (CMDPs) by extending classical MDPs with cost constraints that the agent has to satisfy.

In this dissertation, we push forward the theoretical understanding of *online learning* in *constrained* MDPs. In *online learning* in CMDPs, the objective is twofold. On the one hand, the learner aims at minimizing the *cumulative regret*—that is, the difference in performance between the algorithm and the optimal unknown policy—. On the other hand, the agent wants to keep the violation on the constraints as small as possible.

We study CMDPs where the rewards and the constraints may be either *stochastic*, that is, sampled from fixed distributions at each episode, or *adversarial*, namely, no statistical

assumptions are made. We provide algorithms tailored for every specific setting, and we show the related regret and constraints violation guarantees.

1.1 Related Works

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

1.1.1 Online Learning in Markov Decision Processes

There is a considerable literature on online learning problems (Cesa-Bianchi and Lugosi, 2006) in MDPs (see (Auer et al., 2008; Even-Dar et al., 2009; Neu et al., 2010) for some initial results on the topic). In such settings, two types of feedback are usually investigated: in the *full feedback* model, the entire reward (loss) function is observed after the learner's choice, while in the *bandit feedback* model, the learner only observes the reward due to the chosen action. Azar et al. (2017) study the problem of optimal exploration in episodic MDPs with unknown transitions and stochastic losses when the feedback is bandit. The authors present an algorithm whose regret upper bound is $\widetilde{\mathcal{O}}(\sqrt{T})$,[1] where T is the number of episodes, thus matching the lower bound for this class of MDPs and improving the previous result by Auer et al. (2008). Rosenberg and Mansour (2019b) study the online learning problem in episodic MDPs with adversarial losses and unknown transitions when the feedback is full information. The authors present an online algorithm exploiting entropic regularization and providing a regret upper bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. The same setting is investigated by Rosenberg and Mansour (2019a) when the feedback is bandit. In such a case, the authors provide a regret upper bound of the order of $\widetilde{\mathcal{O}}(T^{3/4})$, which is improved by Jin et al. (2020a) by providing an algorithm that achieves in the same setting a regret upper bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. Finally, Luo et al. (2021) provide the first policy optimization algorithm capable of matching the result of (Jin et al., 2020a), while avoiding the convex projection on the occupancy measure space.

1.1.2 Online Learning in Constrained Markov Decision Processes

The main body of work on constrained MDPs focuses on stochastic settings, that is, rewards and constraints are sampled from fixed distributions. Specifically, Zheng and Ratliff (2020) deal with episodic CMDPs with stochastic losses and constraints, where the transition probabilities are known and the feedback is bandit. The regret upper bound of their algorithm is of the order of $\widetilde{\mathcal{O}}(T^{3/4})$, while the cumulative constraint violation is guaranteed to be below a threshold with a given probability. Bai et al. (2020) provide the first algorithm that achieves sublinear regret[2] when the transition probabilities are unknown, assuming that the rewards are deterministic and the constraints are stochastic with a particular structure. Efroni et al. (2020) propose two approaches to deal with the exploration-exploitation dilemma in episodic CMDPs. These approaches guarantee sublinear regret and constraint violation when transition probabilities, rewards, and constraints are unknown and stochastic, while the feedback is bandit. Precisely, their LP based methods

[1] $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

[2] We refer as sublinear to a quantity that grows as $o(T)$.

guarantee $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative *strong* regret and cumulative *strong* constraints violations, that is, not allowing for cancellations between episodes. Differently, their primal-dual (or dual) algorithms attain $\widetilde{\mathcal{O}}(\sqrt{T})$ weak regret and violations. Liu et al. (2021) study the case where the rewards and the constrained are stochastic with a sub-gaussian form, attaining $\widetilde{O}(\sqrt{T})$ regret and zero violation when a strictly safe policy exists and it is known and bounded violation when the strictly safe policy exists but it is not known *a priori*. Ding et al. (2021) design a primal-dual policy optimization no-regret algorithm for CMDPs with stochastic rewards and stochastic constraints. Wei et al. (2022b) design a model-free, simulator-free reinforcement learning algorithm for CMDPs that achieves regret of order $\widetilde{O}(T^{4/5})$ with zero constraints violation, assuming the number of episodes to be exponentially large in $1/\rho$, where ρ is the Slater's parameter of the offline problem. Müller et al. (2024) provide the first primal-dual procedure capable of achieving sublinear cumulative *strong* regret and cumulative *strong* constraints violations. Finally, Ghosh et al. (2024) propose a model-free primal-dual algorithm for the linear MDP setting. Their algorithm attains $\widetilde{O}(\sqrt{T})$ strong violation if it is allowed to take $\Omega(d^{L-1}T^{1.5L}\log(|A|)^L)$ computational steps in every episode.

In the adversarial setting, Wei et al. (2018) deal with adversarial losses and stochastic constraints, assuming the transition probabilities are known and the feedback is full. The authors provide an algorithm that guarantees an upper bound of the order of $\widetilde{\mathcal{O}}(\sqrt{T})$ on both regret and constraints violation. Qiu et al. (2020) provide a primal-dual approach based on optimism. 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. Wei et al. (2023) , Ding and Lavaei (2023) consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded. Notice that all the works mentioned above focus on constraints violation which allows for cancellations between episodes.

1.1.3 Online Learning with Unknown Constraints

Online leaning with *unknown* constraints has been widely investigated (see, *e.g.*, (Mannor et al., 2009; Liakopoulos et al., 2019; Castiglioni et al., 2022a,b)). Two main settings are usually studied. In canonical settings (see, *e.g.*, (Chen et al., 2022)), the aim is to guarantee that the constraints violation incurred by the algorithm grows sub-linearly. In *hard constraints* settings, the algorithms must satisfy the constraints at every round, by assuming knowledge of a strictly feasible decision (see, *e.g.*, (Pacchiano et al., 2021)).[3] Both soft and hard constraints have been generalized to settings that are more challenging than multi-armed bandits, such as linear bandits (see, *e.g.*, (Gangrade et al., 2024)). A central result is provided by Mannor et al. (2009), who show that it is impossible to suffer from sublinear regret and sublinear constraint violation when an adversary chooses losses and constraints. Liakopoulos et al. (2019) try to overcome such an impossibility result by defining a new notion of regret. They study a class of online learning problems with long-term budget constraints that can be chosen by an adversary. The learner's regret metric is modified by introducing the notion of a *K-benchmark*, *i.e.*, a comparator that meets the problem's allotted budget over any window of length K. Castiglioni et al. (2022a,b); Bernasconi et al.

[3]The definition of *hard constraints* is ambiguous in the online learning literature. Indeed, attaining sublinear *strong* violation is often referred to as a *hard constraint* guarantee.

(2024) focus on online learning with stochastic and adversarial constraints, providing the first *best-of-both-worlds* algorithms for online learning problems with constraints. Finally, some works focus on constrained online convex optimization settings (see, *e.g.*, (Mahdavi et al., 2012; Jenatton et al., 2016; Yu et al., 2017)).

1.2 Original Contributions

The goal of this dissertation is to advance the theoretical understanding of online *constrained Markov decision processes*. Specifically, we study CMDPs where the rewards and the constraints can be either stochastic or adversarial and we provide theoretical guarantees—in terms of *regret* and *violation* bounds—for every specific setting.

In this section, we survey the original contributions of this dissertation. Specifically, in Section 1.2.1 we provide our contributions in the field of online *stochastic* CMDPs, that is, both rewards and constraints are sampled from fixed distributions. In Section 1.2.2, we provide the summary of our results for CMDPS with stochastic constraints, where the rewards (losses) are allowed to change arbitrarily, that is, CMDPs with adversarial rewards and stochastic constraints. In Section 1.2.3, we provide our results in terms of algorithms capable of simultaneously handling stochastic and adversarial CMDPs. Finally, in Section 1.2.4, we provide our contributions in the field of CMDPs with non-stationary rewards and constraints.

The settings studied in this dissertation, and the associated chapters, are summarized in Table 1.1, where we additionally provide a comparison with the settings studied by state-of-the-art works in the field.

Table 1.1: *Comparison between the settings studied by the state-of-the-art works on constrained MDPs and the ones studied in this dissertation. † stands for adversarial environment with bounded non-stationarity.*

	Stochastic Rewards & Stochastic Constraints	Adversarial Rewards & Stochastic Constraints	Adversarial Constraints	Feedback
(Efroni et al., 2020)	✓	✗	✗	Bandit
(Qiu et al., 2020)	✓	✓	✗	Full
(Liu et al., 2021)	✓	✗	✗	Bandit
(Wei et al., 2023)	✓	✗†	✗†	Bandit
Chapter 3	✓	✗	✗	Bandit
Chapter 4	✓	✓	✗	Bandit
Chapter 5	✓	✓	✓	Full
Chapter 6	✓	✓	✓	Bandit
Chapter 7	✓	✓	✓	Bandit
Chapter 9	✓	✗†	✗†	Bandit

1.2.1 CMDPs with Stochastic Rewards and Stochastic Constraints

In the first part of the dissertation, we focus on CMDPs with *stochastic rewards* and *stochastic constraints*, under *bandit feedback*. In such a setting, the classical notions of

regret and violation are usually called *weak*, due to the fact that they allow for negative terms to cancel out positive ones. In CMDPs, this means that the (weak) regret can be easily controlled by using policies achieving large rewards *without* satisfying the constraints. Similarly, the (weak) violation can be controlled by adopting policies satisfying cost constraints by a large margin. The *strong* regret and the *strong* constraints violation are much more reasonable metrics compared to their weak counterparts, as they do *not* allow negative terms to cancel out positive ones. However, achieving sublinear strong regret/violation in CMDPs is much more challenging.

Efroni et al. (2020) were the first to provide a learning algorithm with (optimal) $\widetilde{\mathcal{O}}(\sqrt{T})$ strong regret/violation in general CMDPs. However, their algorithm works by solving linear programs defined over the space of occupancy measures, a task that is highly inefficient in practice. Ideally, one would like learning algorithms that avoid dealing with occupancy measures, by directly optimizing over the policy space. Such policy optimization algorithms are much more efficient and desirable in practice. By leveraging a primal-dual scheme, Efroni et al. (2020) designed a first policy optimization algorithm for CMDPs, though it can only achieve sublinear *weak* regret and *weak* violation, leaving as an open problem whether an analogous result is achievable for the strong metrics.

Very recently, (Müller et al., 2024) partially addressed this problem by proposing a primal-dual policy optimization algorithm attaining $\widetilde{\mathcal{O}}(T^{0.93})$ *strong* regret and *strong* violation. However, the bounds achieved by such an algorithm remain largely suboptimal, leaving a big gap that still needs to be closed.

In the first part of the dissertation, we answer the question left open by (Efroni et al., 2020; Müller et al., 2024): that is, whether it is possible to achieve *optimal* $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds on the *strong* regret and the *strong* constraints violation in CMDPs by using an *efficient* primal-dual *policy optimization* algorithm. We answer the question above affirmatively. To do so, we design a learning algorithm that exploits a novel primal-dual scheme. Specifically, our algorithm adopts, as primal regret minimizer, a state-of-the-art policy optimization algorithm for adversarial (unconstrained) MDPs, while it leverages an approach based on upper confidence bounds in order to build a dual regret minimizer. Crucially, the updates of dual variables performed by our algorithm do *not* resort to optimizing over the space of occupancy measures, making our algorithm a fully policy optimization approach, and, thus, efficient.

1.2.2 CMDPs with Adversarial Rewards and Stochastic Constraints

In the second part of the dissertation, we focus our attention on CMDPs with *adversarial rewards (losses)* and *stochastic constraints*, under *bandit feedback*. We consider three scenarios that differ in the way in which constraints are satisfied and are all usually referred to as *hard* constraints settings in the literature (Liu et al., 2021; Guo et al., 2022). In the first scenario, the learner attains *sublinear* cumulative *strong* constraints violation. In the second one, the learner satisfies constraints at every episode, while in the third one, they achieve *constant* cumulative *strong* constraints violation.

Our work is the first to study CMDPs with both adversarial losses and hard constraints. Indeed, all the works on adversarial CMDPs (see, *e.g.*, (Wei et al., 2018; Qiu et al., 2020)) consider settings with *soft* constraints. These are much weaker than hard constraints, as they are only concerned with the minimization of the *weak* constraints violation. Furthermore, the only few works addressing stochastic hard constraints in CMDPs (Liu et al.,

2021; Shi et al., 2023; Müller et al., 2024) are restricted to *stochastic losses*. Thus, their techniques *cannot* be easily generalized to our setting.

We start by addressing the first scenario, where we design an algorithm that guarantees both sublinear regret and sublinear cumulative strong constraints violation, thus resolving a problem left open by Qiu et al. (2020), *i.e.*, learning with *bandit* feedback in CMDPs with adversarial losses and stochastic constraints. Indeed, we go even further, as Qiu et al. (2020) were only concerned with soft constraints, while our algorithm is capable of managing *strong* constraints violation.

Next, we switch the attention to the second scenario, where we design a *safe* algorithm, *i.e.*, one that satisfies the constraints at every episode. To achieve this, we need to assume that the learner has knowledge about a policy $\pi^\diamond$ strictly satisfying the constraints. We design an algorithm that attains sublinear regret while being safe with high probability.

Then, in the third scenario, we design an algorithm that attains *constant* cumulative strong constraints violation and sublinear regret, by simply assuming that a policy strictly satisfying the constraints exists (Slater's condition holds), but it is *not* known to the learner.

Finally, we provide a lower bound showing that any algorithm attaining $o(\sqrt{T})$ violation cannot avoid a dependence on the Slater's parameter in the regret bound. We believe that this result may be of independent interest, since it is not only applicable to our second and third settings, but also to other settings where a larger violation is allowed.

1.2.3 Best-of-Both-Worlds Algorithms for CMDPs

In the third part of the dissertation, we focus on designing algorithms capable of simultaneously handling *stochastic* and *adversarial* constraints. Specifically, we pioneer the study of CMDPs in which the constraints are selected adversarially.

First, we introduce an algorithm that employs a novel primal-dual approach in CMDPs, allowing it to attain *best-of-both-worlds* guarantees, under *full feedback*, in the flavor of Balseiro et al. (2023). In particular, our algorithm provides optimal (in the number of episodes T) regret and constraints violation bounds when rewards and constraints are selected either *stochastically* or *adversarially*, without requiring any knowledge of the underling process. While best-of-both-worlds algorithms have been recently introduced in online learning settings subject to constraints (see, *e.g.*, (Liakopoulos et al., 2019; Balseiro et al., 2023)), to the best of our knowledge our algorithm is the first of its kind in CMDPs—we underline that, in the literature on online learning in MDPs, the term *best-of-both-worlds* is sometimes referred to algorithms that achieve optimal instance-dependent regret bounds when rewards are selected *stochastically* and $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when rewards are chosen *adversarially* (Jin et al., 2021). In this work, we borrow terminology from the literature on online learning with constraints, where the term usually refers to algorithms that achieve optimal regret and constraints violation bounds when the constraints are selected either *stochastically* or *adversarially* (Balseiro et al., 2023).

When the constraints are selected stochastically, we show that our algorithm provides $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative regret and constraint violation when a suitably-defined Slater-like condition concerning the satisfiability of constraints is satisfied. Moreover, whenever such a condition does *not* hold, our algorithm still ensures $\widetilde{\mathcal{O}}(T^{3/4})$ regret and constraint violation. Instead, whenever the constraints are chosen adversarially, our analysis revolves around the Slater's parameter ρ which is related to the "margin" by which it is possible to strictly satisfy the constraints. Indeed, under adversarial constraints, Mannor et al.

(2009) show that it is impossible to simultaneously achieve sublinear regret and sublinear cumulative constraint violation. We prove that our algorithm achieves sublinear α-regret—that is, sublinear regret with respect to a fraction α of the *constrained* optimum—where $\alpha = \rho/(L+\rho)$, while guaranteeing that the cumulative constraints violation is sublinear in the number of episodes. *This matches the optimal regret guarantees derived for other best-of-both-worlds algorithms in (non-sequential) online learning settings* (Castiglioni et al., 2022a; Balseiro et al., 2023), whenever ρ is a constant independent on T.

Thus, we extend the aforementioned results to the *bandit feedback* setting. This is done employing a primal-dual *policy optimization* method, which is arguably much more efficient than the algorithm developed for the *full feedback* setting, since it does *not* require solving convex programs.

Finally, we propose a novel algorithm that improves the best-of-both-worlds results provided in the previous chapters, under *bandit feedback*. In the stochastic setting, the algorithm attains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation without Slater's condition, that is, a strictly feasible solution may not exist. Furthermore, the algorithm attains $\widetilde{\mathcal{O}}(\sqrt{T})$ *strong* constraint violation, that is, not allowing for cancellations between episodes. In the adversarial setting, the algorithm attains sublinear violation without Slater's condition. Furthermore, employing a stronger notion of Slater's parameter, our algorithm attains sublinear α-regret with respect to the *unconstrained* optimum, instead of the constrained one.

Table 1.2: *Comparison between the theoretical guarantees attained by the state-of-the-art works on constrained MDPs and the ones provided in this dissertation.* † *The results hold taking the expectation on the constraints function in the violation definition.*

	Regret R_T	Violation V_T	*Strong* Regret $\mathcal{R}_T$	*Strong* Violation $\mathcal{V}_T$
(Efroni et al., 2020) *Algorithms 1-2*	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$
(Efroni et al., 2020) *Algorithms 3-4*	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
(Qiu et al., 2020)	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
(Liu et al., 2021) *Algorithm 1*	$\widetilde{\mathcal{O}}(\sqrt{T})$	$0^{\dagger}$	✗	0
(Liu et al., 2021) *Algorithm 2*	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\mathcal{O}(1)^{\dagger}$	✗	✗
Algorithm 3.1	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$
Algorithm 4.1	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	$\widetilde{\mathcal{O}}(\sqrt{T})$
Algorithm 4.2	$\widetilde{\mathcal{O}}(\sqrt{T})$	$0^{\dagger}$	✗	0
Algorithm 4.3	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\mathcal{O}(1)^{\dagger}$	✗	$\mathcal{O}(1)$
Algorithm 5.1 (*Sto. C.*)	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
Algorithm 5.1 (*Adv. C.*)	$\frac{\rho}{\rho+L} R_T$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
Algorithm 6.1 (*Sto. C.*)	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
Algorithm 6.1 (*Adv. C.*)	$\frac{\rho}{\rho+L} R_T$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
Algorithm 7.1 (*Sto. C.*)	$\widetilde{\mathcal{O}}(\sqrt{T})$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	$\widetilde{\mathcal{O}}(\sqrt{T})$
Algorithm 7.1 (*Adv. C.*)	$\frac{\rho}{\rho+L} R_T$	$\widetilde{\mathcal{O}}(\sqrt{T})$	✗	✗
Algorithm 9.2	$\widetilde{\mathcal{O}}(\sqrt{T}+C)$	$\widetilde{\mathcal{O}}(\sqrt{T}+C)$	✗	$\widetilde{\mathcal{O}}(\sqrt{T}+C)$

1.2.4 CMDPs with Non-Stationary Rewards and Constraints

In the fourth part of the dissertation, we focus on CMDPs with *non-stationary* rewards and constraints, under *bandit feedback*. This is primarily motivated by the well-known impossibility result by Mannor et al. (2009), which prevents any learning algorithm from attaining both sublinear regret and sublinear constraints violation, when competing against a best-in-hindsight policy that satisfies the constraints *on average*. In this part, we focus on how to ease the negative result by Mannor et al. (2009).

As a preliminary step, we show that the primal-dual algorithm tailored to attain best-of-both-worlds results with bandit feedback can still recover the desired $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *weak* violation bounds, where C is the measure of adversariality of the constraints.

Then, we consider non-stationary settings that generalize *both* stochastic CMDPs and adversarial ones. Specifically, we address CMDPs where rewards and constraints are selected from probability distributions that are allowed to change *adversarially* from episode to episode. Thus, our CMDPs bridge the gap between fully-stochastic and fully-adversarial ones. We design algorithms whose performances smoothly degrade as a suitable measure of the adverseness of rewards and constraints increases. This is called *(adversarial) corruption*, and it intuitively quantifies how much the distributions of rewards and constraints vary over the episodes with respect to some suitable non-corrupted counterparts. We pro-

Table 1.3: *Comparison on the assumptions employed in the state-of-the-art works on constrained MDPs and in this dissertation.* † *Slater's condition is not necessary to obtain sublinear results, but it is required for the optimal bounds.*

	Slater's Condition	Knowledge of ρ	Knowledge of $\pi^\diamond$
(Efroni et al., 2020) *Algorithms 1-2*	✗	✗	✗
(Efroni et al., 2020) *Algorithms 3-4*	✓	✓	✗
(Qiu et al., 2020)	✓	✗	✗
(Liu et al., 2021) *Algorithm 1*	✓	✓	✓
(Liu et al., 2021) *Algorithm 2*	✓	✓	✗
Algorithm 3.1	✓	✓	✗
Algorithm 4.1	✗	✗	✗
Algorithm 4.2	✓	✓	✓
Algorithm 4.3	✓	✗	✗
Algorithm 5.1 (*Sto. C.*)	✗†	✗	✗
Algorithm 5.1 (*Adv. C.*)	✓	✗	✗
Algorithm 6.1 (*Sto. C.*)	✗†	✗	✗
Algorithm 6.1 (*Adv. C.*)	✓	✗	✗
Algorithm 7.1 (*Sto. C.*)	✗	✗	✗
Algorithm 7.1 (*Adv. C.*)	✓	✗	✗
Algorithm 9.2	✓	✓	✗

pose algorithms that attain $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *strong* constraint violation, where C denotes the corruption of the setting. We remark that $C = \Theta(T)$ in the worst case, and, thus, our bounds are coherent with the impossibility result by Mannor et al. (2009). Moreover, in stochastic CMDPs, our bounds reduce to state-of-the-art $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds (Efroni et al., 2020).

We conclude the section referring to Table 1.2 for a summary of the theoretical guarantees attained by the state-of-the-arts algorithms for online CMDPs and the ones attained by the algorithms provided in this dissertation. Similarly, we refer to Table 1.3 for a summary on the assumptions required by the aforementioned algorithms.

1.3 Structure of the Work

In this section, we provide the structure of the dissertation. Before diving into the technical results of this work, we outline some preliminary concepts needed for the rest of the dissertation. Specifically:

- In Chapter 2, we provide the definition of constrained Markov decision processes. Then, we define the notion of occupancy measure. Additionally, we define the offline optimum and the online performance metrics employed in the rest of the dissertation.

Part I: Stochastic Rewards and Stochastic Constraints

In this part, we focus on constrained Markov decision processes with stochastic rewards and stochastic constraints. Specifically:

- In Chapter 3, we provide the first primal-dual methods capable of attaining optimal *strong* regret and *strong* violation and we derive its theoretical guarantees. We refer to Appendix A for the omitted lemmas and proofs. The results in this chapter appeared in (Stradi et al., 2025a).

Part II: Adversarial Rewards and Stochastic Constraints

In this part, we focus on constrained Markov decision processes with adversarial rewards (losses) and stochastic constraints. Specifically:

- In Chapter 4, we provide the algorithms tailored to handle the three different *hard constraints* settings, while achieving sublinear regret. We derive their theoretical guarantees and we finally provide the associated lower bound. We refer to Appendix B for the omitted lemmas and proofs. The results in this chapter appeared in (Stradi et al., 2025b).

Part III: Best-of-Both-Worlds

In this part, we focus on constrained Markov decision processes where both the rewards and the constraints can be either stochastic or adversarial. Specifically:

- In Chapter 5, we provide the first *best-of-both-worlds* algorithm for CMDPs, under *full feedback*. We first show the specific primal-dual scheme and then we derive its theoretical guarantees. We refer to Appendix C for the omitted lemmas and proofs. The results in this chapter appeared in (Stradi et al., 2024).

- In Chapter 6, we provide the first *best-of-both-worlds* algorithm for CMDPs, under *bandit feedback*, deriving its theoretical guarantees. We refer to Appendix D for the omitted lemmas and proofs. The results in this chapter appeared in (Stradi et al., 2025e).
- In Chapter 7, we provide a novel *best-of-both-worlds* algorithm for CMDPs, under *bandit feedback*, which greatly improves the results attained in Chapter 6. We refer to Appendix E for the omitted lemmas and proofs. The results in this chapter appeared in (Stradi et al., 2025c)

Part IV: Non-Stationary Rewards and Constraints

In this part, we focus on constrained Markov decision processes where both the rewards and the constraints are non-stationary. Specifically:

- In Chapter 8, we extend the results provided in Chapter 6 by showing that the primal-dual scheme employed to attain best-of-both-worlds results can still achieve sublinear results when the environment exhibits partial adversariality.
- In Chapter 9, we provide algorithms capable of attaining sublinear regret and sublinear *strong* constraints violation in non-stationary environment. We first study the case where the *amount* of non-stationarity of the environment is known, deriving the associated theoretical guarantees. Then, we focus on the case in which the *amount* of non-stationarity of the environment is *not* known, attaining comparable results. We refer to Appendix F for the omitted lemmas and proofs. The results in this chapter appeared in (Stradi et al., 2025d).

Finally, in Chapter 10, we conclude the dissertation by providing a final discussion on the results attained and pointing the reader's attention to possible future works.

CHAPTER 2

Preliminaries

In this chapter, we provide the necessary preliminaries on the definitions and the notation that will be employed in the rest of the dissertation. Specifically, in Section 2.1 we formally define the *constrained Markov decision processes* framework and the learner-environment interaction. In Section 2.2, we provide the definition of *occupancy measure*. In Section 2.3, we introduce the *offline optimum* of CMDPs. In Section 2.4, we provide the *online performance metrics* that will be employed to evaluate the performance of our algorithm. Finally, in Section 2.5, we discuss on feasibility parameters and on the necessary assumption to make the problem learnable.

We refer to (Altman, 1999) for a complete discussion on CMDPs.

2.1 Constrained Markov Decision Processes

We study *episodic constrained* MDPs, which we call CMDPs for short and are defined as tuples $M = \left(X, A, m, P, \{r_t\}_{t=1}^{T}, \{G_t\}_{t=1}^{T}, \theta\right)$, where:

- $T \in \mathbb{N}_{>0}$ is the number of episodes of the learning dynamic, with $t \in [T]$ denoting a specific episode.[1]
- X and A are finite state and action spaces, respectively.
- $m \in \mathbb{N}_{>0}$ is the number of constraints.
- $P : X \times A \times X \to [0, 1]$ is the transition function, where, for ease of notation, we denote by $P(x'|x, a)$ the probability of going from state $x \in X$ to $x' \in X$ by taking action $a \in A$.[2]

[1] We denote with $[a, \dots, b]$ the set of all consecutive integers from a to b, while $[b] = [1, \dots, b]$.

[2] W.l.o.g., in this work we consider *loop-free* CMDPs. Formally, this means that X is partitioned into L layers $X_0, \dots, X_L$ such that the first and the last layers are singletons, *i.e.*, $X_0 = \{x_0\}$ and $X_L = \{x_L\}$, and that $P(x'|x, a) > 0$ only if

- $\{r_t\}_{t=1}^T$ is a sequence of vectors describing the rewards at each episode $t \in [T]$, namely $r_t \in [0,1]^{|X \times A|}$. We refer to the reward of a specific state-action pair $x \in X, a \in A$ for an episode $t \in [T]$ as $r_t(x,a)$. Rewards may be *stochastic*, in that case r_t is a random variable distributed according to a distribution $\mathcal{R}$ for every $t \in [T]$, or chosen by an *adversary*.[3]
- $\{G_t\}_{t=1}^T$ is a sequence of matrices describing the m *constraint costs* at each episode $t \in [T]$, namely $G_t \in [0,1]^{|X \times A| \times m}$. For $i \in [m]$, we refer to the violation of the i-th constraint cost for a specific state-action pair $x \in X, a \in A$ at episode $t \in [T]$ as $g_{t,i}(x,a)$. Constraint costs may be *stochastic*, in that case G_t is a random variable distributed according to a probability distribution $\mathcal{G}$ for every $t \in [T]$, or chosen by an *adversary*.
- $\theta \in [0, L]^m$ is a threshold vector, whose component θ_i is for constraint $i \in [m]$.

Algorithm 2.1 Learner-Environment Interaction

for $t = 1, \dots, T$ **do**
  r_t and G_t are chosen *stochastically* or *adversarially*
  The learner chooses a policy $\pi_t : X \times A \to [0,1]$
  The state is initialized to x_0
  for $k = 0, \dots, L-1$ **do**
    The learner plays $a_k \sim \pi_t(\cdot|x_k)$
    The environment evolves to $x_{k+1} \sim P(\cdot|x_k, a_k)$
    The learner observes $r_t(x_k, a_k)$ and $g_{t,i}(x_k, a_k)$, $\forall i \in [m]$ ▷ *bandit feedback*
    The learner observes x_{k+1}
  end for
  The learner is revealed r_t, G_t ▷ *full feedback*
end for

The learner chooses a *policy* $\pi_t : X \times A \to [0,1]$ at each episode, defining a probability distribution over actions at each state. For ease of notation, we denote by $\pi(\cdot|x)$ the probability distribution for a state $x \in X$, with $\pi(a|x)$ denoting the probability of action $a \in A$. We will refer to the set of policies as Π. Thus, the learner traverses the CMDP given the randomization of the policy π_t and the transition P. When *full feedback* is available, the learner observes the entire reward vector and costs matrix. Differently, when only *bandit feedback* is available, the learner observes rewards and costs for the path traversed only. Algorithm 2.1 depicts the interaction between the learner and the environment in a CMDP. As is standard in the RL literature (Sutton and Barto, 1998), we assume that the learner knows X and A, while P is *not* known.

2.1.1 Value Functions

Given a transition function $P : X \times A \times X \to [0,1]$, a policy $\pi : X \times A \to [0,1]$, and a generic vector $v \in [0,1]^{|X \times A|}$ indexed on state-action pairs, we introduce a value function $V^{\pi,P}(\cdot|v) : X \to [0, L]$ that is defined as follows, for every $k \in [0, \dots, L-1]$

$x' \in X_{k+1}$ and $x \in X_k$ for some $k \in [0, \dots, L-1]$. Notice that any episodic CMDP with horizon H that is *not* loop-free can be cast into a loop-free one by suitably duplicating the state space H times, *i.e.*, a state x is mapped to a set of new states (x,k), where $k \in [0, \dots, H]$.

[3] In some chapters of the dissertation, we will make use of losses $\ell_t \in [0,1]^{|X \times A|}$ in place of the rewards. Notice that, by taking $\ell_t(x,a) = 1 - r_t(x,a)$ for all $x \in X, a \in A$ and $t \in [T]$, the two notations are equivalent.

and $x \in X_k$:

$$V^{\pi,P}(x,v) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{k'=k}^{L-1} v(x_{k'},a_{k'}) \mid x_k = x\right].$$

Moreover, we define $V^{\pi,P}(v) \coloneqq V^{\pi,P}(x_0,v)$. Notice that $V^{\pi,P}(\cdot|r)$ and $V^{\pi,P}(\cdot|g_i)$ encode the value functions for the rewards r and the costs g_i of some constraint $i \in [m]$, respectively. In the following, we sometimes omit the dependency of $V^{\pi,P}(\cdot|v)$ on P, when this is clear from context.

2.2 Occupancy Measures

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

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

Moreover, we also define:

$$\begin{aligned} q^{P,\pi}(x,a) &= \sum_{x'\in X_{k+1}} q^{P,\pi}(x,a,x'), \\ q^{P,\pi}(x) &= \sum_{a\in A} q^{P,\pi}(x,a). \end{aligned} \tag{2.1}$$

Then, we can introduce the following lemma, which characterizes *valid* occupancy measures.

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

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

where P is the transition function of the MDP and P^q is the one induced by q (see Equation (2.2)).

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

$$P^q(x'|x,a) = \frac{q(x,a,x')}{q(x,a)}, \quad \pi^q(a|x) = \frac{q(x,a)}{q(x)}. \tag{2.2}$$

To conclude, we underline that the value functions defined in Section 2.1.1 can be easily rewritten in terms of occupancy measures, that is, it holds, for a generic vector $v \in [0,1]^{|X\times A|}$:

$$V^{\pi,P}(v) = v^\top q^{P,\pi},$$

where $q^{P,\pi}$ is the occupancy measure defined as in Equation (2.1).

2.3 Offline CMDPs Optimization

In the following, we characterize the offline optimization problem $\text{LP}_{r,G,\theta}$—with parameters r, G and θ—which is used to define the baselines against which we compare the performances of our algorithm, as:

$$\text{OPT}_{r,G,\theta} := \begin{cases} \max_{q\in\Delta(M)} & r^\top q \\ \text{s.t.} & G^\top q \leq \theta, \end{cases} \tag{2.3}$$

where $q \in [0,1]^{|X\times A|}$ is the occupancy measure vector whose values are defined by the expression in Equation (2.1), $\Delta(M)$ is the set of valid occupancy measures, r is the reward vector, and G is the constraint matrix.

Furthermore, we introduce the following condition.

Condition 2.1 (Slater's condition)**.** *Given a cost matrix G, the Slater's condition holds when there is a strictly feasible solution $q^\diamond$ such that $G^\top q^\diamond < \theta$.*

As we will see throughout the dissertation, Condition 2.1 is not always necessary to make the online CMDP problem learnable. Nonetheless, it is required in harder settings (*e.g.*, the adversarial one) or for the functioning of specific kinds of algorithms (*e.g.*, primal-dual procedures).

Then, we define the Lagrangian function for Problem (2.3).

Definition 2.1 (Lagrangian function)**.** *Given a reward vector r and a cost matrix G, the Lagrangian function $\mathcal{L}_{r,G,\theta} : \Delta(M) \times \mathbb{R}^m_{\geq 0} \to \mathbb{R}$ of Problem (2.3) is defined as:*

$$\mathcal{L}_{r,G,\theta}(q,\lambda) := r^\top q - \lambda^\top (G^\top q - \theta).$$[4]

It is well known (see, *e.g.*, (Altman, 1999)) that strong duality holds for CMDPs assuming Slater's condition. Therefore, we have that the following corollary holds.

Corollary 2.1. *Given a reward vector r and a constraint cost matrix G such that Slater's condition holds, we have:*

$$\begin{aligned} \text{OPT}_{r,G,\theta} &= \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{q\in\Delta(M)} \mathcal{L}_{r,G,\theta}(q,\lambda) \\ &= \max_{q\in\Delta(M)} \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \mathcal{L}_{r,G,\theta}(q,\lambda). \end{aligned}$$

Notice that the min-max problem in Corollary 2.1 corresponds to the optimization problem associated with a zero-sum Lagrangian game.

[4] The Lagrangian function can be easily defined with respect to the policy, too. In such a case, we define the Lagrangian function as $\mathcal{L}_{r,G,\theta}(\pi,\lambda) := r^\top q^{P,\pi} - \lambda^\top (G^\top q^{P,\pi} - \theta)$.

2.4 Cumulative Regret and Constraint Violation

In this section, we introduce the online performance metrics which will be used to evaluate our algorithms. We first introduce the notion of cumulative regret and cumulative constraint violation up to episode T.

The *cumulative regret* is defined as:

$$R_T := T \, \mathrm{OPT}_{\overline{r},\overline{G},\theta} - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t},$$

where:

$$\overline{r} := \begin{cases} \mathbb{E}_{r\sim\mathcal{R}}[r] & \text{if the rewards are stochastic} \\ \frac{1}{T}\sum_{t=1}^{T} r_t & \text{if the rewards are adversarial,} \end{cases}$$

$$\overline{G} := \begin{cases} \mathbb{E}_{G\sim\mathcal{G}}[G] & \text{if the constraint costs are stochastic} \\ \frac{1}{T}\sum_{t=1}^{T} G_t & \text{if the constraint costs are adversarial.} \end{cases}$$

Notice that, in the adversarial case, the regret is computed with respect to an *optimal feasible strategy in hindsight*. We refer to an optimal occupancy measure (*i.e.*, a feasible one achieving value $\mathrm{OPT}_{\overline{r},\overline{G},\theta}$) as q^* and to the associated policy as π^*. Thus, we can write $\mathrm{OPT}_{\overline{r},\overline{G},\theta} = \overline{r}^\top q^*$ and the regret reduces to

$$R_T := \sum_{t=1}^{T} \overline{r}^\top q^* - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t}.$$

The *cumulative constraint violation* is defined as:

$$V_T := \max_{i\in[m]} \sum_{t=1}^{T} \left[G_t^\top q^{P,\pi_t} - \theta\right]_i.$$ [5]

Moreover, for the *stochastic* setting only, we define the following metrics of *cumulative strong regret* and *cumulative strong constraint violation*.

The *cumulative strong regret* is defined as:

$$\mathcal{R}_T := \sum_{t=1}^{T} \left[\overline{r}^\top q^* - \overline{r}^\top q^{P,\pi_t}\right]^+,$$

where $[\cdot]^+ := \max\{0, \cdot\}$. Similarly, we define the *cumulative strong constraint violation* as:

$$\mathcal{V}_T := \max_{i\in[m]} \sum_{t=1}^{T} \left[\overline{G}^\top q^{P,\pi_t} - \theta\right]_i^+.$$

Notably, the strong performance metrics do not allow negative regret (resp. violation) caused by selecting unsafe (resp. safe) policies.

For the sake of notation, we will refer to q^{P,π_t} by using q_t, thus omitting the dependence on P and π.

Our goal is to design online learning algorithms that attain sublinear regret and sublinear violation, that is, $R_T = o(T)$ and $V_T = o(T)$. When possible, we will show how to additionally attain $\mathcal{R}_T = o(T)$ and $\mathcal{V}_T = o(T)$.

[5] Given a generic vector $a \in \mathbb{R}^b$, we denote as $[a]_b$ its b-th component.

2.5 Feasibility Parameter

We introduce a problem-specific parameter $\rho \in [0, L]$, which is strictly related to the feasibility of Problem (2.3), and in particular to "how much" Slater's condition is satisfied. Formally, in settings with *stochastic constraints* chosen from a fixed distribution, the parameter ρ is defined as:

$$\rho := \max_{q \in \Delta(M)} \min_{i \in [m]} - \left[\overline{G}^\top q - \theta\right]_i .$$

Instead, in the *adversarial constraints* setting, ρ is defined as:

$$\rho := \max_{q \in \Delta(M)} \min_{t \in [T]} \min_{i \in [m]} - \left[G_t^\top q - \theta\right]_i .$$

In both cases, the occupancy measure leading to the value of ρ is denoted by $q^\diamond$. Intuitively, ρ represents the "margin" by which the "most feasible" strictly feasible solution satisfies the constraints. Notice that when $\rho > 0$, Slater's condition holds.

Throughout the dissertation, we will assume the following.

Assumption 2.1. *There exists an occupancy measure q°, called feasible solution, such that $\overline{G}^\top q^\circ \leq \theta$, that is, $\rho \geq 0$.*

Assumption 2.1 simply states that the problem is feasible, that is, there exists a safe occupancy measure.

Part I

Stochastic Rewards and Stochastic Constraints

CHAPTER 3

Optimal Strong Regret and Violation via Policy Optimization

In this chapter, we study *online learning* problems in *stochastic* CMDPs, when only *bandit feedback* is available. In such a setting, the classical notions of regret and violation are usually called *weak*, due to the fact that they allow for negative terms to cancel out positive ones. In CMDPs, this means that the (weak) regret can be easily controlled by using policies achieving large rewards *without* satisfying the constraints. Similarly, the (weak) violation can be controlled by adopting policies satisfying cost constraints by a large margin. However, this behavior is most of the times unacceptable in real-world applications. For instance, in autonomous driving, the learner does *not* have the option of being overly safe in some episodes so as to compensate for crashes occurred in previous episodes.

The *strong* regret and the *strong* constraint violation are much more reasonable metrics compared to their weak counterparts, as they do *not* allow negative terms to cancel out positive ones. However, achieving sublinear strong regret/violation in CMDPs is much more challenging.

Efroni et al. (2020) were the first to provide a learning algorithm with (optimal) $\widetilde{\mathcal{O}}(\sqrt{T})$ strong regret/violation in general CMDPs. However, their algorithm works by solving linear programs defined over the space of occupancy measures, a task that is highly inefficient in practice. Ideally, one would like learning algorithms that avoid dealing with occupancy measures, by directly optimizing over the policy space. Such policy optimization algorithms are much more efficient and desirable in practice. By leveraging a primal-dual scheme, Efroni et al. (2020) designed a first policy optimization algorithm for CMDPs, though it can only achieve sublinear *weak* regret and *weak* violation, leaving as an open problem whether an analogous result is achievable for the strong metrics.

Very recently, (Müller et al., 2024) partially addressed this problem by proposing a primal-dual policy optimization algorithm attaining $\widetilde{\mathcal{O}}(T^{0.93})$ *strong* regret and *strong* vi-

olation. However, the bounds achieved by such an algorithm remain largely suboptimal, leaving a big gap that still needs to be closed.

In this chapter, we answer the following question left open by (Efroni et al., 2020; Müller et al., 2024):

*Is it possible to achieve **optimal** $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds on the **strong** regret and the **strong** constraint violation in CMDPs by using an **efficient** primal-dual **policy optimization** algorithm?*

We answer the question above affirmatively. To do so, we design a learning algorithm that exploits a novel primal-dual scheme. Specifically, our algorithm adopts, as primal regret minimizer, a state-of-the-art policy optimization algorithm for adversarial (unconstrained) MDPs, while it leverages an approach based on upper confidence bounds in order to build a dual regret minimizer. Crucially, the updates of dual variables performed by our algorithm do *not* resort to optimizing over the space of occupancy measures, making our algorithm a fully policy optimization approach, and, thus, efficient.

3.1 Setting and Additional Notation

In this chapter, we study CMDPs where rewards and costs are *stochastic*, namely, they are randomly sampled according to some probability distributions $\mathcal{R}$ and $\mathcal{G}_i$, for all $i \in [m]$, whose expected values are $\overline{r}$ and $\overline{g}_i$,for all $i \in [m]$, respectively. As is standard in stochastic RL, we focus on the *bandit feedback* setting, that is, the learner observes the rewards and costs for the path traversed only.

To be inline with the stochastic CMDPs literature, we employ the value functions notation, that is, the cumulative *strong* regret and violation are rewritten as follows:

$$\mathcal{R}_T \coloneqq \sum_{t=1}^{T} \left[\mathrm{OPT}_{\overline{r},\overline{G},\theta} - V^{\pi_t}(\overline{r})\right]^+ \quad ; \quad \mathcal{V}_T \coloneqq \max_{i\in[m]} \sum_{t=1}^{T} \left[V^{\pi_t}(\overline{g}_i) - \theta_i\right]^+ .$$

Similarly, we will rewrite the Lagrangian function $\mathcal{L}_{r,G,\theta} : \Pi \times \mathbb{R}^m_{\geq 0} \to \mathbb{R}$ for a generic CMDP with reward vector $r \in [0,1]^{|X\times A|}$ and cost matrix $G \in [0,1]^{|X\times A|\times m}$, for every policy $\pi \in \Pi$ and vector of Lagrange multipliers $\lambda \in \mathbb{R}^m_{\geq 0}$ as:

$$\mathcal{L}_{r,G,\theta}(\pi,\lambda) \coloneqq V^{\pi}(r) - \sum_{i\in[m]} \lambda_i \left(V^{\pi}(g_i) - \theta_i\right).$$

Finally, throughout the chapter we will assume that Condition 2.1 holds, that is, there exists a strictly feasible policy $\pi^\diamond : X \times A \to [0,1]$ such that $V^{\pi}(\overline{g}_i) < \theta_i$ for every constraint $i \in [m]$. Similarly, it holds $\rho \coloneqq \max_{\pi\in\Pi} \min_{i\in[m]} \left(\theta_i - V^{\pi}(\overline{g}_i)\right) > 0$.

3.2 Parameters Estimation

Let $N_t(x,a)$ be the number of episodes up to $t \in [T]$ in which the pair $(x,a) \in X \times A$ is *visited*. Then, $\widehat{r}_t(x,a) \coloneqq \frac{\sum_{\tau\in[t]} r_\tau(x,a)\mathbb{I}_\tau(x,a)}{\max\{1,N_t(x,a)\}}$, with $\mathbb{I}_\tau(x,a) \in \{0,1\}$ being equal to 1 if and only if (x,a) is visited in episode τ, is an unbiased estimator of the expected reward $\overline{r}(x,a)$. This immediately follows from the fact that $\widehat{r}_t(x,a)$ is defined as the empirical mean of observed rewards for the state-action pair (x,a). Thus, by applying Hoeffding's inequality, the following lemma holds.

Lemma 3.1. *Given a confidence parameter $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for every episode $t \in [T]$ and state-action pair $(x,a) \in X \times A$:*

$$\left|\widehat{r}_t(x,a) - \overline{r}(x,a)\right| \leq \phi_t(x,a), \text{ where } \phi_t(x,a) \coloneqq \min\left\{1, \sqrt{\frac{4\ln(T|X||A|/\delta)}{\max\{1,N_t(x,a)\}}}\right\}.$$

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

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

where $\iota_t(x,a) \coloneqq \sqrt{\frac{\ln(2/\delta')}{2N_t(x,a)}}$.
Now, we are interested in the intersection of the aforementioned events:

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

Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,t}\left\{\left|\widehat{r}_t(x,a) - \overline{r}(x,a)\right| \leq \iota_t(x,a)\right\}\right] && (3.1)\\
&= 1 - \mathbb{P}\left[\bigcup_{x,a,t}\left\{\left|\widehat{r}_t(x,a) - \overline{r}(x,a)\right| \leq \iota_t(x,a)\right\}^c\right]\\
&\geq 1 - \sum_{x,a,t}\mathbb{P}\left[\left\{\left|\widehat{r}_t(x,a) - \overline{r}(x,a)\right| \leq \iota_t(x,a)\right\}^c\right] && (3.2)\\
&\geq 1 - |X||A|T\delta',
\end{aligned}$$

where Inequality (3.2) holds by Union Bound. Noticing that $r_t(x,a) \leq 1$, substituting δ' with $\delta \coloneqq \delta'/|X||A|T$ in $\iota_t(x,a)$ with an additional Union Bound over the possible values of $N_t(x,a)$, and thus obtaining $\phi_t(x,a)$, concludes the proof. □

Similarly, $\widehat{g}_{t,i}(x,a) \coloneqq \frac{\sum_{\tau\in[t]} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a)}{\max\{1,N_t(x,a)\}}$ is clearly an unbiased estimator of the expected cost $\overline{g}_i(x,a)$. Thus, by applying Hoeffding's inequality, it is possible to show the following lemma.

Lemma 3.2. *Given a confidence parameter $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for every $i \in [m]$, episode $t \in [T]$, and state-action pair $(x,a) \in X \times A$:*

$$\left|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)\right| \leq \xi_t(x,a), \text{ where } \xi_t(x,a) \coloneqq \min\left\{1, \sqrt{\frac{4\ln(T|X||A|m/\delta)}{\max\{1,N_t(x,a)\}}}\right\}.$$

Proof. The proof is analogous to the one of Lemma 3.1, with an additional Union Bound over the m constraints. □

Moreover, by letting $M_t(x,a,x')$ be the total number of episodes up to $t \in [T]$ in which the state-action pair $(x,a) \in X \times A$ is visited and the environment evolves to the new state $x' \in X$, the estimated transition probability for the triplet (x,a,x') is $\widehat{P}_t(x'|x,a) \coloneqq$

$\frac{M_t(x,a,x')}{\max\{1,N_t(x,a)\}}$. We refer to Appendix A.5 for additional details and results related to transition probabilities estimation.

3.2.1 Compact Notation

We introduce $\widehat{r}_t \in [0,1]^{|X\times A|}$ to denote the vector whose components are the estimated rewards $\widehat{r}_t(x,a)$. Moreover, we denote by $\phi_t \in [0,1]^{|X\times A|}$ the vector whose entries are the bounds $\phi_t(x,a)$. Similarly, we introduce $\widehat{g}_{t,i} \in [0,1]^{|X\times A|}$ to denote the vector of estimated costs $\widehat{g}_{t,i}(x,a)$, while we denote by $\xi_t \in [0,1]^{|X\times A|}$ the vector of the bounds $\xi_t(x,a)$. Finally, we let $\widetilde{r}_t := \widehat{r}_t + \phi_t$ and $\widetilde{g}_{t,i} := \widehat{g}_{t,i} - \xi_t$.

3.3 A Novel Primal-Dual Algorithm

Next, we introduce a novel primal-dual algorithm, called *constrained primal-dual policy optimization* (CPD-PO), which allows to efficiently achieve $\widetilde{O}(\sqrt{T})$ *strong* regret and *strong* violation.

3.3.1 The CPD-PO Algorithm

Algorithm 3.1 Constrained Primal-Dual Policy Optimization (CPD-PO)

Require: number of rounds $T \in \mathbb{N}_{>0}$, problem-specific parameter $\rho \in [0,L]$, confidence $\delta \in (0,1)$

$\pi_1 \leftarrow$ first policy prescribed by PO-DB
Initialize all the estimators, counters, and bounds
for $t = 1,\dots,T$ **do**
  Interact as in Algorithm 2.1
  Observe (x_k,a_k), $r_t(x_k,a_k)$, and $g_{t,i}(x_k,a_k)$ for every $k \in [0,\dots,L-1]$ and $i \in [m]$, as feedback from the interaction in Algorithm 2.1
  Build estimators $\widehat{r}_t$, $\widehat{g}_{t,i}$, ϕ_t, ξ_t, and $\widehat{P}_t$ as prescribed in Section 3.2
  $\lambda_{t,i} \leftarrow \arg\max_{\lambda \in \left\{0,\frac{L+1}{\rho}\right\}} \lambda\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right), \quad \forall i \in [m]$
  Build artificial loss for every pair (x_k,a_k) received as feedback:

$$\ell_t(x_k,a_k) \leftarrow \frac{(L+1)m}{\rho} - \left[\widetilde{r}_t(x_k,a_k) - \sum_{i\in[m]} \lambda_{t,i}\left(\widetilde{g}_{t,i}(x_k,a_k) - \frac{\theta_i}{L}\right)\right]$$

  Update policy $\pi_{t+1} \leftarrow$ PO-DB$\left(\{(x_k,a_k),\ell_t(x_k,a_k)\}_{k=0}^{L-1}\right)$
end for

Algorithm 3.1 shows the pseudocode of CPD-PO. It employs an instance of the *policy optimization with dilated bonus* (PO-DB) algorithm by Luo et al. (2021), which is the state-of-the-art algorithm for learning in adversarial (unconstrained) MDPs under *bandit* feedback. Notice that this algorithm employs a policy optimization approach, by efficiently optimizing state by state. Thus, PO-DB does *not* resort to any optimization step performed over the space of occupancy measures.

Algorithm 3.1 initializes an instance of PO-DB as prescribed in (Luo et al., 2021), and it immediately uses it to get the policy π_1 to be used at $t = 1$. Furthermore, it initializes

the estimators $\widehat{r}_t$, $\widehat{g}_{t,i}$ to vectors of zeros and all the counters $N_t(x,a), M_t(x,a,x')$ to zero (Line 2). Then, at every episode $t \in [T]$, the algorithm plays policy π_t (Line 4) and observes the feedback associated with the trajectory traversed during the episode (Line 5). The estimators and the confidence bounds are then updated as shown in Section 3.2 (Line 6). The update of dual variables (Line 7) is performed as a binary choice between two values, namely zero and $^{L+1}/_{\rho}$, for each $i \in [m]$. The value zero is selected when the optimistic estimation of the i-th constraint is *not* violated by the selected policy π_t (with respect to the estimated transition $\widehat{P}_t$). In such a case, in the next primal update, the algorithm will *not* focus on minimizing that specific constraint violation. On the contrary, if the optimistic estimation of the i-th constraint is *not* satisfied, then the dual update selects the value $^{L+1}/_{\rho}$. This quantity is chosen to be large enough to guarantee that, with respect to the deterministic Lagrangian game defined by the true reward and cost distributions, any policy cannot gain more than $\text{OPT}_{\overline{r},\overline{G},\theta}$ (see Section 3.4.1 for further discussion on these aspects). We remark that the value of $V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i})$ in Line 7 of Algorithm 3.1 can be efficiently computed by means of a simple dynamic programming procedure. The primal update is performed by the policy update of `PO-DB`. Notice that `PO-DB` is tailored for adversarial MDPs (in which no-statistical assumption is made on the loss functions) with bandit feedback. Thus, for every $t \in [T]$, the primal algorithm expects to receive a loss value for any state-action traversed by the policy previously chosen (and the trajectory itself). We feed `PO-DB` by building a Lagrangian loss, employing the Lagrangian vector selected in Line 7. Notice that the estimated Lagrangian is subtracted and scaled by the state-action maximum (negative) loss $^{(L+1)m}/_{\rho}$, since `PO-DB` is tailored for positive loss (see Line 8). Finally, the Lagrangian loss and the trajectory are given to the `PO-DB` instance (Line 9). We remark that `PO-DB` minimizes the loss function by employing state-by-state optimization updates. Thus, it is computationally much more efficient than algorithms resorting to a projection on the occupancy measure space.

3.3.2 Algorithm Comparison with (Efroni et al., 2020) and (Müller et al., 2024)

In the following, we highlight the main differences between Algorithm 3.1 and the primal-dual formulations employed by Efroni et al. (2020) and Müller et al. (2024).

Efroni et al. (2020) were the first to introduce online primal-dual methods that achieve sublinear regret and sublinear constraint violation in CMDPs (allowing for cancellations). Müller et al. (2024) were the first to achieve sublinear *strong* regret and *strong* violation via primal-dual methods.

Efroni et al. (2020) propose two primal-dual algorithms.[1] The first algorithm, that is `OptDual-CMDP`, performs a UCB-like primal update given an optimistic estimation of the Lagrangian function. Such estimation shares some similarities with ours. As concerns the dual update, `OptDual-CMDP` performs a gradient descent update on the optimistic Lagrangian. The second algorithm (`OptPrimalDual-CMDP`) employs a multiplicative-weight-kind of update for the primal update. Precisely, this update is performed in closed form given the Lagrangified action-value function as objective. The dual update is gradient-descent-like, with the additional modification that the Lagrangian multipliers space is bounded (similarly to our work) since the primal regret minimizer needs the loss/reward

[1] It is important to highlight that `OptDual-CMDP` is often referred to as a *dual* method since the primal is only updated given a UCB-like procedure on the Lagrangian. Nevertheless, since the difficulty in attaining *strong* regret and violation holds for dual methods, too, we will refer to this kind of algorithms as primal-dual.

functions to be bounded.

Müller et al. (2024) employ similar primal and dual regret minimizers as the ones of the `OptPrimalDual-CMDP` algorithm, namely, multiplicate weights for the primal and gradient descent for the dual. Differently from Efroni et al. (2020), Müller et al. (2024) employ a regularized scheme to define the underlying Lagrangian game played by the primal and the dual regret minimizers. Indeed, the algorithm employs a regularized Lagrangian formulation in order to make the primal objective strictly concave (in the state-action visit distribution) and the dual one strongly convex (in the Lagrangian variable). The strict concavity/strong convexity is necessary to converge in last-iterate to the optimal values of the Lagrangian game and, thus, to attain sublinear *strong* regret and *strong* violation. In our work, we do *not* resort to any regularization scheme, thus simply employing the standard Lagrangian formulation of CMDPs. The main difference between our algorithm and those in (Efroni et al., 2020; Müller et al., 2024) lies in the dual update. Indeed, the black-box primal update employed in our work is multiplicative weight like as the works described above.[2] Differently, in the dual update, we use a UCB-like update, thus not resorting to any adversarial regret minimizers. The UCB-like kind of update is performed between the minimum and the maximum reasonable value for the Lagrange variables. This modification allows us to play the adversarial primal regret minimizer on the deterministic Lagrangian game (up to confidence terms factor) associated with the "best" Lagrangian variable for the violations previously attained. Moreover, notice that our algorithm, since we employ an adversarial primal regret minimizer and differently from algorithms that employ UCB-like updates on the primal (such as `OptDual-CMDP`), may choose non-deterministic policies, which are often optimal in online constrained problem.

3.4 Theoretical Analysis

In the following section, we discuss the theoretical guarantees attained by Algorithm 3.1. Specifically, in Section 3.4.1 we provide fundamental results on the Lagrangian formulation employed by `CPD-PO`. In Section 3.4.2 we discuss the guarantees attained by the primal algorithm. Finally, in Section 3.4.3 we state the regret and violations guarantees attained by Algorithm 3.1.

3.4.1 Results on the Lagrangian Formulation

In this section, we state some useful preliminary results attained by Algorithm 3.1. Specifically, the following results concern the primal-dual scheme employed by `CPD-PO`. We start by showing that the dual variables decision space is well-defined. Indeed, it is fundamental to show that, given any policy π_t selected by the black-box primal algorithm, the dual decision space is sufficient to upper-bound the Lagrangian function value with the constrained optimum. This is done by means of the following lemma.

Lemma 3.3. *Given a CMDP with reward vector $r \in [0,1]^{|X \times A|}$ and cost matrix $G \in [0,1]^{|X \times A| \times m}$, for every policy $\pi \in \Pi$, the following holds:*

$$V^{\pi}(r) - \max_{\lambda \in \left[0, \frac{L+1}{\rho}\right]^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi}(g_i) - \theta_i\right) \leq \mathrm{OPT}_{r,G,\theta}.$$

[2] Nevertheless, it is an enhanced version with exploration bonus, which allows to have independent adversarial regret minimizer guarantees (see (Luo et al., 2021) for further discussion)

Proof. We analyze separately the case in which π satisfies the constraints *(i)* and the case π is not safe *(ii)*.
(i). We start analyzing π s.t. $V^{\pi}(g_i) \leq \theta_i$ for all $i \in [m]$. In such a case, it is easy to check that:

$$\arg\max_{\lambda \in \left[0, \frac{L+1}{\rho}\right]^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi}(g_i) - \theta_i\right) = \underline{0}.$$

Thus, employing Lemma A.3 gives the result.
(ii). We then consider the case π is not safe. In such a case, π may either partially satisfy the constraints or violating all of the them.
First of all we notice that, in such a scenario:

$$\arg\max_{\lambda \in \left[0, \frac{L+1}{\rho}\right]^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi}(g_i) - \theta_i\right) = \bar{\lambda},$$

where $\bar{\lambda}$ is the Lagrangian vector composed by 0 values for the constraints which are violated and $^{L+1}/_{\rho}$ values for the others.
Indeed, it holds:

$$\begin{aligned}
\mathcal{L}_{r,G,\theta}\left(\pi, \bar{\lambda}\right) &= V^{\pi}(r) - \sum_{i \in [m]} \bar{\lambda}_i \left(V^{\pi}(g_i) - \theta_i\right) \\
&\leq \max_{\pi \in \Pi} V^{\pi}(r) - \frac{L+1}{\rho} \sum_{i \in [m]} \left[V^{\pi}(g_i) - \theta_i\right]^+ && \text{(3.3a)} \\
&\leq \max_{\pi \in \Pi} V^{\pi}(r) - \frac{L}{\rho} \sum_{i \in [m]} \left[V^{\pi}(g_i) - \theta_i\right]^+ \\
&\leq \max_{\pi \in \Pi} \min_{\|\lambda\|_1 \in [0, L/\rho]} V^{\pi}(r) - \sum_{i \in [m]} \lambda_i \left[V^{\pi}(g_i) - \theta_i\right]^+ \\
&\leq \min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{\pi \in \Pi} V^{\pi}(r) - \sum_{i \in [m]} \lambda_i \left[V^{\pi}(g_i) - \theta_i\right]^+ && \text{(3.3b)} \\
&\leq \min_{\|\lambda\|_1 \in [0, L/\rho]} \max_{\pi \in \Pi} \mathcal{L}_{r,G,\theta}\left(\pi, \lambda\right) \\
&= \mathrm{OPT}_{r,G,\theta}, && \text{(3.3c)}
\end{aligned}$$

where the Inequality (3.3a) holds by definition of $\bar{\lambda}$, the Inequality (3.3b) holds by the *max-min inequality* and Inequality (3.3c) follows from Lemma A.2.
This concludes the proof. □

As previously specified, Lemma 3.3 states that it is not convenient for the primal algorithm instantiated by Algorithm 3.1 to play non-safe policies in order to gain more rewards than the one possibly attained by safe policies. Again, this is true given the dual update performed by `CPD-PO`.

It is important to notice that Lemma 3.3 holds for an optimization problem where rewards, constraints and transitions are known a-priori. Indeed, this is not the case in an online learning setting, where all the aforementioned parameter are unknown and must be estimated in an online fashion. Specifically, if the dual algorithm were aware of the unknown distributions mentioned above, the dual update of Algorithm 3.1 combined with

Lemma 3.3 would lead to:

$$\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) \geq \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, \lambda_t), \tag{3.4}$$

for all $t \in [T]$. Equation (3.4) would be crucial to prove *strong* regret guarantees, since, it shows that the policy chosen by the algorithm cannot outperform π^*, thus, making the standard regret definition (on the Lagrangian) to collapse to the *strong* one. Nevertheless, as previously specified, Equation (3.4) cannot be attained without complete knowledge of the environment. To overcome this issue, we prove the Equation (3.4) holds *up to* sublinear term, which depends on the uncertainty on the environment estimation. This is done in the following lemma.

Lemma 3.4. *With probability at least $1-\delta$, Algorithm 3.1 guarantees that, for every episode $t \in [T]$:*

$$\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) \geq \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, \lambda_t) - \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) - \frac{8Lm}{\rho} \|q^{\widehat{P}_t,\pi_t} - q^{P,\pi_t}\|_1.$$

Proof. First of all, we notice that, when $\lambda_t = \underline{0}$, $\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) = \text{OPT}_{\overline{r},\overline{G},\theta}$ by definition. Furthermore when λ_t is the $\frac{L+1}{\rho}$ vector, it holds that $\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) \geq \text{OPT}_{\overline{r},\overline{G},\theta}$ since π^* is feasible. Same reasoning holds when λ_t is any vector in $\left\{0, \frac{L+1}{\rho}\right\}^m$. Thus, it is always the case that $\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) \geq \text{OPT}_{\overline{r},\overline{G},\theta}$.
Thus, we proceed as follows:

$$\begin{aligned}
&\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, \lambda_t) \\
&\leq V^{\pi_t}(\overline{r}) - \sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i\right) && \text{(3.5a)} \\
&= V^{\pi_t}(\overline{r}) - \sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i\right) \pm \sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) \\
&\leq V^{\pi_t}(\overline{r}) - \sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) \\
&\qquad + \frac{4Lm}{\rho} \sum_{x\in X, a\in A} \left|q^{\widehat{P},\pi_t}(x,a) - q^{P,\pi_t}(x,a)\right| && \text{(3.5b)} \\
&= V^{\pi_t}(\overline{r}) - \max_{\lambda\in\left\{0,\frac{L+1}{\rho}\right\}^m} \sum_{i\in[m]} \lambda_i \left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) \\
&\qquad + \frac{4Lm}{\rho} \sum_{x\in X, a\in A} \left|q^{\widehat{P},\pi_t}(x,a) - q^{P,\pi_t}(x,a)\right| \\
&\leq V^{\pi_t}(\overline{r}) - \max_{\lambda\in\left\{0,\frac{L+1}{\rho}\right\}^m} \sum_{i\in[m]} \lambda_i \left(V^{\pi_t,\widehat{P}_t}(\overline{g}_i - 2\xi_t) - \theta_i\right) \\
&\qquad + \frac{4Lm}{\rho} \sum_{x\in X, a\in A} \left|q^{\widehat{P},\pi_t}(x,a) - q^{P,\pi_t}(x,a)\right| && \text{(3.5c)} \\
&= V^{\pi_t}(\overline{r}) - \max_{\lambda\in\left\{0,\frac{L+1}{\rho}\right\}^m} \sum_{i\in[m]} \lambda_i \left(V^{\pi_t,\widehat{P}_t}(\overline{g}_i - 2\xi_t) - \theta_i\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{4Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right| \\
&\pm \max_{\lambda \in \left\{0, \frac{L+1}{\rho}\right\}^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi_t}\left(\overline{g}_i - 2\xi_t\right) - \theta_i \right) \\
&\leq V^{\pi_t}(\overline{r}) - \max_{\lambda \in \left\{0, \frac{L+1}{\rho}\right\}^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi_t}\left(\overline{g}_i - 2\xi_t\right) - \theta_i \right) \\
&\quad + \frac{8Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right| \qquad (3.5d) \\
&\leq V^{\pi_t}(\overline{r}) - \max_{\lambda \in \left\{0, \frac{L+1}{\rho}\right\}^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi_t}\left(\overline{g}_i\right) - \theta_i \right) + \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\quad + \frac{8Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right| \qquad (3.5e) \\
&= V^{\pi_t}(\overline{r}) - \max_{\lambda \in \left[0, \frac{L+1}{\rho}\right]^m} \sum_{i \in [m]} \lambda_i \left(V^{\pi_t}\left(\overline{g}_i\right) - \theta_i \right) + \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\quad + \frac{8Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right| \\
&\leq \mathrm{OPT}_{\overline{r}, \overline{G}, \theta} + \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) + \frac{8Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right|, \qquad (3.5f)
\end{aligned}
$$

where Inequality (3.5a) holds with probability at least $1 - \delta$ by Lemma 3.2, Inequality (3.5b) holds by Hölder inequality, Inequality (3.5c) holds with probability at least $1 - \delta$ by Lemma 3.2, Inequality (3.5d) holds by Hölder inequality, Inequality (3.5e) holds by definition of λ_t and Inequality (3.5f) holds by Lemma 3.3.
Thus, it holds:

$$
\begin{aligned}
&\mathcal{L}_{\overline{r}, \overline{G}, \theta}(\pi^*, \lambda_t) \\
&\geq \mathrm{OPT}_{\overline{r}, \overline{G}, \theta} \\
&\geq \mathcal{L}_{\overline{r}, \overline{G}, \theta}(\pi_t, \lambda_t) - \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) - \frac{8Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right|,
\end{aligned}
$$

which concludes the proof. □

The interpretation of Lemma 3.4 is the following: Equation (3.4) holds up to two terms. The first term $4Lm/\rho \cdot V^{\pi_t}(\xi_t)$ encompasses the uncertainty on the constraints. Indeed, it is the expectation over policy π_t and transitions of the confidence intervals on the costs. This quantity decreases as the algorithm collects cost samples, thus it is sublinear in T when summed over the episodes. Similarly, the term

$$
\frac{8Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}_t, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right|
$$

encompasses the uncertainty related to the transitions. This term is sublinear in T when

summed over the episode. Finally, we remark that the uncertainty about the rewards does not affect the dual update, thus, Lemma 3.4 is completely independent on that.

3.4.2 Primal Algorithm

In the following section, we state the theoretical guarantees attained by the primal algorithm employed by Algorithm 3.1. Precisely, we have the following guarantees.

Lemma 3.5. *Given any $\delta \in (0,1)$, the instance of* PO-DB *employed by Algorithm 3.1 guarantees that, for every policy $\pi \in \Pi$, the primal regret $R_T^{\mathcal{L}}(\pi)$ defined as:*

$$\sum_{t=1}^{T}\left[V^{\pi}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi}(\widetilde{g}_{t,i}) - \theta_i\right)\right] - \sum_{t=1}^{T}\left[V^{\pi_t}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i\right)\right],$$

is bounded as follows:

$$R_T^{\mathcal{L}}(\pi) \leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right),$$

with probability at least $1 - \mathcal{O}(\delta)$.

Lemma 3.5 is obtained by the following considerations. First of all, the analysis employs the no-regret guarantees of PO-DB. Indeed, by simple computation and using the linearity of expectation, it is possible to recover the $R_T^{\mathcal{L}}(\pi)$ definition by the loss function fed into PO-DB by Algorithm 3.1. Moreover, notice that the no-regret property of PO-DB cannot be directly applied to our setting, since the range of the losses is clearly different. Indeed PO-DB is tailored for standard episodic MDPs where the losses are bounded in $[0, L]$. This is not true for the Lagrangified MDPs where the losses are bounded in $[0, {}^{2L(L+1)m}/_{\rho}]$. Nevertheless, it is sufficient to multiply the original bound for a $\mathcal{O}({}^{Lm}/_{\rho})$ factor to obtain the result.

Proof. We first notice that, by simple computation, it holds:

$$\begin{aligned}
&\sum_{t=1}^{T} V^{\pi_t}\left(\frac{(L+1)m}{\rho}\right) - \sum_{t=1}^{T}\left[V^{\pi_t}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}V^{\pi_t}\left(\widetilde{g}_{t,i} - \frac{\theta_i}{L}\right)\right] \\
&\qquad - \sum_{t=1}^{T} V^{\pi}\left(\frac{(L+1)m}{\rho}\right) - \sum_{t=1}^{T}\left[V^{\pi}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}V^{\pi}\left(\widetilde{g}_{t,i} - \frac{\theta_i}{L}\right)\right] \\
&= -\sum_{t=1}^{T}\left[V^{\pi_t}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}V^{\pi_t}\left(\widetilde{g}_{t,i} - \frac{\theta_i}{L}\right)\right] \\
&\qquad + \sum_{t=1}^{T}\left[V^{\pi}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}V^{\pi}\left(\widetilde{g}_{t,i} - \frac{\theta_i}{L}\right)\right] \\
&= \sum_{t=1}^{T}\left[V^{\pi}(\widetilde{r}_t) - \sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi}(\widetilde{g}_{t,i}) - \theta_i\right)\right]
\end{aligned}$$

$$- \sum_{t=1}^{T} \left[V^{\pi_t}(\widetilde{r}_t) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i \right) \right],$$

where the first step holds since $(L+1)m/\rho$ is constant for any state-action pair and the second step holds since $V^{\pi}(\theta_i/L) = \theta_i$.
Then, we notice that, any regret minimizer algorithm $\mathcal{A}$, which attains $\mathcal{E}_{\mathcal{A}}$ regret upper-bound when the losses are bounded in $[0, 1]$ (or in $[0, L]$ for MDPs), may achieve $C \cdot \mathcal{E}_{\mathcal{A}}$ regret upper-bound when the range of the losses is C and it is known to the algorithm. The result is intuitive, since it is always possible to apply an affine transformation to the losses received and, thus, to scale them in $[0, 1]$. Then, the algorithm can be fed with the scaled loss as expected, but it will suffer a loss multiplied by a C factor, attaining the aforementioned $C \cdot \mathcal{E}_{\mathcal{A}}$ bound.
Noticing that the loss vector received by PO-DB is multiplied by an additional $\mathcal{O}(Lm/\rho)$ factor given by the definition of the Lagrangian variable in Algorithm 3.1, we can employ the standard regret bound PO-DB (see Lemma A.7) with the additional $\mathcal{O}(Lm/\rho)$ factor. Thus, it holds, with probability at least $1 - \mathcal{O}(\delta)$:

$$\begin{aligned}
&\sum_{t=1}^{T} \left[V^{\pi}(\widetilde{r}_t) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi}(\widetilde{g}_{t,i}) - \theta_i \right) \right] \\
&\qquad - \sum_{t=1}^{T} \left[V^{\pi_t}(\widetilde{r}_t) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i \right) \right] \\
&= \sum_{t=1}^{T} V^{\pi_t} \left(\frac{(L+1)m}{\rho} \right) - \sum_{t=1}^{T} \left[V^{\pi_t}(\widetilde{r}_t) - \sum_{i \in [m]} \lambda_{t,i} V^{\pi_t} \left(\widetilde{g}_{t,i} - \frac{\theta_i}{L} \right) \right] \\
&\qquad - \sum_{t=1}^{T} V^{\pi} \left(\frac{(L+1)m}{\rho} \right) - \sum_{t=1}^{T} \left[V^{\pi}(\widetilde{r}_t) - \sum_{i \in [m]} \lambda_{t,i} V^{\pi} \left(\widetilde{g}_{t,i} - \frac{\theta_i}{L} \right) \right] \\
&\leq \widetilde{\mathcal{O}} \left(\frac{L^3 m}{\rho} |X| \sqrt{|A|T} + \frac{L^5 m}{\rho} \right),
\end{aligned}$$

which concludes the proof. □

3.4.3 Regret and Violation

In the following section, we provide the cumulative *strong* regret and cumulative *strong* violation guarantees attained by Algorithm 3.1.

We start showing that a regret of order $\widetilde{\mathcal{O}}(\sqrt{T})$ is attainable by employing primal-dual method which does not optimize over the occupancy measure space. This is done in the following theorem.

Theorem 3.1. *Given any $\delta \in (0, 1)$, Algorithm 3.1 attains:*

$$\mathcal{R}_T \leq \widetilde{\mathcal{O}} \left(\frac{L^3 m}{\rho} |X| \sqrt{|A|T} + \frac{L^5 m}{\rho} \right),$$

with probability at least $1 - \mathcal{O}(\delta)$.

Theorem 3.1 is proved by combining Lemma 3.4 and Lemma 3.5. Specifically, we start from the theoretical guarantees attained by PO-DB, that is, using the bound provided in Lemma 3.5. Next, employing the optimism of the confidence bound, we recover the true Lagrangian function, excluding sublinear terms. As previously specified, given Lemma 3.4 it is then possible to recover the cumulative *strong* regret definition on the regret formulated with respect to the Lagrangian function. Finally, to get the final bound, it is sufficient to notice that the optimal solution π^* is safe and that:

$$\sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi_t}(\overline{g}_i) - \theta_i\right) \geq -\frac{4Lm}{\rho} \sum_{x\in X, a\in A} \left| q^{\widehat{P}_t,\pi_t}(x,a) - q^{P,\pi_t}(x,a) \right|.$$

Proof. We first notice that, by Lemma 3.4, we have, for all $t \in [T]$:

$$\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) \geq \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t,\lambda_t) - \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) - \frac{8Lm}{\rho} \|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1,$$

with probability at least $1-\delta$ and where π^* is the optimal solution corresponding to $\text{OPT}_{\overline{r},\overline{G},\theta}$. Thus, we have that an upper bound on:

$$\sum_{t=1}^{T} \left(\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) - \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t,\lambda_t) + \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) + \frac{8Lm}{\rho} \|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \right) \tag{3.6}$$

enforces the same bound on

$$\sum_{t=1}^{T} \left[\mathcal{L}_{r,G}(\pi^*,\lambda_t) - \mathcal{L}_{r,G}(\pi_t,\lambda_t) + \frac{4Lm}{\rho} V^{\pi_t}(\xi_t) + \frac{8Lm}{\rho} \|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \right]^+,$$

since the term in the summation is always non-negative.
Now, we employ Lemma 3.5, which implies that with probability at least $1 - \mathcal{O}(\delta)$, for any $\pi \in \Pi$:

$$\begin{aligned} &\sum_{t=1}^{T} \left(V^{\pi}(\widetilde{r}_t) - \sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi}(\widetilde{g}_{t,i}) - \theta_i\right) \right) \\ &\quad - \sum_{t=1}^{T} \left(V^{\pi_t}(\widetilde{r}_t) - \sum_{i\in[m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i\right) \right) \\ &\quad \leq \widetilde{\mathcal{O}} \left(\frac{L^3 m}{\rho} |X| \sqrt{|A|T} + \frac{L^5 m}{\rho} \right), \end{aligned} \tag{3.7}$$

which implies the same bound for π^*.
We now show that the left-hand side of the aforementioned equation reduce to Equation (3.6) up to sublinear terms. By Lemma 3.1,with probability at least $1-\delta$, it holds:

$$\overline{r} \preceq \widetilde{r}_t.$$

Thus, taking the expectation over policy and transition we obtain:

$$V^{\pi}(\overline{r}) \leq V^{\pi}\left(\widetilde{r}_t\right).$$

Thus, Equation (3.7) can be rewritten as:

$$\sum_{t=1}^{T}\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\widetilde{g}_{t,i})-\theta_i\right)\right)$$
$$-\sum_{t=1}^{T}\left(V^{\pi_t}(\widetilde{r}_t)-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\widetilde{g}_{t,i})-\theta_i\right)\right)$$
$$\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right).$$

Similarly, it holds that $\widehat{r}_t\preceq\overline{r}+\phi_t$ which implies:

$$\sum_{t=1}^{T}\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\widetilde{g}_{t,i})-\theta_i\right)\right)$$
$$-\sum_{t=1}^{T}\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\widetilde{g}_{t,i})-\theta_i\right)\right)$$
$$\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right)+2\sum_{t=1}^{T}V^{\pi_t}(\phi_t).$$

We apply Lemma A.5 to obtain, with probability at least $1-\mathcal{O}(\delta)$, by Union Bound:

$$\sum_{t=1}^{T}\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\widetilde{g}_{t,i})-\theta_i\right)\right)$$
$$-\sum_{t=1}^{T}\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\widetilde{g}_{t,i})-\theta_i\right)\right)$$
$$\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right).$$

We proceed similarly for the constraints cost noticing that $\widetilde{g}_{t,i}\preceq\overline{g}_i$, and $\overline{g}_i-2\xi_t\preceq\widetilde{g}_{t,i}$ for all $i\in[m]$. Thus, we have:

$$\sum_{t=1}^{T}\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i)-\theta_i\right)\right)$$
$$-\sum_{t=1}^{T}\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\right)$$
$$\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right)+2m\sum_{t=1}^{T}V^{\pi_t}(\xi_t).$$

Finally, employing Lemma A.6, we have that the regret defined over the Lagrangian is bounded by,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i)-\theta_i\right)\right)\\
&\quad-\sum_{t=1}^{T}\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\right)\\
&\quad\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right).
\end{aligned}$$

Now, we notice the following inequality,

$$\begin{aligned}
&\sum_{t=1}^{T}\Bigg(\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i)-\theta_i\right)\right)\\
&\quad-\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\right)\\
&\quad\pm\left(\frac{4Lm}{\rho}V^{\pi_t}(\xi_t)+\frac{8Lm}{\rho}\|q^{\widehat{P},\pi_t}-q^{P,\pi_t}\|_1\right)\Bigg)\\
&\quad\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right),
\end{aligned}$$

which implies the following chain of inequalities,

$$\begin{aligned}
&\sum_{t=1}^{T}\Bigg(\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i)-\theta_i\right)\right)\\
&\qquad-\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\right)\\
&\qquad+\frac{4Lm}{\rho}V^{\pi_t}(\xi_t)+\frac{8Lm}{\rho}\|q^{\widehat{P},\pi_t}-q^{P,\pi_t}\|_1\Bigg)\\
&\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right)+\sum_{t=1}^{T}\frac{4Lm}{\rho}V^{\pi_t}(\xi_t)\\
&\qquad+\sum_{t=1}^{T}\frac{8Lm}{\rho}\|q^{\widehat{P},\pi_t}-q^{P,\pi_t}\|_1\\
&\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right),
\end{aligned}$$

where in the last step we employed Lemma A.6 and Lemma A.9.

We then employ Lemma 3.4 to get the following bound:

$$\sum_{t=1}^{T}\left[\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t)-\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t,\lambda_t)+\frac{4Lm}{\rho}V^{\pi_t}(\xi_t)+\frac{8Lm}{\rho}\|q^{\widehat{P},\pi_t}-q^{P,\pi_t}\|_1\right]^+ \leq \widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right),$$

which, since $\frac{4Lm}{\rho}V^{\pi_t}(\xi_t)$ and $\frac{8Lm}{\rho}\left|q^{\widehat{P},\pi_t}(x,a)-q^{P,\pi_t}(x,a)\right|$ are non-negative by construction, implies:

$$\begin{aligned}\sum_{t=1}^{T}\Bigg[&\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i)-\theta_i\right)\right)\\&-\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\right)\Bigg]^+\\&\leq\widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T}+\frac{L^5m}{\rho}\right).\end{aligned}$$

To get the final bound, we first notice that, by definition of the constrained optimum π^*, the quantity $V^{\pi^*}(\overline{g}_i)-\theta_i$ is always non-positive for every $i\in[m]$. Moreover, by the Lagrangian update of Algorithm 3.1, we have that:

$$\begin{aligned}&\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\\&\quad\geq\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\widetilde{g}_{t,i})-\theta_i\right)\\&\quad=\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\widetilde{g}_{t,i})-\theta_i\right)\pm\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i})-\theta_i\right)\\&\quad\geq\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i})-\theta_i\right)-\frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t}-q^{P,\pi_t}\|_1\\&\quad\geq-\frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t}-q^{P,\pi_t}\|_1.\end{aligned}$$

Thus, following the reasoning above, we bound the strong cumulative regret as follows:

$$\begin{aligned}\mathcal{R}_T&:=\sum_{t=1}^{T}\left[V^{\pi^*}(\overline{r})-V^{\pi_t}(\overline{r})\right]^+\\&\leq\sum_{t=1}^{T}\Bigg[\left(V^{\pi^*}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i)-\theta_i\right)\right)\\&\qquad-\left(V^{\pi_t}(\overline{r})-\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i)-\theta_i\right)\right)\end{aligned}$$

$$
\begin{aligned}
&+ \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1\Bigg]^+ \\
&\leq \sum_{t=1}^{T}\Bigg[\left(V^{\pi^*}(\overline{r}) - \sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi^*}(\overline{g}_i) - \theta_i\right)\right) \\
&\quad - \left(V^{\pi_t}(\overline{r}) - \sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i) - \theta_i\right)\right)\Bigg]^+ \\
&\quad + \sum_{t=1}^{T}\left[\frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1\right]^+ \\
&\leq \widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T} + \frac{L^5m}{\rho}\right) + \sum_{t=1}^{T}\frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \\
&\leq \widetilde{\mathcal{O}}\left(\frac{L^3m}{\rho}|X|\sqrt{|A|T} + \frac{L^5m}{\rho}\right),
\end{aligned}
$$

which holds with probability $1-\mathcal{O}(\delta)$ by Lemma A.9 and Union Bound. This concludes the proof. □

We underline that, from a regret bound perspective, our result obtains the optimal regret order in T, while Müller et al. (2024) achieves a suboptimal $T^{0.93}$ only. Furthermore, our bound is not worse than the one in (Müller et al., 2024) w.r.t. the dependency on ρ, m and L.

We conclude by stating the result related to the cumulative *strong* violations. Even in this case, we show that the $\widetilde{\mathcal{O}}(\sqrt{T})$ order is attainable.

Theorem 3.2. *For any $\delta \in (0,1)$, Algorithm 3.1 attains:*

$$
\mathcal{V}_T \leq \widetilde{\mathcal{O}}\left(L^3m|X|\sqrt{|A|T} + L^5m\right),
$$

with probability at least $1-\mathcal{O}(\delta)$.

To prove the result, we proceed similarly to Theorem 3.1, namely, we employ Lemma 3.5 and we apply the following equality

$$
\begin{aligned}
&\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) = \\
&\qquad \frac{L}{L+1}\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) + \frac{1}{\rho}\sum_{i\in[m]}\left[V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right]^+
\end{aligned}
$$

to retrieve the *strong* violation definition (up to confidence terms). Notice that, given the $L/L+1$ factor which allows us to retrieve the *strong* violation term, it is not possible to directly apply Lemma 3.4 anymore. Nevertheless, it is indeed possible to prove a similar inequality. Specifically, it holds:

$$
\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) \geq \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, \lambda_t L/L+1)
$$

$$- \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) - \frac{4Lm}{\rho} \sum_{x \in X, a \in A} \left| q^{\widehat{P}, \pi_t}(x,a) - q^{P, \pi_t}(x,a) \right|.$$

Indeed, the inequality above can be employed to bound sublinearly the following quantity:

$$\sum_{t=1}^{T} \mathcal{L}_{\overline{r}, \overline{G}, \theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) - \sum_{t=1}^{T} \mathcal{L}_{\overline{r}, \overline{G}, \theta}(\pi^*, \lambda_t),$$

which in turn allows to bound sublinearly $\frac{1}{\rho} \cdot \sum_{t=1}^{T} \sum_{i \in [m]} \left[V^{\pi_t}(\overline{g}_i) - \theta_i \right]^+$, once excluded the sublinear terms associated to the confidence bounds. Finally, simplifying in the PO-DB regret bound the ${}^1/_\rho$ factor associated to the violation term gives the desired result.

Proof. We first notice the following equations:

$$\begin{aligned}
&\sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right) \\
&\quad = \frac{L}{L+1} \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right) + \frac{1}{L+1} \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right) \\
&\quad = \frac{L}{L+1} \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right) + \frac{1}{\rho} \sum_{i \in [m]} \left[V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right]^+, \quad (3.8)
\end{aligned}$$

where Equation (3.8) holds by definition λ_t.
Employing similar steps to Theorem 3.1, it holds, with probability at least $1 - \mathcal{O}(\delta)$:

$$\begin{aligned}
&\sum_{t=1}^{T} \left(V^{\pi^*}(\overline{r}) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi^*}(\overline{g}_i) - \theta_i \right) \right) \\
&\qquad - \sum_{t=1}^{T} \left(V^{\pi_t}(\overline{r}) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta_i \right) \right) \\
&\qquad \leq \widetilde{\mathcal{O}} \left(\frac{L^3 m}{\rho} |X| \sqrt{|A|T} + \frac{L^5 m}{\rho} \right).
\end{aligned}$$

Adding and subtracting the estimated violation the following inequality holds:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathcal{L}_{\overline{r}, \overline{G}, \theta}(\pi^*, \lambda_t) - \sum_{t=1}^{T} \left(V^{\pi_t}(\overline{r}) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t}(\widetilde{g}_{t,i}) - \theta \right) \right. \\
&\qquad \left. \pm \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right) \right) \\
&\qquad \leq \widetilde{\mathcal{O}} \left(\frac{L^3 m}{\rho} |X| \sqrt{|A|T} + \frac{L^5 m}{\rho} \right),
\end{aligned}$$

which by Hölder inequality and Lemma A.9 implies:

$$\sum_{t=1}^{T} \mathcal{L}_{\overline{r}, \overline{G}, \theta}(\pi^*, \lambda_t) - \sum_{t=1}^{T} \left(V^{\pi_t}(\overline{r}) - \sum_{i \in [m]} \lambda_{t,i} \left(V^{\pi_t, \widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i \right) \right)$$

$$\leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right). \tag{3.9}$$

Thus we substitute Equation (3.8) in Inequality (3.9) to obtain,

$$\begin{aligned}\sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) - \sum_{t=1}^{T}\Bigg(V^{\pi_t}(\overline{r}) - \frac{L}{L+1}\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) \\ - \frac{1}{\rho}\sum_{i\in[m]}\left[V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right]^+\Bigg) \\ \leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right).\end{aligned}$$

To get back to the Lagrangian function of the offline problem, we notice that:

$$\begin{aligned}&\frac{L}{L+1}\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right) \\ &\geq \frac{L}{L+1}\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t,\widehat{P}_t}(\overline{g}_i - 2\xi_t) - \theta_i\right) \\ &\geq \frac{L}{L+1}\sum_{i\in[m]}\lambda_{t,i}\left(V^{\pi_t}(\overline{g}_i) - \theta_i\right) - \frac{2Lm}{\rho}V^{\pi_t}(\xi_t) - \frac{2Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1,\end{aligned}$$

and we employ Lemma A.6 Lemma A.9 to obtain:

$$\begin{aligned}\sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) - \sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) + \frac{1}{\rho}\sum_{t=1}^{T}\sum_{i\in[m]}\left[V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right]^+ \\ \leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right).\end{aligned}$$

Thus, similarly to the steps above, we notice that:

$$\begin{aligned}&\frac{1}{\rho}\sum_{i\in[m]}\left[V^{\pi_t,\widehat{P}_t}(\widetilde{g}_{t,i}) - \theta_i\right]^+ \\ &\geq \frac{1}{\rho}\sum_{i\in[m]}\left[V^{\pi_t,\widehat{P}_t}(\overline{g}_i - 2\xi_t) - \theta_i\right]^+ \\ &\geq \frac{1}{\rho}\sum_{i\in[m]}\left[V^{\pi_t}(\overline{g}_i) - \theta_i\right]^+ - \frac{2m}{\rho}V^{\pi_t}(\xi_t) - \frac{2m}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1.\end{aligned}$$

Employing Lemma A.6 and Lemma A.9 we obtain:

$$\sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) - \sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) + \frac{1}{\rho}\sum_{t=1}^{T}\sum_{i\in[m]}\left[V^{\pi_t}(\overline{g}_i) - \theta_i\right]^+$$

$$\leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right),$$

which can be rewritten as:

$$\frac{1}{\rho}\sum_{t=1}^{T}\sum_{i\in[m]}[V^{\pi_t}(\overline{g}_i) - \theta_i]^+$$
$$\leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right) + \sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) - \sum_{t=1}^{T}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t).$$

Thus, we employ Lemma A.4 with Lemma A.6 and Lemma A.9 to obtain:

$$\frac{1}{\rho}\sum_{t=1}^{T}\sum_{i\in[m]}[V^{\pi_t}(\overline{g}_i) - \theta_i]^+ \leq \widetilde{\mathcal{O}}\left(\frac{L^3 m}{\rho}|X|\sqrt{|A|T} + \frac{L^5 m}{\rho}\right),$$

from which we obtain the final bound:

$$\begin{aligned}
\mathcal{V}_T &:= \max_{i\in[m]}\sum_{t=1}^{T}[V^{\pi_t}(\overline{g}_i) - \theta_i]^+ \\
&\leq \sum_{t=1}^{T}\sum_{i\in[m]}[V^{\pi_t}(\overline{g}_i) - \theta_i]^+ \\
&= \rho\cdot\frac{1}{\rho}\sum_{t=1}^{T}\sum_{i\in[m]}[V^{\pi_t}(\overline{g}_i) - \theta_i]^+ \\
&\leq \widetilde{\mathcal{O}}\left(L^3 m|X|\sqrt{|A|T} + L^5 m\right),
\end{aligned}$$

with probability at least $1 - \mathcal{O}(\delta)$, by Union Bound. This concludes the proof. □

Comparing our violations result with the one in Müller et al. (2024), Algorithm 3.1 attains the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ order in T, while Müller et al. (2024) achieves a suboptimal $T^{0.93}$. For the violations, our bound is not worse than the one in (Müller et al., 2024) w.r.t. the dependency on m and L. Moreover, we do not have the dependency on the Slater's parameter ρ.

Part II

Adversarial Rewards and Stochastic Constraints

CHAPTER 4

Regret Minimization Under Hard Constraints

In this chapter, we study online learning problems in *episodic* CMDPs with *adversarial losses* and *stochastic hard constraints*, under *bandit* feedback. In such settings, the goal of the learner is to minimize their *regret*—the difference between their cumulative loss and what they would have obtained by always selecting a best-in-hindsight policy—, while at the same time guaranteeing that the constraints are satisfied during the learning process.

We consider three scenarios that differ in the way in which constraints are satisfied and are all usually referred to as *hard* constraints settings in the literature (Liu et al., 2021; Guo et al., 2022). In the first scenario, the learner attains *sublinear* cumulative *strong* constraints violation. In the second one, the learner satisfies constraints at every episode, while, in the third one, they achieve *constant* cumulative *strong* constraints violation.

To the best of our knowledge, this work is the first to study CMDPs that involve both adversarial losses and hard constraints. Indeed, all the works on adversarial CMDPs (see, *e.g.*, (Wei et al., 2018; Qiu et al., 2020)) consider settings with *soft* constraints. These are much weaker than hard constraints, as they are only concerned with the minimization of the cumulative constraints violation. As a result, they allow negative violations to cancel out positive ones across different episodes. Such cancellations are unreasonable in real-world applications. For instance, in autonomous driving, avoiding a collision clearly does *not* "repair" a crash occurred previously. Furthermore, the only few works addressing stochastic hard constraints in CMDPs (Liu et al., 2021; Shi et al., 2023; Müller et al., 2024; Stradi et al., 2025a) are restricted to *stochastic losses*. Thus, their techniques cannot be easily generalized to our setting. Our CMDP settings capture many more applications than theirs, since being able to deal with adversarial losses allows to tackle general non-stationary environments, which are ubiquitous in the real world.

We start by addressing the first scenario, where we design an algorithm—called Sublin-

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

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

Then, in the third scenario, we design an algorithm that attains *constant* cumulative strong constraints violation and sublinear regret, by simply assuming that a policy strictly satisfying the constraints exists, but it is *not* known to learner. Our algorithm—called Constant Violation Optimistic Policy Search (CV-OPS)— estimates such a policy and its associated constraints violation in a constant number of episodes. This is done by employing two no-regret algorithms. The first one with the objective of minimizing violation, and the second one with the goal of selecting the most violated constraint. A stopping condition that depends on the guarantees of both the no-regret algorithms enforces that the number of episodes used to estimate the desired policy is sufficient, while still being constant. After that, CV-OPS runs S-OPS with the estimated policy, attaining the desired results. Finally, we provide a lower bound showing that any algorithm attaining $o(\sqrt{T})$ violation cannot avoid a dependence on the Slater's parameter in the regret bound. We believe that this result may be of independent interest, since it is not only applicable to our second and third settings, but also to other settings where a larger violation is allowed.

4.1 Setting and Additional Notation

In this chapter, we study CMDPs with *adversarial losses* and *stochastic constraints*, under *bandit feedback*. Specifically, $\{\ell_t\}_{t=1}^T$ is the sequence of vectors of losses at each episode,

namely $\ell_t \in [0,1]^{|X \times A|}$.[1] We refer to the loss for a state-action pair $(x,a) \in X \times A$ as $\ell_t(x,a)$. Losses are adversarial, *i.e.*, no statistical assumption on how they are selected is made. Differently, costs are stochastic, *i.e.*, the matrices G_t are i.i.d. random variables distributed according to an (unknown) probability distribution $\mathcal{G}$. Thus, the performance of the learner is evaluated in terms of the *cumulative regret* defined as:

$$R_T := \sum_{t=1}^{T} \ell_t^\top q^{P,\pi_t} - T \cdot \text{OPT}_{\overline{\ell},\overline{G},\theta},$$

where $\overline{\ell} := \frac{1}{T}\sum_{t=1}^{T} \ell_t$ is the average of the adversarial losses over the T episodes, $\overline{G} := \mathbb{E}_{G \sim \mathcal{G}}[G]$ is the expected value of the stochastic cost matrices and $\text{OPT}_{\overline{\ell},\overline{G},\theta}$ is the equivalent of Program (2.3) for losses and it is computed as:

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

4.1.1 Guaranteeing Sublinear Violation

In this setting, we consider the *cumulative strong constraints violation*:

$$\mathcal{V}_T := \max_{i \in [m]} \sum_{t=1}^{T} \left[\overline{G}^\top q_t - \theta\right]_i^+,$$

where $[x]^+ := \max\{0, x\}$. Our goal is to design algorithms with $\mathcal{V}_T = o(T)$.

4.1.2 Guaranteeing Safety

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

Definition 4.1 (Safe algorithm)**.** *An algorithm is* safe *if and only if* $\overline{G}^\top q_t \leq \theta, \ \forall t \in [T]$.

As shown by Bernasconi et al. (2022), without further assumptions, it is *not* possible to achieve $R_T = o(T)$ while at the same time guaranteeing that the safety property holds with high probability, even in simple stochastic multi-armed bandit instances. To design safe learning algorithms, we need the following two assumptions. The first one is the Slater's condition, that is, Condition 2.1 is satisfied. The second assumption is related to learner's knowledge about a strictly feasible policy.

Assumption 4.1. *Both the strictly safe policy* $\pi^\diamond$ *and its costs* $\beta = [\beta_1, \dots, \beta_m] := \overline{G}^\top q^\diamond$ *are known to the learner.*

Intuitively, Assumption 4.1 is needed to guarantee that safety holds during the first episodes, when the leaner's uncertainty about the costs is high. Conditions 2.1 and 4.1 are often employed in CMDPs (see, *e.g.*, (Liu et al., 2021)), as they are usually met in real-world applications of interest, where it is common to have access to a "do-nothing" policy resulting in *no* constraint being violated.

[1] As pointed out in Chapter 2, the notion of loss is equivalent to the one of reward by taking $\ell_t(x,a) = 1 - r_t(x,a)$.

4.1.3 Guaranteeing Constant Violation

In this setting, we relax Assumption 4.1, and we only assume Slater's condition (Condition 2.1). We show that it is possible to achieve constant violation, namely $\mathcal{V}_T$ is upper bounded by a constant independent of T, while attaining similar regret guarantees compared to the second setting.

4.2 Concentration Bounds

In this section, we provide concentration bounds for the estimates of unknown stochastic parameter of the CMDP. Let $N_t(x,a)$ be the number of episodes up to $t \in [T]$ in which the state-action pair $(x,a) \in X \times A$ is *visited*. Then, $\widehat{g}_{t,i}(x,a) := \frac{\sum_{\tau \in [t]} g_{\tau,i}(x,a) \mathbb{I}_\tau(x,a)}{\max\{1, N_t(x,a)\}}$, with $\mathbb{I}_\tau(x,a) = 1$ if and only if (x,a) is visited in episode τ, is an unbiased estimator of the expected cost of constraint $i \in [m]$ for (x,a), which we denote by $\overline{g}_i(x,a) := \mathbb{E}_{G \sim \mathcal{G}}[g_{t,i}(x,a)]$.

Thus, by applying Hoeffding's inequality, it holds, with probability at least $1-\delta$ that $|\widehat{g}_{t,i}(x,a) - \overline{g}_i(x,a)| \leq \xi_t(x,a)$, where we let the confidence bound $\xi_t(x,a) := \min\{1, \sqrt{4 \ln(T|X||A|m/\delta) / \max\{1, N_t(x,a)\}}\}$ (refer to Lemma 3.2 for the formal result). For ease of notation, we let $\widehat{G}_t \in [0,1]^{|X \times A| \times m}$ be the matrix of the estimated costs $\widehat{g}_{t,i}(x,a)$. Moreover, we denote by $\xi_t \in [0,1]^{|X \times A|}$ the vector whose entries are the bounds $\xi_t(x,a)$, and we let $\Xi_t \in [0,1]^{|X \times A| \times m}$ be a matrix built by concatenating vectors ξ_t in such a way that the statement of Lemma 3.2 becomes: $|\widehat{G}_t - \overline{G}| \preceq \Xi_t$ holds with probability at least $1-\delta$, where $|\cdot|$ and $\preceq$ are applied component wise. In the following, given any $\delta \in (0,1)$, we refer to the event defined in Lemma 3.2 as $\mathcal{E}^G(\delta)$.

Similarly, we define *confidence sets* for the transition function of a CMDP, by exploiting suitable concentration bounds for estimated transition probabilities. By letting $M_t(x,a,x')$ be the total number of episodes up to $t \in [T]$ in which $(x,a) \in X \times A$ is visited and the environment transitions to $x' \in X$, the estimated transition probability for (x,a,x') is defined as $\widehat{P}_t(x' \mid x,a) := \frac{M_t(x,a,x')}{\max\{1, N_t(x,a)\}}$. Then, at episode $t \in [T]$, the confidence set for the transitions is $\mathcal{P}_t := \bigcap_{(x,a,x') \in X \times A \times X} \mathcal{P}_t^{x,a,x'}$, with $\mathcal{P}_t^{x,a,x'} := \left\{ \overline{P} : |\overline{P}(x'|x,a) - \widehat{P}_t(x'|x,a)| \leq \epsilon_t(x,a,x') \right\}$, where we let $\epsilon_t(x,a,x') := 2\sqrt{\frac{\widehat{P}_t(x'|x,a) \ln(T|X||A|/\delta)}{\max\{1, N_t(x,a)-1\}}} + \frac{14 \ln(T|X||A|/\delta)}{3 \max\{1, N_t(x,a)-1\}}$ for some confidence $\delta \in (0,1)$. It is well known that, with probability at least $1-4\delta$, it holds that the transition function P belongs to $\mathcal{P}_t$ for all $t \in [T]$ (see (Jin et al., 2020a) and Lemma A.8 for the formal statement). At each $t \in [T]$, given a confidence set $\mathcal{P}_t$, it is possible to efficiently build a set $\Delta(\mathcal{P}_t)$ that comprises all the occupancy measures that are valid with respect to every transition function $\overline{P} \in \mathcal{P}_t$. We defer the formal definition of $\Delta(\mathcal{P}_t)$ to Appendix B.2. Lemma A.8 implies that, with high probability, the set $\Delta(M)$ of valid occupancy measures is included in all the "estimated" sets $\Delta(\mathcal{P}_t)$, for every $t \in [T]$. In the following, given any confidence $\delta \in (0,1)$, we refer to the event that $\Delta(M) \subseteq \bigcap_{t \in [T]} \Delta(\mathcal{P}_t)$ as $\mathcal{E}^\Delta(\delta)$, which holds with probability at least $1-4\delta$ thanks to Lemma A.8.

Finally, for ease of presentation, given $\delta \in (0,1)$ we define a *clean event* $\mathcal{E}^{G,\Delta}(\delta)$ under which all the concentration bounds for costs and transitions correctly hold. Formally, $\mathcal{E}^{G,\Delta}(\delta) := \mathcal{E}^G(\delta) \cap \mathcal{E}^\Delta(\delta)$, which holds with probability at least $1-5\delta$ by a union bound (and Lemmas 3.2 and A.8).

Algorithm 4.1 Sublinear Violation Optimistic Policy Search (SV-OPS)

Require: X, A, θ, T, δ, η, γ
for $k \in [0, \dots, L-1]$, $(x, a, x') \in X_k \times A \times X_{k+1}$ **do**
  $N_0(x,a) \leftarrow 0$; $M_0(x,a,x') \leftarrow 0$
  $\widehat{q}_1(x,a,x') \leftarrow 1/|X_k||A||X_{k+1}|$
end for
$\pi_1 \leftarrow \pi^{\widehat{q}_1}$
for $t \in [T]$ **do**
  Choose π_t and receives *bandit feedback* as depicted in Algorithm 2.1
  Build *upper occupancy bounds* for $k \in [0, \dots, L-1]$:

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

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

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

  for $k \in [0, \dots, L-1]$ **do**
    $N_t(x_k, a_k) \leftarrow N_{t-1}(x_k, a_k) + 1$
    $M_t(x_k, a_k, x_{k+1}) \leftarrow M_{t-1}(x_k, a_k, x_{k+1}) + 1$
  end for
  Build $\mathcal{P}_t$, $\widehat{G}_t$, and Ξ_t as in Section 4.2
  Build *unconstrained occupancy* for all (x, a, x'):

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

  if PROJ$\left(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\right)$ is *feasible* **then**
    $\widehat{q}_{t+1} \leftarrow$ PROJ$\left(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\right)$
  else
    $\widehat{q}_{t+1} \leftarrow$ any $q \in \Delta(\mathcal{P}_t)$
  end if
  $\pi_{t+1} \leftarrow \pi^{\widehat{q}_{t+1}}$
end for

4.3 Guaranteeing Sublinear Violation

We start by designing the SV-OPS algorithm, guaranteeing that both the regret R_T and the *strong* constraints violation $\mathcal{V}_T$ are sublinear in T. Dealing with adversarial losses while limiting strong constraints violation begets considerable challenges, which go beyond classical exploration-exploitation trade-offs faced in unconstrained settings. On the one hand, using state-of-the-art algorithms for online learning in adversarial, unconstrained MDPs would lead to sublinear regret, but violation would grow linearly. On the other hand, a naïve approach that randomly explores to compute a set of policies satisfying the constraints with high probability can lead to sublinear violation, at the cost of linear regret. Thus, a clever adaptation of the techniques for unconstrained settings is needed.

Our algorithm—called Sublinear Violation Optimistic Policy Search (SV-OPS)—works by selecting policies derived from a set of occupancy measures that *optimistically* satisfy cost constraints.

This ensures that the set is always non-empty with high probability and that it collapses to the (true) set of constraint-satisfying occupancy measures as the number of episodes increases, enabling SV-OPS to attain sublinear constraints violation. The fundamental property preserved by SV-OPS is that, even though the "optimistic" set changes during the execution of the algorithm, it always subsumes the (true) set of constraint-satisfying occupancy measures. This crucially allows SV-OPS to employ classical policy-selection methods for unconstrained MDPs.

Algorithm 4.1 provides the pseudocode of SV-OPS. At each episode $t \in [T]$, SV-OPS plays policy π_t and receives feedback as described in Algorithm 2.1 (Line 7). Then, SV-OPS computes an *upper occupancy bound* $u_t(x_k, a_k)$ for every state-action pair (x_k, a_k) visited during Algorithm 2.1, by using the confidence set for the transition function $\mathcal{P}_{t-1}$ computed in the previous episode, namely, it sets $u_t(x_k, a_k) := \max_{\overline{P} \in \mathcal{P}_{t-1}} q^{\overline{P}, \pi_t}(x, a)$ for every $k \in [0 \dots L-1]$ (Line 8). Intuitively, $u_t(x_k, a_k)$ represents the maximum probability with which (x_k, a_k) is visited when using policy π_t, given the confidence set for the transition function built so far. The upper occupancy bounds are combined with the exploration factor γ to compute an *optimistic loss estimator* $\widehat{\ell}_t(x, a)$ for every state-action pair $(x, a) \in X \times A$ (see Line 9). After that, SV-OPS updates all the counters given the path traversed in Algorithm 2.1 (Lines 11–12), it builds the new confidence set $\mathcal{P}_t$, and it computes the matrices $\widehat{G}_t$ and Ξ_t of estimated costs and their bounds, respectively, by using the feedback (Line 14). To choose a policy π_{t+1}, SV-OPS first computes an *unconstrained occupancy measure* $\tilde{q}_{t+1}$ according to an unconstrained OMD update (Orabona, 2019) (see Line 15). Then, $\tilde{q}_{t+1}$ is projected onto a suitably-defined set of occupancy measures that *optimistically* satisfy the constraints. Next, we formally define the projection (Line 16).

$$\text{PROJ}\big(\tilde{q}_{t+1}, \widehat{G}_t, \Xi_t, \mathcal{P}_t\big) := \begin{cases} \underset{q \in \Delta(\mathcal{P}_t)}{\arg\min} \quad D(q || \tilde{q}_{t+1}) \\ \text{s.t.} \quad \big(\widehat{G}_t - \Xi_t\big)^\top q \leq \theta, \end{cases} \tag{4.1}$$

where $D(q||\tilde{q}_{t+1})$ is the unnormalized KL-divergence between q and $\tilde{q}_{t+1}$. Problem (4.1) is a linearly-constrained convex mathematical program, and, thus, it can be solved efficiently, that is, in polynomial time, for an arbitrarily-good approximate solution.[2] Intuitively, Problem (4.1) performs a projection onto the set of $q \in \Delta(\mathcal{P}_t)$ that additionally satisfy $\big(\widehat{G}_t - \Xi_t\big)^\top q \leq \theta$, where lower confidence bounds $\widehat{G}_t - \Xi_t$ for the costs are used in order to take an optimistic approach with respect to constraints satisfaction. Finally, if Problem (4.1) is feasible, then at the next episode SV-OPS selects the $\pi^{\widehat{q}_{t+1}}$ induced by a solution $\widehat{q}_{t+1}$ to Problem (4.1) (Line 17), otherwise it chooses a policy induced by any $q \in \Delta(\mathcal{P}_t)$ (Line 19).

The optimistic approach adopted in Problem (4.1) crucially allows to prove the following lemma.

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

Proof. To prove the lemma we show that under the event $\mathcal{E}^{G,\Delta}(\delta)$, which holds the probability at least $1 - 5\delta$, Program (4.1) admits a feasible solution. Precisely, under

[2] As customary in adversarial MDPs, we assume that an optimal solution to Problem (4.1) can be computed efficiently. Otherwise, we can still derive all of our results up to small approximations.

the event $\mathcal{E}^{\Delta}(\delta)$, the true transition function P belongs to $\mathcal{P}_t$ at each episode. Moreover, under the event $\mathcal{E}^{G}(\delta)$, we have, for any feasible solution $q^{\square}$ of the offline optimization problem, for any $t \in [T]$,

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

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

Lemma 4.1 holds since, under the event $\mathcal{E}^{G,\Delta}(\delta)$, projection is performed on a set subsuming the (true) set of constraints-satisfying occupancies. Lemma 4.1 is fundamental, as it allows to show that SV-OPS attains sublinear $\mathcal{V}_T$ and R_T.

4.3.1 Cumulative Strong Constraints Violation

To prove that the strong constraints violation achieved by SV-OPS is sublinear, we exploit the fact that the concentration bounds for costs and transitions shrink at a rate of $\mathcal{O}(1/\sqrt{T})$. This allows us to show the following result.

Theorem 4.1. *Given $\delta \in (0,1)$, Algorithm 4.1 attains:*

$$\mathcal{V}_T \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(T|X||A|m/\delta\right)}\right),$$

with prob. at least $1-8\delta$.

Proof. The key point of the problem is to relate the constraints satisfaction with the convergence rate of both the confidence bound on the constraints and the transitions. First, we notice that under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, all the following reasoning hold for every constraint $i \in [m]$. Thus, we focus on the bound of a single constraint violation problem defined as follows:

$$\mathcal{V}_T := \sum_{t=1}^{T} \left[\overline{g}^{\top} q_t - \theta\right]^{+}$$

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

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

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

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

Rearranging the equation, it holds:

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

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

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

from which we obtain:

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

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

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

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

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

$$\begin{aligned}\sum_{t=1}^T \|\widehat{q}_t - q_t\|_1 &= \sum_{t=1}^T \sum_{x,a} |\widehat{q}_t(x,a) - q_t(x,a)| \\ &\leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \qquad (4.2)\end{aligned}$$

where Inequality (4.2) holds since, by Lemma A.9, under the clean event, with probability at least $1-2\delta$, we have $\sum_{t=1}^T \sum_{x,a} |\widehat{q}_t(x,a) - q_t(x,a))| \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right)$, when $\widehat{q}_t \in \Delta(\mathcal{P}_t)$. Please notice that the condition $\widehat{q}_t \in \Delta(\mathcal{P}_t)$ is verified since the constrained space defined by Program (4.1) is contained in $\Delta(\mathcal{P}_t)$.

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

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

$$\leq \sum_{t=1}^{T} \xi_{t-1}^{\top} q_t + \sum_{t=1}^{T} \|\widehat{q}_t - q_t\|_1.$$

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

$$\begin{aligned}
&\sum_{t=1}^{T} \xi_{t-1}^{\top} q_t \\
&= \sum_{t=1}^{T} \sum_{x,a} \xi_{t-1}(x,a) q_t(x,a) \\
&\leq \sum_{t=1}^{T} \sum_{x,a} \xi_{t-1}(x,a) \mathbb{I}_t(x,a) + L\sqrt{2T \ln \frac{1}{\delta}} \qquad (4.3) \\
&= \sqrt{4 \ln \left(\frac{T|X||A|m}{\delta} \right)} \sum_{t=1}^{T} \sum_{x,a} \sqrt{\frac{1}{\max\{1, N_{t-1}(x,a)\}}} \mathbb{I}_t(x,a) + L\sqrt{2T \ln \frac{1}{\delta}} \\
&\leq 3\sqrt{4 \ln \left(\frac{T|X||A|m}{\delta} \right)} \sum_{x,a} \sqrt{N_T(x,a)} + L\sqrt{2T \ln \frac{1}{\delta}} \qquad (4.4) \\
&\leq 6\sqrt{L|X||A|T \ln \left(\frac{T|X||A|m}{\delta} \right)} + L\sqrt{2T \ln \frac{1}{\delta}}, \qquad (4.5)
\end{aligned}$$

where Inequality (4.3) follows from Azuma inequality and noticing that $\sum_{x,a} \xi_{t-1}(x,a) q_t(x,a) \leq L$ (with probability at least $1-\delta$), Inequality (4.4) holds since $1 + \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} + 1 \leq 3\sqrt{T}$ and Inequality (4.5) follows from Cauchy-Schwarz inequality and noticing that $\sqrt{\sum_{x,a} N_T(x,a)} \leq \sqrt{LT}$.
We combine the previous bounds as follows:

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

The results holds with probability at least at least $1 - 8\delta$ by union bound over the clean event, Lemma A.9 and the Azuma-Hoeffding inequality. This concludes the proof. □

4.3.2 Cumulative Regret

The crucial observation that allows us to prove that the regret attained by SV-OPS grows sublinearly is that the set on which the algorithm perform its projection step (Problem (4.1)) always contains the (true) set of occupancy measures that satisfy the constraints, and, thus, it also always contains the best-in-hindsight constraint-satisfying occupancy measure q^*. As a result, even though cost estimates may be arbitrarily bad, SV-OPS is still guaranteed

to select policies resulting in losses that are smaller than or equal to those incurred by q^*. This allows us to show the following:

Theorem 4.2. *Given $\delta \in (0,1)$, by setting $\eta = \gamma = \sqrt{L \ln(L|X||A|/\delta)/T|X||A|}$, Algorithm 4.1 attains:*

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

with prob. at least $1 - 10\delta$.

Proof. We first rewrite the regret definition as follows:

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

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

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

$$\begin{aligned} ① &= \sum_{t=1}^{T} \ell_t^\top (q_t - \widehat{q}_t) \\ &= \sum_{t=1}^{T} \sum_{x,a} \ell_t(x,a)(q_t(x,a) - \widehat{q}_t(x,a)) \\ &\leq \sum_{t=1}^{T} \sum_{x,a} |(q_t(x,a) - \widehat{q}_t(x,a)|, \end{aligned} \tag{4.6}$$

where the Inequality (4.6) holds by Hölder inequality noticing that $\|\ell_t\|_\infty \leq 1$ for all $t \in [T]$. Then, noticing that the projection of Algorithm 4.1 is performed over a subset of $\Delta(\mathcal{P}_t)$ and employing Lemma A.9, we obtain:

$$① \leq \mathcal{O}\left(L|X|\sqrt{|A|T \ln\left(\frac{T|X||A|}{\delta}\right)}\right), \tag{4.7}$$

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

Bound on ②. To bound the second term, we underline that, under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, the estimated safe occupancy $\widehat{q}_t$ belongs to $\Delta(\mathcal{P}_t)$ and the optimal safe solution q^* is included in the constrained decision space for each $t \in [T]$. Moreover we

notice that, for each $t \in [T]$, the constrained space is convex and linear, by construction of Program (4.1). Thus, following the standard analysis of online mirror descent Orabona (2019) and from Lemma B.4, we have, under the clean event:

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

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

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

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

$$② \leq \mathcal{O}\left(L\sqrt{|X||A|T \ln\left(\frac{|X||A|}{\delta}\right)}\right), \tag{4.8}$$

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

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

$$③ \leq \mathcal{O}\left(L|X|\sqrt{|A|T \ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{4.9}$$

Bound on ④. We bound the fourth term employing Corollary B.1 and obtaining,

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

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

$$④ \leq \mathcal{O}\left(L\sqrt{|X||A|T \ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{4.10}$$

Final result. Finally, combining Equation (4.7), Equation (4.8), Equation (4.9) and Equation (4.10) and applying a union bound, we obtain, with probability at least $1-10\delta$,

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

This concludes the proof. □

4.4 Guaranteeing Safety

We design another algorithm, called S-OPS, attaining sublinear regret and enjoying the safety property with high probability. To do this, we work under Conditions 2.1 and 4.1. Designing safe algorithms raises many additional challenges compared to the case studied in Section 4.3. Indeed, adapting techniques for adversarial, unconstrained MDPs does *not* work anymore, and, thus, *ad hoc* approaches are needed. This is because safety extremely limits exploration.

Our algorithm—Safe Optimistic Policy Search (S-OPS)—builds on top of the SV-OPS algorithm developed in Section 4.3. Selecting policies derived from the "optimistic" set of occupancy measures, as done by SV-OPS, is *not* sufficient anymore, as it would clearly result in the safety property being unsatisfied during the first episodes. Our new algorithm circumvents such an issue by employing, at each episode, a suitable randomization between the policy derived from the "optimistic" set (the one SV-OPS would select) and the strictly feasible policy $\pi^\diamond$. Crucially, as we show next, such a randomization accounts for constraints satisfaction by taking a *pessimistic* approach, namely, by considering upper confidence bounds on the costs characterizing the constraints. This is needed in order to guarantee the safety property. Moreover, having access to the strictly feasible policy $\pi^\diamond$ and its expected costs β (Assumption 4.1) allows S-OPS to always place a sufficiently large probability on the policy derived from the "optimistic" set, so that a sufficient amount of exploration is guaranteed, and, in its turn, sublinear regret is attained. Notice that S-OPS effectively selects *non-Markovian* policies, as it employs a randomization between two Markovian policies at each episode.

Algorithm 4.2 provides the pseudocode of S-OPS. Differently from SV-OPS, the policy selected at the first episode is obtained by randomizing a uniform occupancy measure with $\pi^\diamond$ (Line 5). The probability λ_0 of selecting $\pi^\diamond$ is chosen pessimistically. Intuitively, in the first episode, being pessimistic means that λ_0 must guarantee that the constraints are satisfied for any possible choice of costs and transitions, and, thus, $\lambda_0 := \max_{i\in[m]}\{{}^{L-\theta_i}/_{L-\beta_i}\}$. Thanks to Conditions 2.1 and 4.1, it is always the case that $\lambda_0 < 1$. Thus, $\pi_1 \neq \pi^\diamond$ with positive probability and some exploration is performed even in the first episode. Analogously to SV-OPS, at each $t \in [T]$, S-OPS selects a policy π_t and receives feedback as described in Algorithm 2.1, it computes optimistic loss estimators, it updates the confidence set for the transitions, and it computes the matrices of estimated costs and their bounds. Then, as in SV-OPS, an update step of unconstrained OMD is performed. Although identical to the update done in SV-OPS, the one in S-OPS uses loss estimators computed when using a randomization between the policy obtained by solving Problem (4.1) and the strictly feasible policy $\pi^\diamond$. Thus, there is a mismatch between the occupancy measure used to estimate losses and the one computed by the pro-

Algorithm 4.2 Safe Optimistic Policy Search (`S-OPS`)

Require: X, A, θ, T, δ, η, γ, $\pi^\diamond$, β

1: **for** $k \in [0, \dots, L-1]$, $(x, a, x') \in X_k \times A \times X_{k+1}$ **do**
2: $N_0(x,a) \leftarrow 0$; $M_0(x,a,x') \leftarrow 0$
3: $\widehat{q}_1(x,a,x') \leftarrow \frac{1}{|X_k||A||X_{k+1}|}$
4: **end for**
5: $\pi_1 \leftarrow \begin{cases} \pi^\diamond & \text{w. probability } \lambda_0 := \max_{i\in[m]} \left\{ \frac{L-\theta_i}{L-\beta_i} \right\} \\ \pi^{\widehat{q}_1} & \text{w. probability } 1-\lambda_0 \end{cases}$
6: **for** $t \in [T]$ **do**
7: Select π_t in Algorithm 2.1 and receive feedback
8: Build *upper occupancy bounds* for $k \in [0, \dots, L-1]$:

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

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

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

10: **for** $k \in [0, \dots, L-1]$ **do**
11: $N_t(x_k, a_k) \leftarrow N_{t-1}(x_k, a_k) + 1$
12: $M_t(x_k, a_k, x_{k+1}) \leftarrow M_{t-1}(x_k, a_k, x_{k+1}) + 1$
13: **end for**
14: Build $\mathcal{P}_t$, $\widehat{G}_t$, and Ξ_t as in Section 4.2
15: Build *unconstrained occupancy* for all (x, a, x'):

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

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

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

20: Define $\overline{m} := \{i \in [m] : (\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} > \theta_i\}$
21: $\sigma \leftarrow \max_{i \in \overline{m}} \left\{ \frac{\min\left\{(\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1}, L\right\} - \theta_i}{\min\left\{(\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1}, L\right\} - \beta_i} \right\}$
22: $\lambda_t \leftarrow \begin{cases} \sigma & \text{if } \exists i \in [m] : (\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} > \theta_i \\ 0 & \text{if } \forall i \in [m] : (\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} \le \theta_i \end{cases}$
23: **else**
24: $\widehat{q}_{t+1} \leftarrow$ take any $q \in \Delta(\mathcal{P}_t)$; $\lambda_t \leftarrow 1$
25: **end if**
26: $\pi_{t+1} \leftarrow \begin{cases} \pi^\diamond & \text{with probability } \lambda_t \\ \pi^{\widehat{q}_{t+1}} & \text{with probability } 1-\lambda_t \end{cases}$
27: **end for**

jection step. The projection step performed by `S-OPS` (Line 16) is the same as the one in

SV-OPS. Specifically, the algorithm projects the unconstrained occupancy measure $\tilde{q}_{t+1}$ onto an "optimistic" set by solving Problem (4.1), which, if the problem is feasible, results in occupancy measure $\widehat{q}_{t+1}$. However, differently from SV-OPS, when the problem is feasible, S-OPS does *not* select the policy $\pi^{\widehat{q}_{t+1}}$ derived from $\widehat{q}_{t+1}$, but it rather uses a randomization between such a policy and the strictly feasible policy $\pi^{\diamond}$ (Line 26). The probability λ_t of selecting $\pi^{\diamond}$ is chosen pessimistically with respect to constraints satisfaction, by using upper confidence bounds for the costs and upper occupancy bounds given policy $\pi^{\widehat{q}_{t+1}}$ (Lines 19 and 22). Such a pessimistic approach ensures that the constraints are satisfied with high probability, thus making the algorithm safe with high probability. If Problem (4.1) is *not* feasible, then any occupancy measure in $\Delta(\mathcal{P}_t)$ can be selected (Line 24).

4.4.1 Safety Property

We show that S-OPS is safe with high probability.

Theorem 4.3. *Given a confidence $\delta \in (0,1)$, Algorithm 4.2 is safe with probability at least $1-5\delta$.*

Intuitively, Theorem 4.3 follows from the way in which the randomization probability λ_t is defined. Indeed, λ_t relies on two crucial components:

(i) a pessimistic estimate of the costs for state-action pairs, namely, the upper confidence bounds $\widehat{g}_{t,i} + \xi_t$,

(ii) a pessimistic choice of transition probabilities, encoded by the upper occupancy bounds defined by the vector $\widehat{u}_t$.

Notice that the $\max_{i \in \overline{m}}$ operator allows to be conservative with respect to all the (non satisfied) constraints.

Proof. We show that, under event $\mathcal{E}^{G,\Delta}(\delta)$, the *non-Markovian* policy defined by the probability λ_t satisfies the constraints. Intuitively, the result follows from the construction of the convex combination parameter λ_t. Indeed, λ_t is built using a pessimist estimated of the constraints cost, namely, $\widehat{g}_{t,i} + \xi_t$. Moreover, the upper occupancy bound $\widehat{u}_t$ introduces pessimism in the choice of the transition function. Finally, the $\max_{i \in \overline{m}}$ operator allows to be conservative for all the non satisfied constraints.
We split the analysis in the two possible cases defined by λ_t, namely, $\lambda_t = 0$ and $\lambda_t \in (0,1)$. Please notice that $\lambda_t < 1$, by construction.

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

$$\begin{aligned} \theta_i &\geq (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{u}_t \\ &\geq (\widehat{g}_{t-1,i} + \xi_{t-1})^\top \widehat{q}_t && (4.11) \\ &= (\widehat{g}_{t-1,i} + \xi_{t-1})^\top q_t \\ &\geq \overline{g}_i^\top q_t, && (4.12) \end{aligned}$$

where Inequality (4.11) holds by definition of $\widehat{u}_t$ and Inequality (4.12) by the pessimistic definition of the constraints.

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

$$\lambda_{t-1}\overline{g}_i^\top q^\diamond + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t} = \lambda_{t-1}\beta_i + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t}. \tag{4.13}$$

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

First case. It holds that:

$$\begin{aligned}
\lambda_{t-1}\beta_i + (1-\lambda_{t-1})\overline{g}_i^\top q^{\widehat{\pi}_t,P} &\leq \lambda_{t-1}\beta_i + (1-\lambda_{t-1})L \qquad (4.14)\\
&= \frac{L-\theta_i}{L-\beta_i}(\beta_i - L) + L\\
&= \frac{\theta_i - L}{\beta_i - L}(\beta_i - L) + L\\
&= \theta_i,
\end{aligned}$$

where Inequality (4.14) holds by definition of the constraints.

Second case. It holds that:

$$\begin{aligned}
&\lambda_{t-1}\beta_i + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t}\\
&\quad\leq \lambda_{t-1}\beta_i + (1-\lambda_{t-1})\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top q^{P,\widehat{\pi}_t} \qquad (4.15)\\
&\quad\leq \lambda_{t-1}\beta_i + (1-\lambda_{t-1})\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t \qquad (4.16)\\
&\quad= \lambda_{t-1}\beta_i - \lambda_{t-1}\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t + \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t\\
&\quad= \lambda_{t-1}(\beta_i - \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t) + \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t\\
&\quad\leq \frac{\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t - \theta_i}{\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t - \beta_i}(\beta_i - \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t) + \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t\\
&\quad= \frac{\theta_i - \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t}{\beta_i - \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t}(\beta_i - \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t) + \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t\\
&\quad= \theta_i - \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t + \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t\\
&\quad= \theta_i,
\end{aligned}$$

where Inequality (4.15) holds by the definition of the event and Inequality (4.16) holds by the definition of $\widehat{u}_t$.

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

4.4.2 Cumulative Regret

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

Theorem 4.4. *Given $\delta \in (0,1)$, by setting $\eta = \gamma = \sqrt{L \ln(L|X||A|/\delta)/T|X||A|}$, Algorithm 4.2 attains:*

$$R_T \leq \mathcal{O}\left(\Psi L^3 |X| \sqrt{|A| T \ln\left(T|X||A|m/\delta\right)}\right),$$

with prob. at least $1 - 11\delta$, where $\Psi := \max_{i \in [m]} \{1/\min\{(\theta_i - \beta_i), (\theta_i - \beta_i)^2\}\}$.

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

$$\underbrace{\sum_{t=1}^{T} \ell_t^\top \left(q_t - q^{P_t,\pi_t}\right)}_{①} + \underbrace{\sum_{t=1}^{T} \widehat{\ell}_t^\top \left(q^{P_t,\widehat{\pi}_t} - q^*\right)}_{②} + \underbrace{\sum_{t=1}^{T} \ell_t^\top \left(q^{P_t,\pi_t} - q^{P_t,\widehat{\pi}_t}\right)}_{③} +$$
$$+ \underbrace{\sum_{t=1}^{T} \left(\ell_t - \widehat{\ell}_t\right)^\top q^{P_t,\widehat{\pi}_t}}_{④} + \underbrace{\sum_{t=1}^{T} \left(\widehat{\ell}_t - \ell_t\right)^\top q^*}_{⑤},$$

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

Bound on ① We bound the first term as follows:

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

where the last inequality holds by Hölder inequality noticing that $\|\ell_t\|_\infty \leq 1$ for all $t \in [T]$. Then we can employ Lemmas A.9, since π_t is the policy that guides the exploration and $P_t \in \mathcal{P}_t$, obtaining:

$$① \leq \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right), \tag{4.17}$$

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

Bound on ② The second term is bounded similarly to the second part of Theorem 4.2. Precisely, we notice that under the clean event $\mathcal{E}^{G,\Delta}(\delta)$, the optimal safe solution q^* is included in the constrained decision space for each $t \in [T]$. Moreover we notice that, for each $t \in [T]$, the constrained space is convex and linear, by construction of the convex program. Thus, following the standard analysis of online mirror descent Orabona (2019) and from Lemma B.4, we have, under the clean event:

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

Guaranteeing the safety property makes bounding the biased estimator more complex with respect to Theorem 4.2. Thus, noticing that $\lambda_{t-1} \leq \max_{i\in[m]}\left\{\frac{L-\theta_i}{L-\beta_i}\right\}$ and by definition of π_t, we proceed as follows:

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

The previous result is intuitive. Paying an additional $\max_{i\in[m]}\left\{\frac{L}{\theta_i-\beta_i}\right\}$ factor allows to relate the loss estimator $\widehat{\ell}_t$ with the policy that guides the exploration, namely, π_t. Thus, following the same steps as Theorem 4.2 we obtain, with probability $1-\delta$, under the clean event:

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

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

$$② \leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{1}{\theta_i-\beta_i}\right\}L\sqrt{L|X||A|T\ln\left(\frac{|X|^2|A|}{\delta}\right)}\right), \tag{4.18}$$

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

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

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

where we used that $\ell_t^\top q^{P_t,\pi^\diamond} \leq L$ for any $t \in [T]$. Thus, we proceed bounding $\sum_{t=1}^{T} \lambda_{t-1}$.
We focus on a single episode $t \in [T]$, in which we assume without loss of generality that the i-th constraint is the hardest to satisfy.
Precisely,

$$\begin{aligned}\lambda_t &= \frac{\min\left\{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1}, L\right\} - \theta_i}{\min\left\{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1}, L\right\} - \beta_i} \\ &\leq \frac{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} - \theta_i}{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} - \beta_i} \\ &\leq \frac{(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} - \theta_i}{\theta_i - \beta_i} \qquad (4.19) \\ &= \frac{(\widehat{g}_{t,i} - \xi_t)^\top \widehat{u}_{t+1} + 2\xi_t^\top \widehat{u}_{t+1} - \theta_i}{\theta_i - \beta_i} \\ &= \frac{(\widehat{g}_{t,i} - \xi_t)^\top \widehat{q}_{t+1} + (\widehat{g}_{t,i} - \xi_t)^\top (\widehat{u}_{t+1} - \widehat{q}_{t+1}) + 2\xi_t^\top \widehat{u}_{t+1} - \theta_i}{\theta_i - \beta_i} \\ &\leq \frac{(\widehat{g}_{t,i} - \xi_t)^\top \widehat{q}_{t+1} + \widehat{g}_{t,i}^\top (\widehat{u}_{t+1} - \widehat{q}_{t+1}) + 2\xi_t^\top \widehat{u}_{t+1} - \theta_i}{\theta_i - \beta_i} \\ &\leq \frac{\widehat{g}_{t,i}^\top (\widehat{u}_{t+1} - \widehat{q}_{t+1}) + 2\xi_t^\top \widehat{u}_{t+1}}{\theta_i - \beta_i} \qquad (4.20) \\ &= \frac{\widehat{g}_{t,i}^\top (\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}) + \widehat{g}_{t,i}^\top (q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}) + 2\xi_t^\top \widehat{u}_{t+1}}{\theta_i - \beta_i} \\ &\leq \frac{\|\widehat{g}_{t,i}\|_\infty ||\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}||_1 + \|\widehat{g}_{t,i}\|_\infty \|q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}\|_1 + 2\xi_t^\top \widehat{u}_{t+1}}{\theta_i - \beta_i} \\ &\leq \frac{||\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}||_1 + \|q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}\|_1 + 2\xi_t^\top \widehat{u}_{t+1}}{\theta_i - \beta_i} \\ &\leq \frac{L(1-\lambda_t)\|\widehat{u}_{t+1} - q^{P,\widehat{\pi}_{t+1}}\|_1 + L(1-\lambda_t)\|q^{P,\widehat{\pi}_{t+1}} - q^{P_{t+1},\widehat{\pi}_{t+1}}\|_1}{\min\left\{(\theta_i - \beta_i), (\theta_i - \beta_i)^2\right\}}\end{aligned}$$

$$+ \frac{2L(1-\lambda_t)\xi_t^\top \widehat{u}_{t+1}}{\min\{(\theta_i-\beta_i),(\theta_i-\beta_i)^2\}}, \tag{4.21}$$

where Inequality (4.19) holds since, for the hardest constraint, when $\lambda_t \neq 0$, $(\widehat{g}_{t,i}+\xi_t)^\top \widehat{u}_{t+1} > \theta_i$, Inequality (4.20) holds since, under the clean event, $(\widehat{g}_{t,i}-\xi_t)^\top \widehat{q}_{t+1} \leq \theta_i$ and Inequality (4.21) holds since $\lambda_t \leq \frac{L-\theta_i}{L-\beta_i}$. Intuitively, Inequality (4.21) shows that, to guarantee the safety property, Algorithm 4.2 has to pay a factor proportional to the pessimism introduced on the transition and cost functions, plus the constraints satisfaction gap of the strictly feasible solution given as input to the algorithm.
We need to generalize the result summing over t, taking into account that the hardest constraints may vary. Thus, we bound the summation as follows,

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

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

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

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

$$\begin{aligned}
&\sum_{t=1}^{T}(1-\lambda_{t-1})\xi_{t-1}^\top q^{P,\widehat{\pi}_t} \\
&\quad\leq \sum_{t=1}^{T}\xi_{t-1}^\top q_t && (4.22)\\
&\quad\leq \sum_{t=1}^{T}\sum_{x,a}\xi_{t-1}(x,a)\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} && (4.23)\\
&\quad= \sqrt{4\ln\left(\frac{T|X||A|m}{\delta}\right)}\sum_{t=1}^{T}\sum_{x,a}\sqrt{\frac{1}{\max\{1,N_{t-1}(x,a)\}}}\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} \\
&\quad\leq 6\sqrt{\ln\left(\frac{T|X||A|m}{\delta}\right)}\sqrt{|X||A|\sum_{x,a}N_T(x,a)} + L\sqrt{2T\ln\frac{1}{\delta}} && (4.24)\\
&\quad\leq 6\sqrt{L|X||A|T\ln\left(\frac{T|X||A|m}{\delta}\right)} + L\sqrt{2T\ln\frac{1}{\delta}},
\end{aligned}$$

where Inequality (4.22) follows from the definition of π_t, Inequality (4.23) follows from Azuma-Hoeffding inequality and Inequality (4.24) holds since $1 + \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} + 1 \leq 3\sqrt{T}$ and Cauchy-Schwarz inequality.
Thus, we obtain,

$$③ \leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{1}{\min\left\{(\theta_i-\beta_i),(\theta_i-\beta_i)^2\right\}}\right\} L^3|X|\sqrt{|A|T\ln\left(\frac{T|X||A|m}{\delta}\right)}\right), \tag{4.25}$$

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

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

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

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

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

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

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\ell_t-\mathbb{E}_t[\widehat{\ell}_t]\right)^\top q^{P_t,\widehat{\pi}_t}\\
&= \sum_{t,x,a} q^{P_t,\widehat{\pi}_t}(x,a)\ell_t(x,a)\left(1-\frac{\mathbb{E}_t\left[\mathbb{I}_t(x,a)\right]}{u_t(x,a)+\gamma}\right)\\
&= \sum_{t,x,a} q^{P_t,\widehat{\pi}_t}(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right)\\
&\leq \max_{i\in[m]}\left\{\frac{L}{\theta_i-\beta_i}\right\}\sum_{t,x,a}(1-\lambda_{t-1})q^{P_t,\widehat{\pi}_t}(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right)
\end{aligned}$$

$$\leq \max_{i\in[m]}\left\{\frac{L}{\theta_i-\beta_i}\right\}\sum_{t,x,a} q^{P_t,\pi_t}(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right)$$

$$= \max_{i\in[m]}\left\{\frac{L}{\theta_i-\beta_i}\right\}\sum_{t,x,a} \frac{q^{P_t,\pi_t}(x,a)}{u_t(x,a)+\gamma}\left(u_t(x,a)-q_t(x,a)+\gamma\right)$$

$$\leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{L}{\theta_i-\beta_i}\right\}L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right)+\max_{i\in[m]}\left\{\frac{L}{\theta_i-\beta_i}\right\}\gamma|X||A|T,$$

where the last steps holds by Lemma A.9. Thus, combining the previous equations, we have, with probability at least $1-3\delta$, under the clean event:

$$④ \leq \mathcal{O}\left(\max_{i\in[m]}\left\{\frac{1}{\theta_i-\beta_i}\right\}L^2|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{4.26}$$

Bound on ⑤ The last term is bounded as in Theorem 4.2. Thus, setting $\gamma = \sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$, we obtain, with probability at least $1-\delta$, under the clean event:

$$⑤ \leq \mathcal{O}\left(L\sqrt{|X||A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right). \tag{4.27}$$

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

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

which concludes the proof. □

The regret bound in Theorem 4.4 is in line with the one of SV-OPS in the bounded violation setting, with an additional ΨL^2 factor. Such a factor comes from the mismatch between loss estimators and the occupancy measure chosen by the OMD-like update. Notice that Ψ depends on the violation gap $\min_{i\in[m]}\{\theta_i-\beta_i\}$, which represents how much the strictly feasible solution satisfies the constraints. Such a dependence is expected, since the better the strictly feasible solution (in terms of constraints satisfaction), the larger the exploration performed during the first episodes.

4.5 Guaranteeing Constant Violation

In this section, we provide an algorithm that attains *constant* cumulative strong violation. To achieve this goal, we only need that a strictly feasible policy exists (Condition 2.1). Algorithm 4.3 provides the pseudocode of Constant Violation Optimistic Policy Search (CV-OPS). The key idea of CV-OPS is to estimate, in a constant number of episodes, a strictly feasible policy and its associated violation, and then run S-OPS with such estimates. Algorithm 4.3 needs access to two anytime no-regret algorithms, one for adversarial MDPs with bandit feedback that learns an estimated strictly feasible pol-

Algorithm 4.3 Constant Violation Optimistic Policy Search (CV-OPS)

Require: Anytime adversarial MDPs regret minimizer $\mathcal{A}^P$, online linear optimizer $\mathcal{A}^D$

1: **for** $t \in [T]$ **do**
2: Select $\pi_t \leftarrow \mathcal{A}^P$
3: Select $\phi_t \leftarrow \mathcal{A}^D$
4: Play π_t and observe *bandit feedback* as prescribed in Algorithm 2.1
5: Feed $\{x_k, a_k, \sum_{i\in[m]} \phi_{t,i}(g_{t,i}(x_k,a_k) - \frac{\theta_i}{L})\}_{k=0}^{L-1}$ to $\mathcal{A}^P$
6: Feed $\{-\sum_{k=0}^{L-1}(g_{t,i}(x_k,a_k) - \frac{\theta_i}{L})\}_{i\in[m]}$ to $\mathcal{A}^D$
7: **if** $-\max_{i\in[m]} \sum_{\tau\in[t]} \sum_{k=0}^{L-1}(g_{t,i}(x_k,a_k) - \frac{\theta_i}{L}) \geq 2C_{\mathcal{A}}^P\sqrt{t\ln(t)} + 8L\sqrt{2t\ln\frac{1}{\delta}} + 2C_{\mathcal{A}}^D\sqrt{t}$ **then**
8: Go to Line 11
9: **end if**
10: **end for**
11: $\widehat{\rho} \leftarrow -\frac{1}{t}\max_{i\in[m]} \sum_{\tau\in[t]} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k) - \theta_i\right) - \frac{2L}{t}\sqrt{2t\ln 1/\delta}$
12: $\widehat{\pi}^\diamond \leftarrow \pi_\tau$ with probability $1/t$, for $\tau \in [t]$
13: **Run** S-OPS with $\beta_i = \theta_i - \widehat{\rho}$ for all $i \in [m]$ and $\pi^\diamond = \widehat{\pi}^\diamond$

icy and one for the full feedback setting on the simplex, which learns the most violated constraint. Specifically, CV-OPS employs an anytime regret minimizer for adversarial MDPs—called the primal algorithm $\mathcal{A}^P$—that attains, with probability at least $1 - C_P^\delta\delta$, for all $\tau \in [T], q \in \Delta(M)$, and for any sequence of loss functions the following regret bound $\sum_{t=1}^{\tau} \ell_t^\top(q_t - q) \leq C_{\mathcal{A}}^P\sqrt{\tau\ln(\tau)}$,where $C_{\mathcal{A}}^P$ encompass constant terms. This kind of guarantees are attained by state-of-the-arts algorithms for adversarial MDPs (e.g., (Jin et al., 2020a)) after applying a standard doubling trick (Lattimore and Szepesvári, 2020). Algorithm 4.3 also employs an anytime online linear optimizer—called the dual algorithm $\mathcal{A}^D$)—that attains for all $\tau \in [T], \phi \in \Delta_m$, and for any sequence of loss functions the following regret bound $\sum_{t=1}^{\tau} \ell_t^\top(\phi_t - \phi) \leq C_{\mathcal{A}}^D\sqrt{\tau}$, where $C_{\mathcal{A}}^D$ encompasses constant terms. This bound can be easily obtained by an online gradient descent algorithm (Orabona, 2019).

At each episode $t \in [T]$, Algorithm 4.3 requests a policy and a distribution over the m constraints to $\mathcal{A}^P$ and $\mathcal{A}^D$, respectively (Lines 2-3). Thus, the algorithm plays the policy received by $\mathcal{A}^P$ and observes the usual bandit feedback for CMDPs (Line 4). Next, the loss functions for both the primal and the dual algorithm are built. Specifically, $\mathcal{A}^P$ receives the violation attained by the policy π_t where any constraint is weighted given ϕ_t (Line 5), while $\mathcal{A}^D$ receives the negative of the violation attained for all $i \in [m]$ (Line 6). The estimation phase stops when the violation attained by the algorithm exceeds twice the regret bounds attained by both the primal and the dual algorithm plus the uncertainty on the estimation (Line 7). This condition is suitably chosen to ensure that the number of episodes are sufficient to estimate an approximation of $\pi^\diamond$, while still being constant. After the estimation, Algorithm 4.3 computes pessimistically the estimated Slater's parameter $\widehat{\rho}$ as the average violation attained during the estimation phase minus a quantity associated with the uncertainty of the estimation (Line 11). Finally, CV-OPS computes the strictly feasible solution as the uniform policy with respect to all the policies played in the estimation phase (Line 12) and runs S-OPS with $\beta_i = \theta_i - \widehat{\rho}$, for all $i \in [m]$ and $\pi^\diamond = \widehat{\pi}^\diamond$ in input (Line 13).

4.5.1 Cumulative Strong Constraints Violation

First, we show that the stopping condition at Line 7 allows to run the estimation phase for no more than a constant number of episodes. This is done in the following lemma.

Lemma 4.2. *Given any $\delta \in (0,1)$, the episodes that Algorithm 4.3 uses to compute $\widehat{\rho}$ and $\widehat{\pi}^{\diamond}$ are $\bar{t} \leq {}^{1}\!/_{\rho^4}(3C_{\mathcal{A}}^{P} + 10L\ln\frac{1}{\delta} + 3C_{\mathcal{A}}^{D} + L)^4$, with prob. at least $1-(C_P^{\delta}+2)\delta$.*

Proof. By the no-regret property of Algorithm $\mathcal{A}^P$, it holds:

$$\sum_{t=1}^{\bar{t}} \sum_{i\in[m]} \phi_{t,i}\left(g_{t,i}^{\top}(q_t - q^{\diamond}) - \theta_i\right) \leq C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})},$$

with probability at least $1 - C_P^{\delta}\delta$.
Thus, applying the Azuma inequality, we get:

$$\begin{aligned}
&\sum_{t=1}^{\bar{t}} \sum_{i\in[m]} \phi_{t,i}(g_{t,i} - {}^{\theta_i}\!/_{L})^{\top} q_t \\
&\qquad \leq C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})} + L\sqrt{2\bar{t}\ln\frac{1}{\delta}} + \sum_{t=1}^{\bar{t}} \sum_{i\in[m]} \phi_{t,i}(\overline{g}_i - {}^{\theta_i}\!/_{L})^{\top} q^{\diamond} \\
&\qquad = C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})} + L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - \sum_{t=1}^{\bar{t}} \sum_{i\in[m]} \phi_{t,i}(\theta_i - \beta_i) \\
&\qquad \leq C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})} + L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - \bar{t}\min_{i\in[m]}(\theta_i - \beta_i) \\
&\qquad = C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})} + L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - \bar{t}\rho,
\end{aligned}$$

with probability $1-(C_P^{\delta}+1)\delta$, by Union Bound. Similarly, applying the Azuma inequality, it holds:

$$\sum_{t=1}^{\bar{t}_i} \sum_{i\in[m]} \phi_{t,i} \sum_{k=0}^{L-1}(g_{t,i}(x_k,a_k) - {}^{\theta_i}\!/_{L}) \leq C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})} + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - \bar{t}\rho,$$

with probability at least $1-(C_P^{\delta}+2)\delta$.
By the no-regret property of Algorithm $\mathcal{A}^D$, it holds:

$$-\sum_{t=1}^{\bar{t}} \sum_{i\in[m]} \phi_{t,i}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k) - \theta_i\right) - \sum_{t=1}^{\bar{t}} -\left(\sum_{k=0}^{L-1} g_{t,i^*}(x_k,a_k) - \theta_{i^*}\right) \leq C_{\mathcal{A}}^{D}\sqrt{\bar{t}},$$

where $i^* = \arg\max_{i\in[m]} \sum_{t=1}^{\bar{t}} \sum_{k=0}^{L-1}(g_{t,i}(x_k,a_k) - {}^{\theta_i}\!/_{L})$, from which we obtain, with probability at least $1-(C_P^{\delta}+2)\delta$:

$$\max_{i\in[m]} \sum_{t=1}^{\bar{t}}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k) - \theta_i\right) \leq C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})} + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}} + C_{\mathcal{A}}^{D}\sqrt{\bar{t}} - \bar{t}\rho.$$

By the stopping condition of Algorithm 4.3, it holds:

$$-\max_{i\in[m]}\sum_{t=1}^{\bar{t}}\sum_{k=0}^{L-1}(g_{t,i}(x_k,a_k)-{}^{\theta_i}/_{L})\geq 2C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}+8L\sqrt{2\bar{t}\ln\frac{1}{\delta}}+2C_{\mathcal{A}}^{D}\sqrt{\bar{t}}.$$

Equivalently, by the same stopping condition, it holds:

$$\begin{aligned}-\max_{i\in[m]}\sum_{t=1}^{\bar{t}-1}\sum_{k=0}^{L-1}&(g_{t,i}(x_k,a_k)-{}^{\theta_i}/_{L})\\&\leq 2C_{\mathcal{A}}^{P}\sqrt{\bar{t}-1\ln(\bar{t}-1)}+8L\sqrt{2\bar{t}-1\ln\frac{1}{\delta}}+2C_{\mathcal{A}}^{D}\sqrt{\bar{t}-1}\\&\leq 2C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}+8L\sqrt{2\bar{t}\ln\frac{1}{\delta}}+2C_{\mathcal{A}}^{D}\sqrt{\bar{t}},\end{aligned}$$

which implies:

$$\max_{i\in[m]}\sum_{t=1}^{\bar{t}-1}\sum_{k=0}^{L-1}(g_{t,i}(x_k,a_k)-{}^{\theta_i}/_{L})\geq -2C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}-8L\sqrt{2\bar{t}\ln\frac{1}{\delta}}-2C_{\mathcal{A}}^{D}\sqrt{\bar{t}}.$$

Thus, we get the following inequality:

$$-2C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}-8L\sqrt{2\bar{t}\ln\frac{1}{\delta}}-2C_{\mathcal{A}}^{D}\sqrt{\bar{t}}-L\leq C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}+2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}+C_{\mathcal{A}}^{D}\sqrt{\bar{t}}-\bar{t}\rho.$$

Hence,

$$\begin{aligned}\bar{t}\rho&\leq 3C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}+10L\sqrt{2\bar{t}\ln\frac{1}{\delta}}+3C_{\mathcal{A}}^{D}\sqrt{\bar{t}}+L\\&\leq 3C_{\mathcal{A}}^{P}\bar{t}^{3/4}+10L\sqrt{2\ln\frac{1}{\delta}}\bar{t}^{3/4}+3C_{\mathcal{A}}^{D}\bar{t}^{3/4}+L\bar{t}^{3/4},\end{aligned}$$

which implies:

$$\bar{t}\leq\frac{\left(3C_{\mathcal{A}}^{P}+10L\sqrt{2\ln\frac{1}{\delta}}+3C_{\mathcal{A}}^{D}+L\right)^{4}}{\rho^{4}}.$$

This concludes the proof. □

The bound could be reduced to $\bar{t}\leq {}^{1}/_{\rho^2}(3C_{\mathcal{A}}^{P}+10L\ln\frac{1}{\delta}+3C_{\mathcal{A}}^{D}+L)^2$, with access to a no-regret algorithm without the logarithmic dependence on T in the bound. Next, we show that Algorithm 4.3 estimates a strictly feasible policy whose constraints margin is in $[\widehat{\rho},\rho]$, as follows.

Lemma 4.3. *Given any $\delta\in(0,1)$, Algorithm 4.3 guarantees:*

$$\widehat{\rho}\leq\min_{i\in[m]}(\theta_i-\overline{g}_i^{\top}q^{P,\widehat{\pi}^{\diamond}})\leq\rho,$$

with prob. at least $1-2\delta$.

Proof. It holds:

$$\begin{aligned}
\widehat{\rho} &= -\frac{1}{\bar{t}}\left(\max_{i\in[m]}\sum_{t=1}^{\bar{t}}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k) - {}^{\theta_i}/_{L}\right) + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right) \\
&= \min_{i\in[m]}\left(\theta_i - \frac{1}{\bar{t}}\left(\sum_{t=1}^{\bar{t}}\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k) + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\right) \\
&\leq \min_{i\in[m]}\left(\theta_i - \frac{1}{\bar{t}}\left(\sum_{t=1}^{\bar{t}}\sum_{k=0}^{L-1} \overline{g}_i(x_k,a_k) - L\sqrt{2\bar{t}\ln\frac{1}{\delta}} + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\right) \\
&\leq \min_{i\in[m]}\left(\theta_i - \frac{1}{\bar{t}}\left(\sum_{t=1}^{\bar{t}} \overline{g}_i^\top q_t - 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}} + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\right) \\
&= \min_{i\in[m]}\left(\theta_i + \frac{1}{\bar{t}}\left(\sum_{t=1}^{\bar{t}} -\overline{g}_i^\top q_t + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\right) \\
&\leq \min_{i\in[m]}\left(\theta_i - \beta_i\right) \\
&= \rho,
\end{aligned}$$

where the first steps hold with probability at least $1-2\delta$ by Azuma inequality and a Union Bound and the last inequality holds by definition of $q^\diamond$.
To prove the second result we first notice that, by definition of $q^\diamond$:

$$\min_{i\in[m]}(\theta_i - \overline{g}_i^\top q^{P,\widehat{\pi}^\diamond}) \leq \min_{i\in[m]}(\theta_i - \overline{g}_i^\top q^\diamond) = \min_{i\in[m]}(\theta_i - \beta_i) = \rho.$$

Furthermore, by definition of $\widehat{\pi}^\diamond$, it holds:

$$q^{P,\widehat{\pi}^\diamond} = \frac{1}{\bar{t}}\sum_{t=1}^{\bar{t}} q^{P,\pi_t}.$$

Hence,

$$\begin{aligned}
\min_{i\in[m]}(\theta_i - \overline{g}_i^\top q^{P,\widehat{\pi}^\diamond}) &= \min_{i\in[m]}\left(\theta_i - \frac{1}{\bar{t}}\left(\sum_{t=1}^{\bar{t}} \overline{g}_i^\top q^{P,\pi_t}\right)\right) \\
&\geq \min_{i\in[m]}\left(\theta_i - \frac{1}{\bar{t}}\left(\sum_{t=1}^{\bar{t}}\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k) + 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\right) \\
&= \widehat{\rho},
\end{aligned}$$

where the first inequality holds with probability at least $1-2\delta$ employing the Azuma inequality and a Union Bound. This concludes the proof. □

Finally, we provide the result in term of cumulative strong constraints violation attained by our algorithm.

Theorem 4.5. *Given $\delta \in (0,1)$, Algorithm 4.3 attains:*

$$\mathcal{V}_T \leq \mathcal{O}\left(\frac{L}{\rho^4}\left(C_{\mathcal{A}}^P + L\sqrt{\ln\frac{1}{\delta}} + C_{\mathcal{A}}^D + L\right)^4\right),$$

with probability at least $1-(C_P^\delta+7)\delta$.

Theorem 4.5 is proved by employing Lemma 4.2 to bound the episodes in the estimation phase and Lemma 4.3 to state that S-OPS with $\widehat{\rho}, \widehat{\pi}^\diamond$ is safe with high probability.

Proof. We split the violations between the two phases of Algorithm 4.3 as:

$$\begin{aligned}
\mathcal{V}_T &\leq \max_{i\in[m]} \sum_{t=1}^{\bar{t}} \left[\overline{g}_i^\top q_t - \theta_i\right]^+ + \max_{i\in[m]} \sum_{t=\bar{t}+1}^{T} \left[\overline{g}_i^\top q_t - \theta_i\right]^+ \\
&\leq L\bar{t} + \max_{i\in[m]} \sum_{t=\bar{t}}^{T} \left[\overline{g}_i^\top q_t - \theta_i\right]^+ \\
&\leq \mathcal{O}\left(\frac{L\left(C_{\mathcal{A}}^P + L\sqrt{\ln\frac{1}{\delta}} + C_{\mathcal{A}}^D + L\right)^4}{\rho^4}\right) + \max_{i\in[m]} \sum_{t=\bar{t}+1}^{T} \left[\overline{g}_i^\top q_t - \theta_i\right]^+,
\end{aligned}$$

where the last step holds with probability at least $1-(C_P^\delta+2)\delta$ by Lemma 4.2.
In the following we show that, after $\bar{t}$ episodes, Algorithm 4.3 is safe with high probability. Similarly to Theorem 4.3 there are two possible scenarios defined by λ_t, namely, $\lambda_t = 0$ and $\lambda_t \in (0,1)$. When $\lambda_t = 0$, applying the same reasoning of Theorem 4.3 gives the result.
In the following analysis we consider a generic constraints $i \in \overline{m}$. Thus, we notice that the constraints cost attained by the *non-Markovian* policy π_t, is equal to $\lambda_{t-1}\overline{g}_i^\top q^{P,\widehat{\pi}^\diamond} + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t}$.
Then, we consider both the cases when $L < \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t$ (first case) and $L > \left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t$ (second case). If the two quantities are equivalent, the proof still holds breaking the ties arbitrarily.
First case. It holds that:

$$\begin{aligned}
\lambda_{t-1}\overline{g}_i^\top q^{P,\widehat{\pi}^\diamond} + (1-\lambda_{t-1})\overline{g}_i^\top q^{\widehat{\pi}_t,P} &= \frac{L-\theta_i}{L-\theta_i+\widehat{\rho}}(\overline{g}_i^\top q^{P,\widehat{\pi}^\diamond} - L) + L \\
&\leq \frac{\theta_i - L}{\theta_i - \widehat{\rho} - L}(\theta_i - \widehat{\rho} - L) + L \qquad (4.28) \\
&= \theta_i,
\end{aligned}$$

where Inequality (4.28) holds with probability at least $1-2\delta$ thanks to Lemma 4.3.
Second case. Similarly to the first case, it holds that, under the clean event:

$$\begin{aligned}
&\lambda_{t-1}\overline{g}_i^\top q^{P,\widehat{\pi}^\diamond} + (1-\lambda_{t-1})\overline{g}_i^\top q^{P,\widehat{\pi}_t} \\
&\leq \lambda_{t-1}\overline{g}_i^\top q^{P,\widehat{\pi}^\diamond} + (1-\lambda_{t-1})\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top q^{P,\widehat{\pi}_t} \qquad (4.29) \\
&\leq \lambda_{t-1}\overline{g}_i^\top q^{P,\widehat{\pi}^\diamond} + (1-\lambda_{t-1})\left(\widehat{g}_{t-1,i} + \xi_{t-1}\right)^\top \widehat{u}_t \qquad (4.30)
\end{aligned}$$

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

where Inequality (4.29) holds by the definition of the event and Inequality (4.30) holds by the definition of $\widehat{u}_t$.
To conclude the proof, we underline that λ_t is chosen taking the maximum over the constraints, which implies that the more conservative λ_t (the one which takes the combination nearer to the strictly feasible solution) is chosen. Thus, all the constraints are satisfied and a final Union Bound concludes the proof. □

4.5.2 Cumulative Regret

We provide the theoretical guarantees attained by Algorithm 4.3 in terms of cumulative regret. To do so, we show that $\widehat{\rho}$ is not too small. This is done in the following lemma.

Lemma 4.4. *Given any $\delta \in (0,1)$, Algorithm 4.3 guarantees $\widehat{\rho} \geq \rho/2$ with probability at least $1 - (C_P^\delta + 2)\delta$.*

Proof. Similarly to Lemma 4.2, we get, with probability at least $1 - (C_P^\delta + 2)\delta$:

$$
-\max_{i\in[m]} \sum_{t=1}^{\bar{t}} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k, a_k) - \theta_i \right) \geq -C_{\mathcal{A}}^P \sqrt{\bar{t}\ln(\bar{t})} - 2L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - C_{\mathcal{A}}^D \sqrt{\bar{t}} + \bar{t}\rho.
$$

Then, notice that, by the stopping condition of Algorithm 4.3, it holds:

$$
\begin{aligned}
&-\max_{i\in[m]} \sum_{t=1}^{\bar{t}} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k, a_k) - \theta_i \right) - C_{\mathcal{A}}^P \sqrt{\bar{t}\ln(\bar{t})} - 4L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - C_{\mathcal{A}}^D \sqrt{\bar{t}} \\
&\quad = -\frac{1}{2}\max_{i\in[m]} \sum_{t=1}^{\bar{t}} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k, a_k) - \theta_i \right) - \frac{1}{2}\max_{i\in[m]} \sum_{t=1}^{\bar{t}} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k, a_k) - \theta_i \right) \\
&\qquad\qquad - C_{\mathcal{A}}^P \sqrt{\bar{t}\ln(\bar{t})} - 4L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - C_{\mathcal{A}}^D \sqrt{\bar{t}} \\
&\quad \geq -\frac{1}{2}\max_{i\in[m]} \sum_{t=1}^{\bar{t}} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k, a_k) - \theta_i \right) + C_{\mathcal{A}}^P \sqrt{\bar{t}\ln(\bar{t})} - C_{\mathcal{A}}^P \sqrt{\bar{t}\ln(\bar{t})} \\
&\qquad\qquad + 4L\sqrt{2\bar{t}\ln\frac{1}{\delta}} - 4L\sqrt{2\bar{t}\ln\frac{1}{\delta}} + C_{\mathcal{A}}^D \sqrt{\bar{t}} - C_{\mathcal{A}}^D \sqrt{\bar{t}}
\end{aligned}
$$

$$\geq -\frac{1}{2}\max_{i\in[m]}\sum_{t=1}^{\bar{t}}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k)-\theta_i\right).$$

Hence,

$$\begin{aligned}
\widehat{\rho} &= -\frac{1}{\bar{t}}\left(\max_{i\in[m]}\sum_{t=1}^{\bar{t}}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k)-{}^{\theta_i}/_{L}\right)+2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\\
&= -\frac{1}{\bar{t}}\left(\max_{i\in[m]}\sum_{t=1}^{\bar{t}}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k)-{}^{\theta_i}/_{L}\right)\pm C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}\pm 4L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right.\\
&\qquad\qquad \left.\pm C_{\mathcal{A}}^{D}\sqrt{\bar{t}}+2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\\
&\geq \frac{1}{\bar{t}}\left(-\frac{1}{2}\max_{i\in[m]}\sum_{t=1}^{\bar{t}}\left(\sum_{k=0}^{L-1} g_{t,i}(x_k,a_k)-\theta_i\right)+C_{\mathcal{A}}^{P}\sqrt{\bar{t}\ln(\bar{t})}\right.\\
&\qquad\qquad \left.+4L\sqrt{2\bar{t}\ln\frac{1}{\delta}}+C_{\mathcal{A}}^{D}\sqrt{\bar{t}}-2L\sqrt{2\bar{t}\ln\frac{1}{\delta}}\right)\\
&\geq \frac{\rho}{2}.
\end{aligned}$$

This concludes the proof. □

Finally, we state the regret attained by CV-OPS.

Theorem 4.6. *Given any $\delta\in(0,1)$, with $\eta=\gamma=\sqrt{{}^{L\ln(L|X||A|/\delta)}/_{T|X||A|}}$, Algorithm 4.3 attains regret:*

$$R_T\leq\mathcal{O}\left(\Theta L^3|X|\sqrt{|A|T\ln\left(\frac{T|X||A|m}{\delta}\right)}+\frac{L}{\rho^4}\left(C_{\mathcal{A}}^{P}+L\ln\left(\frac{1}{\delta}\right)+C_{\mathcal{A}}^{D}+L\right)^4\right),$$

with probability at least $1-(C_P^{\delta}+13)\delta$, where we let $\Theta\coloneqq{}^{1}/_{\min\{\rho,\rho^2\}}$.

As Theorem 4.5, Theorem 4.6 is proved by employing Lemma 4.2 to bound the episodes in the estimation phase. Then, the result follows from the regret guarantees of S-OPS and Lemma 4.4 for ${}^{1}/_{\widehat{\rho}}\leq{}^{2}/_{\rho}$.

Proof. We split the regret between the two phases of Algorithm 4.3 as:

$$\begin{aligned}
R_T &\leq \sum_{t=1}^{\bar{t}}\ell_t^{\top}(q^{P,\pi_t}-q^*)+\sum_{t=\bar{t}+1}^{T}\ell_t^{\top}(q^{P,\pi_t}-q^*)\\
&\leq L\bar{t}+\sum_{t=\bar{t}+1}^{T}\ell_t^{\top}(q^{P,\pi_t}-q^*)
\end{aligned}$$

$$\leq \mathcal{O}\left(\frac{L\left(C_{\mathcal{A}}^{P}+L\sqrt{\ln\frac{1}{\delta}}+C_{\mathcal{A}}^{D}+L\right)^{4}}{\rho^{4}}\right)+\sum_{t=\bar{t}+1}^{T}\ell_{t}^{\top}(q^{P,\pi_{t}}-q^{*}),$$

where the last step holds with probability at least $1-(C_{P}^{\delta}+2)\delta$ by Lemma 4.2.
To bound the second terms we follow the steps of Theorem 4.4 after noticing that, for all $t\in[T]$:

$$\lambda_{t}\leq\max_{i\in[m]}\left\{\frac{L-\theta_{i}}{L-\theta_{i}+\widehat{\rho}}\right\}<1,$$

and that:

$$\begin{aligned}
\lambda_{t} &= \frac{\min\left\{(\widehat{g}_{t,i}+\xi_{t})^{\top}\widehat{u}_{t+1},L\right\}-\theta_{i}}{\min\left\{(\widehat{g}_{t,i}+\xi_{t})^{\top}\widehat{u}_{t+1},L\right\}-\theta_{i}+\widehat{\rho}} \\
&\leq \frac{(\widehat{g}_{t,i}+\xi_{t})^{\top}\widehat{u}_{t+1}-\theta_{i}}{(\widehat{g}_{t,i}+\xi_{t})^{\top}\widehat{u}_{t+1}-\theta_{i}+\widehat{\rho}} \\
&\leq \frac{(\widehat{g}_{t,i}+\xi_{t})^{\top}\widehat{u}_{t+1}-\theta_{i}}{\widehat{\rho}} \qquad (4.31) \\
&= \frac{(\widehat{g}_{t,i}-\xi_{t})^{\top}\widehat{u}_{t+1}+2\xi_{t}^{\top}\widehat{u}_{t+1}-\theta_{i}}{\widehat{\rho}} \\
&= \frac{(\widehat{g}_{t,i}-\xi_{t})^{\top}\widehat{q}_{t+1}+(\widehat{g}_{t,i}-\xi_{t})^{\top}(\widehat{u}_{t+1}-\widehat{q}_{t+1})+2\xi_{t}^{\top}\widehat{u}_{t+1}-\theta_{i}}{\widehat{\rho}} \\
&\leq \frac{(\widehat{g}_{t,i}-\xi_{t})^{\top}\widehat{q}_{t+1}+\widehat{g}_{t,i}^{\top}(\widehat{u}_{t+1}-\widehat{q}_{t+1})+2\xi_{t}^{\top}\widehat{u}_{t+1}-\theta_{i}}{\widehat{\rho}} \\
&\leq \frac{\widehat{g}_{t,i}^{\top}(\widehat{u}_{t+1}-\widehat{q}_{t+1})+2\xi_{t}^{\top}\widehat{u}_{t+1}}{\widehat{\rho}} \qquad (4.32) \\
&= \frac{\widehat{g}_{t,i}^{\top}(\widehat{u}_{t+1}-q^{P,\widehat{\pi}_{t+1}})+\widehat{g}_{t,i}^{\top}(q^{P,\widehat{\pi}_{t+1}}-q^{P_{t+1},\widehat{\pi}_{t+1}})+2\xi_{t}^{\top}\widehat{u}_{t+1}}{\widehat{\rho}} \\
&\leq \frac{\|\widehat{g}_{t,i}\|_{\infty}||\widehat{u}_{t+1}-q^{P,\widehat{\pi}_{t+1}}||_{1}+\|\widehat{g}_{t,i}\|_{\infty}\|q^{P,\widehat{\pi}_{t+1}}-q^{P_{t+1},\widehat{\pi}_{t+1}}\|_{1}+2\xi_{t}^{\top}\widehat{u}_{t+1}}{\widehat{\rho}} \\
&\leq \frac{||\widehat{u}_{t+1}-q^{P,\widehat{\pi}_{t+1}}||_{1}+\|q^{P,\widehat{\pi}_{t+1}}-q^{P_{t+1},\widehat{\pi}_{t+1}}\|_{1}+2\xi_{t}^{\top}\widehat{u}_{t+1}}{\widehat{\rho}} \\
&\leq \frac{L(1-\lambda_{t})\|\widehat{u}_{t+1}-q^{P,\widehat{\pi}_{t+1}}\|_{1}+L(1-\lambda_{t})\|q^{P,\widehat{\pi}_{t+1}}-q^{P_{t+1},\widehat{\pi}_{t+1}}\|_{1}}{\min\left\{\widehat{\rho},\widehat{\rho}^{\,2}\right\}} \\
&\qquad + \frac{2L(1-\lambda_{t})\xi_{t}^{\top}\widehat{u}_{t+1}}{\min\left\{\widehat{\rho},\widehat{\rho}^{\,2}\right\}} \qquad (4.33) \\
&\leq \frac{4L(1-\lambda_{t})\|\widehat{u}_{t+1}-q^{P,\widehat{\pi}_{t+1}}\|_{1}+4L(1-\lambda_{t})\|q^{P,\widehat{\pi}_{t+1}}-q^{P_{t+1},\widehat{\pi}_{t+1}}\|_{1}}{\min\left\{\rho,\rho^{2}\right\}} \\
&\qquad + \frac{8L(1-\lambda_{t})\xi_{t}^{\top}\widehat{u}_{t+1}}{\min\left\{\rho,\rho^{2}\right\}}, \qquad (4.34)
\end{aligned}$$

where Inequality (4.31) holds since, for the hardest constraint, when $\lambda_t \neq 0$, $(\widehat{g}_{t,i} + \xi_t)^\top \widehat{u}_{t+1} > \theta_i$, Inequality (4.32) holds since, under the clean event, $(\widehat{g}_{t,i} - \xi_t)^\top \widehat{q}_{t+1} \leq \theta_i$, Inequality (4.33) holds since $\lambda_t \leq \frac{L-\theta_i}{L-\theta_i+\widehat{\rho}}$ and Inequality (4.34) holds with probability at least $1 - (C_P^\delta + 2)\delta$ by Lemma 4.4. A Union Bound concludes the proof. □

Differently from S-OPS, the bound of CV-OPS scales as the inverse of the Slater's parameter ρ, not as the (possibly smaller) margin of a generic strictly feasible policy. Thus, the bound of Algorithm 4.3 is asymptotically smaller than the one of S-OPS.

4.5.3 Lower Bound on the Regret

We conclude by showing that a dependency on the feasibility of the strictly feasible solution in the regret bound is unavoidable to guarantee violation of order $o(\sqrt{T})$, *which is the case of both the second and third setting*.

This is done by means of the following lower bound.

Theorem 4.7. *There exist two instances of CMDPs (with a single state and one constraint) such that, if in the first instance an algorithm suffers from a violation $\mathcal{V}_T = o(\sqrt{T})$ probability at least $1 - n\delta$ for any $\delta \in (0, 1)$ and $n > 0$, then, in the second instance, it must suffer from a regret $R_T = \Omega(\frac{1}{\rho}\sqrt{T})$ with probability $3/4 - n\delta$.*

Proof. We consider two instances defined as follows. Both of them are characterized by a CMDP with one state (which is omitted for simplicity), two actions a_1, a_2, one constraint and $\theta = 1/2$. For the sake of simplicity we consider CMDP with rewards in place of losses. Notice that this is without loss of generality since any losses can be converted to an associated reward. We assume that the rewards are deterministic while the constraints are Bernoulli distributions with means defined in the following. Specifically, instance i^1 and instance i^2 are defined as:

$$i^1 := \begin{cases} \overline{r}(a_1) = \frac{1}{2}, & \overline{g}(a_1) = \frac{1}{2} + \epsilon \\ \overline{r}(a_2) = 0, & \overline{g}(a_2) = \frac{1}{2} - \rho \end{cases},$$

$$i^2 := \begin{cases} \overline{r}(a_1) = \frac{1}{2}, & \overline{g}(a_1) = \frac{1}{2} \\ \overline{r}(a_2) = 0, & \overline{g}(a_2) = \frac{1}{2} - \rho \end{cases},$$

where ϵ is a parameter to be defined later. Thus, since the algorithm must suffer $o(\sqrt{T})$ violation, for any constant $c > 0$, it holds:

$$\mathbb{P}^1 \left\{ q_t(a_2) \geq \frac{\epsilon}{\epsilon + \rho} - c\frac{1}{\sqrt{T}}, \quad \forall t \in [T] \right\} \geq 1 - n \cdot \delta,$$

where $q(a_2)$ is the occupancy measure associated to action a_2 and $\mathbb{P}^1$ is the probability measure of instance i_1 which encompasses the randomness of both environment and algorithm. Thus we can rewrite the inequality above as:

$$\mathbb{P}^1 \left\{ \sum_{t=1}^{T} q_t(a_2) \geq T\frac{\epsilon}{\epsilon + \rho} - c\sqrt{T} \right\} \geq 1 - n \cdot \delta.$$

By means of the Pinsker's inequality we can relate the probability measures $\mathbb{P}^1$ and $\mathbb{P}^2$ as follows:

$$\begin{aligned}&\mathbb{P}^2\left\{\sum_{t=1}^{T} q_t(a_2) \geq T\frac{\epsilon}{\epsilon+\rho} - c\sqrt{T}\right\}\\ &\geq \mathbb{P}^1\left\{\sum_{t=1}^{T} q_t(a_2) \geq T\frac{\epsilon}{\epsilon+\rho} - c\sqrt{T}\right\} - \sqrt{\frac{1}{2}KL(i^1, i^2)},\end{aligned}$$

where $KL(i^1, i^2)$ is the the KL-divergence between the probability measures of instance i_1 and i_2.
Noticing that by standard KL-decomposition argument, $KL(i^1, i^2) \leq \epsilon^2 T$, we have:

$$\mathbb{P}^2\left\{\sum_{t=1}^{T} q_t(a_2) \geq T\frac{\epsilon}{\epsilon+\rho} - c\sqrt{T}\right\} \geq 1 - n\cdot\delta - \epsilon\sqrt{\frac{T}{2}}.$$

We then notice that, since the rewards are deterministic, the regret of the second instance R_T^2 is bounded as:

$$\begin{aligned}R_T^2 &= \frac{1}{2}\sum_{t=1}^{T} q_t(a_2)\\ &\geq \frac{1}{2}T\frac{\epsilon}{\epsilon+\rho} - \frac{c}{2}\sqrt{T}\\ &\geq \frac{1}{4\rho}T\epsilon - c\sqrt{T}\\ &= \frac{1}{16\rho}\sqrt{2T} - c\sqrt{T}\\ &\geq \frac{1}{32\rho}\sqrt{2T},\end{aligned}$$

with probability $\frac{3}{4} - n\cdot\delta$, taking $\epsilon = \frac{1}{4}\sqrt{\frac{2}{T}}$, $\rho \geq \epsilon$ and $c \leq \frac{\sqrt{2}}{32}$.
This concludes the proof. □

Notice that this lower bound holds for *any* algorithm attaining a violation bound that is $o(\sqrt{T})$, thus, it is still applicable to settings where the violations are allowed to be much larger than the ones attained by Algorithm 4.2 and Algorithm 4.3.

Part III

Best-of-Both-Worlds

CHAPTER 5

A Best-of-Both-Worlds Algorithm for Full Feedback

In this chapter, we study *online learning* in episodic CMDPs under *full feedback*, when both the rewards and the constraints can be either stochastic or adversarial. All the existing works studying online learning problems in CMDPs address settings in which the constraints are selected stochastically according to an unknown (stationary) probability distribution. While these works address both the case where the rewards are stochastic (see, *e.g.*, (Zheng and Ratliff, 2020; Efroni et al., 2020)) and the one in which they are adversarial (see, *e.g.*, (Wei et al., 2018; Qiu et al., 2020)), to the best of our knowledge there is no work addressing settings with adversarially-selected constraints. Some works (see, *e.g.*, (Ding and Lavaei, 2023; Wei et al., 2023)) consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded. However, these results are *not* applicable to general settings with adversarial constraints.

In this chapter, we pioneer the study of CMDPs in which the constraints are selected adversarially. In doing so, we introduce an algorithm that employs a novel primal-dual approach in CMDPs, allowing it to attain *best-of-both-worlds* guarantees, in the flavor of Balseiro et al. (2023). In particular, our algorithm provides optimal (in the number of episodes T) regret and constraint violation bounds when rewards and constraints are selected either *stochastically* or *adversarially*, without requiring any knowledge of the underling process. While best-of-both-worlds algorithms have been recently introduced in online learning settings subject to constraints (see, *e.g.*, (Liakopoulos et al., 2019; Balseiro et al., 2023)), to the best of our knowledge our algorithm is the first of its kind in CMDPs.[1]

[1]Notice that, in the literature on online learning in MDPs, the term *best-of-both-worlds* is sometimes referred to algorithms that achieve optimal instance-dependent regret bounds when rewards are selected *stochastically* and $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when rewards are chosen *adversarially* (Jin et al., 2021). In this work, we borrow terminology from the literature on online learning with constraints, where the term usually refers to algorithms that achieve optimal regret and constraint violation bounds when the constraints are selected either *stochastically* or *adversarially* (Balseiro et al., 2023).

When the constraints are selected stochastically, we show that our algorithm provides $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative regret and constraint violation when a suitably-defined Slater-like condition concerning the satisfiability of constraints is satisfied. Moreover, whenever such a condition does *not* hold, our algorithm still ensures $\widetilde{\mathcal{O}}(T^{3/4})$ regret and constraint violation. Instead, whenever the constraints are chosen adversarially, our analysis revolves around the parameter ρ which is related to our Slater-like condition, and in particular to the "margin" by which it is possible to strictly satisfy the constraints. Indeed, under adversarial constraints, Mannor et al. (2009) show that it is impossible to simultaneously achieve sublinear regret and sublinear cumulative constraint violation. We prove that our algorithm achieves no-α-regret with $\alpha = \rho/(L+\rho)$, which is optimal (Bernasconi et al., 2025), while guaranteeing that the cumulative constraint violation is sublinear in the number of episodes. *This matches the regret guarantees derived for other best-of-both-worlds algorithms in (non-sequential) online learning settings* (Castiglioni et al., 2022a; Balseiro et al., 2023), which are optimal whenever ρ is a constant independent on T.

Differently from previous works on online learning with adversarial constraints, in this work we *relax the strong assumption* that the algorithm has to know the value of the parameter ρ related to Slater's condition. This assumption is ubiquitous in the adversarially-constrained online optimization literature (see, *e.g.*, (Castiglioni et al., 2022b)), but it is *extremely* unreasonable in practice. Indeed, in real-world scenarios, the learner has usually no clue about the "margin" by which a strictly feasible solution satisfies the constraints. Relaxing such an assumption is a non-trivial task from a technical perspective. This is done by proving that our primal-dual algorithm guarantees that dual variables are automatically bounded, by showing that both the primal and the dual regret minimizers attain a strong no-regret property, called *no-interval regret*. This is crucial since the classical (weaker) no-regret property is *not* enough to ensure that dual variables are automatically bounded, thus preventing from employing *primal-dual* methods.

5.1 Setting and Additional Notation

We study *episodic constrained* MDPs, when *full feedback* is available. Both the rewards and the costs can be either *stochastic* or *adversarial*. To be coherent with the *best-of-both-worlds* constrained online learning literature, we will define the costs as follows. $\{G_t\}_{t=1}^T$ is a sequence of constraint matrices describing the m *constraint* violations at each episode $t \in [T]$, namely $G_t \in [-1,1]^{|X\times A|\times m}$, where non-strictly positive violation values stand for constraints satisfaction. Thus, the optimum definition reduces to:

$$\text{OPT}_{r,G} := \begin{cases} \max_{q\in\Delta(M)} & r^\top q \\ \text{s.t.} & G^\top q \leq \underline{0}. \end{cases}$$

Similarly, the *cumulative constraint violation* reduces to:

$$V_T := \max_{i\in[m]} \sum_{t=1}^{T} \left[G_t^\top q^{P,\pi_t}\right]_i.$$

5.1.1 Feasibility Parameter

In this chapter, we will make use of the following condition on the value of ρ, which plays a central role when proving algorithm guarantees in the following sections.

Condition 5.1. *It holds that* $\rho \geq T^{-\frac{1}{8}} L\sqrt{20m}$.

While Condition 5.1 may seem unusual in the CMDPs literature, we remark that state-of-the-art primal-dual methods assume that the parameter ρ is a constant, and, thus, they simply hide the dependence on $1/\rho$ in the regret bound or in the $\mathcal{O}$-notation (see, *e.g.*, (Efroni et al., 2020; Qiu et al., 2020)). Thus, if the parameter ρ is arbitrarily small, their regret bounds may be superlinear in T. As we show next, our primal-dual method for CMDPs is the first to work even in degenerate scenarios, namely, when ρ is arbitrarily small.

5.2 Constrained MDP Optimization Algorithm

In this section, we present our algorithm named *primal-dual gradient descent online policy search* (`PDGD-OPS`). Its rationale is to instantiate two no-regret algorithms, referred to as *primal* and *dual player*, respectively. Precisely, the primal player optimizes on the primal variable space of the Lagrangian function, namely on the set $\Delta(M)$, while the dual player does it on the dual variable space $\mathbb{R}^m_{\geq 0}$, which, in our algorithm, is properly shrunk to $\left[0, T^{1/4}\right]^m$. As concerns the objective functions, the primal player aims at maximizing the Lagrangian function, while the dual one at minimizing it, as described in the Lagrangian zero-sum game defined in Corollary 2.1. Notice that, while the space of the dual variables is known *apriori*, the occupancy measure space needs be estimated online as the transition probabilities are unknown. Thus, it is necessary to employ a no-regret algorithm working with adversarial MDPs for the primal player. Moreover, in order to provide guarantees on the dynamics of the Lagrange multipliers—necessary to bound the cumulative regret and cumulative constraint violation—we require that the primal player satisfies the weak no-interval regret property (see the following Definition 5.2 for a formalization of such a property).

5.2.1 PDGD-OPS Algorithm

Algorithm 5.1 provides the pseudo-code of the `PDGD-OPS` algorithm. As mentioned before, the algorithm employs two regret minimizers, named `UC-O-GDPS` and `OGD`, working on the space of the primal and dual variables, respectively. The occupancy measure is initialized uniformly (Line 1) by the primal player. We refer to Section 5.3 for the description of the `UC-O-GDPS` initialization. The dual player is initialized by the `OGD.INIT` procedure which takes as input the decision space $\mathcal{D}$ and a learning rate η, and it returns the vector $\lambda_1 = \underline{0}$ associated with the dual variable (Line 2).

During the learning process, at each episode $t \in [T]$, the `PDGD-OPS` algorithm plays the policy $\pi^{\widehat{q}_t}$ induced by the occupancy measure $\widehat{q}_t$ computed in the previous episode (Line 4). The feedback received by the learner once the episode is concluded concerns the trajectory $(x_k, a_k)_{k=0}^{L-1}$ traversed in the CMDP, the reward vector, and the constraint matrix for that specific episode.

Given the observed feedback, the algorithm builds the Lagrangian objective function (Line 5), namely $\ell_t = G_t\lambda_t - r_t$, which is fed in the form of a loss into the primal player along with the trajectory and the adaptive learning rate (Line 6). The trajectory is needed to estimate the transition probabilities, while the rationale of the adaptive learning rate is to remove the quadratic dependence from $\|\lambda\|_1$ in the regret bound of the primal player. See Section 5.3 for the description of `UC-O-GDPS.UPDATE` (Line 7).

To conclude, we notice that the dual player receives only the loss $-G_t^\top \widehat{q}_t$, since the $r_t^\top \widehat{q}_t$ factor has no dependence on the optimization variable λ_t, and, thus, it does not affect the optimization process. For the sake of completeness, we report the OGD update of the dual player, namely OGD.UPDATE (Line 8), defined as follows:

$$\lambda_{t+1} := \Pi_{\mathcal{D}} \left(\lambda_t + \eta G_t^\top \widehat{q}_t \right), \tag{5.1}$$

where $\Pi_{\mathcal{D}}$ is the Euclidean projection on the decision space $\mathcal{D}$, $\eta = {}^1\!/_K \sqrt{T \ln\left(\frac{T^2}{\delta}\right)}$ and K is an instance-dependent quantity that does not depend on T and δ. From here on, we refer to the regret suffered by OGD with respect to a general Lagrange multiplier λ as $R_T^{\mathsf{D}}(\lambda)$, where D stands for *dual*. Notice that, thanks to the properties of OGD (Orabona, 2019), by using the aforementioned learning rate η, we obtain $R_T^{\mathsf{D}}(\lambda) \leq \widetilde{\mathcal{O}}\left((1 + ||\lambda||_2^2)\sqrt{T}\right)$.

Algorithm 5.1 Primal-Dual Gradient Descent Online Policy Search (PDGD-OPS)

Require: T, X, A, δ

1: $\widehat{q}_1 \leftarrow$ UC-O-GDPS.INIT (X, A, δ)
2: $\lambda_1 \leftarrow$ OGD.INIT $\left(\left[0, T^{1/4}\right]^m, \eta\right)$
3: **for** $t = 1$ to T **do**
4: Play $\pi^{\widehat{q}_t}$ and observe trajectory $(x_k, a_k)_{k=0}^{L-1}$, reward vector r_t, and constraint matrix G_t
5: $\ell_t \leftarrow G_t \lambda_t - r_t$
6: $\eta_t \leftarrow \frac{1}{\overline{\ell}_t C \sqrt{T}}$ with $\overline{\ell}_t = \max\{||\ell_\tau||_\infty\}_{\tau=1}^t$
7: $\widehat{q}_{t+1} \leftarrow$ UC-O-GDPS.UPDATE $\left(\ell_t, \eta_t, (x_k, a_k)_{k=0}^{L-1}\right)$
8: $\lambda_{t+1} \leftarrow$ OGD.UPDATE $\left(-G_t^\top \widehat{q}_t\right)$
9: **end for**

5.3 Adversarial MDP Optimization Algorithm

We focus on the algorithm employed by the primal player. As previously discussed, this algorithm resorts to online learning techniques, since the decision space of the primal player is not known beforehand. In particular, the algorithm is a regret minimizer for adversarial MDPs, as Algorithm 5.1 deals with both stochastic and adversarial settings.

5.3.1 UC-O-GDPS Algorithm

Upper confidence online gradient descent policy search (UC-O-GDPS) follows the rationale of the UC-O-REPS algorithm by Rosenberg and Mansour (2019b), from which we highlight two major differences. The first difference concerns the update step. In particular, while in UC-O-REPS the update is performed by online mirror descent when the unnormalized KL is used as Bregman divergence, in UC-O-GDPS such a step is performed by online gradient descent. The use of online gradient descent allows the UC-O-GDPS algorithm to satisfy the weak no-interval regret property (see Definition 5.2) which plays a central role in our regret analysis. We also notice that, to the best of our knowledge, the weak no-interval regret property has never been studied in episodic adversarial MDPs, and thus our result may be of independent interest. The second difference concerns the design of an adaptive learning rate which depends on the losses previously observed. The satisfaction of weak no-interval regret property and the adoption of our adaptive learning rate allow us to attain a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ for PDGD-OPS in place of $\widetilde{\mathcal{O}}\left(T^{3/4}\right)$.

Transitions confidence set Initially, we discuss how UC-O-GDPS updates the confidence set, denoted with $\mathcal{P}$, on the transition probabilities P. In particular, the update of the confidence set requires a non-negligible computational effort, however it is possible to update the confidence set at a subset of episodes to make the UC-O-GDPS algorithm more efficient without worsening the regret bounds. More precisely, the episodes are divided dynamically in epochs depending of the observed feedback, and the update of the confidence bound is only performed at the first episode of every epoch. UC-O-GDPS adopts counters of visits for each state-action pair (x, a) and each state-action-state triple (x, a, x') to estimate the empirical transition function as:

$$\overline{P}_j\left(x' \mid x, a\right) = \frac{M_j\left(x' \mid x, a\right)}{\max\left\{1, N_j(x, a)\right\}},$$

where $N_j(x, a)$ and $M_j\left(x' \mid x, a\right)$ are the initial values of the counters, that is, the total number of visits of pair (x, a) and triple (x, a, x'), respectively, observed in the epochs preceding epoch j. Furthermore, a new epoch starts whenever there is a state-action pair whose counter is doubled compared to its initial value at the beginning of the epoch. The confidence set $\mathcal{P}_j$ is updated at every epoch j as, for every $(x, a) \in X \times A$:

$$\mathcal{P}_j = \left\{\widehat{P} : \left\|\widehat{P}\left(\cdot|x, a\right) - \overline{P}_j\left(\cdot|x, a\right)\right\|_1 \leq \epsilon_j\left(x, a\right)\right\}, \tag{5.2}$$

where $\epsilon_j\left(x, a\right)$ is defined as:

$$\epsilon_j\left(x, a\right) = \sqrt{\frac{2|X_{k(x)+1}| \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\left\{1, N_j(x, a)\right\}}},$$

and $k(x)$ denotes the index of the layer to which x belongs and $\delta \in (0, 1)$ is the given confidence. The next result, which follows from (Rosenberg and Mansour, 2019b), shows that the cumulative error due to the estimation of the transition probabilities grows sublinearly.

Lemma 5.1 (Rosenberg and Mansour (2019b)). *If the confidence set $\mathcal{P}$ is updated as in Equation (5.2), with probability at least $1 - 2\delta$, it holds that:*

$$\sum_{t=1}^{T} ||q_t - \widehat{q}_t||_1 \leq \mathcal{E}_\delta^q,$$

where $\mathcal{E}_\delta^q \leq \widetilde{\mathcal{O}}(\sqrt{T})$.

Initialization Algorithm 5.1 employs the procedure called UC-O-GDPS.INIT (Line 1) to initialize the epoch index as $j = 1$ and the confidence set $\mathcal{P}_1$ as the set of all possible transition functions. For all $k \in [0, \ldots, L-1]$ and $(x, a, x') \in X_k \times A \times X_{k+1}$, the counters are initialized as $N_0(x, a) = N_1(x, a) = M_0\left(x' \mid x, a\right) = M_1\left(x' \mid x, a\right) = 0$. Finally, the following occupancy measure

$$\widehat{q}_1\left(x, a, x'\right) = \frac{1}{|X_k||A||X_{k+1}|}$$

is returned by the initialization procedure, for every $k \in [0, \ldots, L-1]$ and $(x, a, x') \in X_k \times A \times X_{k+1}$.

Update The pseudo-code of the `UC-O-GDPS.UPDATE` procedure (used in Line 7 of Algorithm 5.1) is provided in Algorithm 5.2. Initially, it updates the estimate of the confidence set $\mathcal{P}$ (Lines 1–7) as described above, and, subsequently, it performs an update step according to projected online gradient descent (Line 9).

Algorithm 5.2 Upper Confidence OGD Policy Search (`UC-O-GDPS.UPDATE`)

Require: $\ell_t, \eta_t, (x_k, a_k)_{k=0}^{L-1}$

1: **for** $k \in [0, \dots, L-1]$ **do**

2: Update counters:

$$N_j(x_k, a_k) \leftarrow N_j(x_k, a_k) + 1,$$
$$M_j(x_{k+1} \mid x_k, a_k) \leftarrow M_j(x_{k+1} \mid x_k, a_k) + 1$$

3: **end for**

4: **if** $\exists k : N_j(x_k, a_k) \geq \max\{1, 2N_{j-1}(x_k, a_k)\}$ **then**

5: Increase epoch index $j \leftarrow j + 1$

6: Initialize new counters: for all (x, a, x'),

$$N_j(x, a) \leftarrow N_{j-1}(x, a)$$
$$M_j(x' \mid x, a) \leftarrow M_{j-1}(x' \mid x, a)$$

7: Update confidence set $\mathcal{P}_j$ as in Equation (5.2)

8: **end if**

9: Update occupancy measure:

$$\widehat{q}_{t+1} \leftarrow \Pi_{\Delta(\mathcal{P}_j)}(\widehat{q}_t - \eta_t \ell_t)$$

5.3.2 Interval Regret

Initially, we provide the definition of interval regret for adversarial online MDPs.

Definition 5.1 (Interval regret)**.** *Given an interval of consecutive episodes $[t_1, \dots, t_2] \subseteq [1, \dots, T]$, the interval regret with respect to a general occupancy measure q is defined as:*

$$R_{t_1,t_2}(q) := \sum_{t=t_1}^{t_2} \ell_t^\top (q_t - q).$$

Now, we define the notion of weak no-interval regret. This notion plays a crucial role when proving the properties of Algorithm 5.1, and it is defined as follows.

Definition 5.2 (Weak no-interval regret)**.** *An online MDP optimizer satisfies the weak no-interval regret property if:*

$$R_{t_1,t_2}(q) \leq \widetilde{\mathcal{O}}\left(\sqrt{T}\right), \qquad \forall [t_1, \dots, t_2] \subseteq [1, \dots, T].$$

For ease of presentation, in the following we use the superscript P in the regret to distinguish the regret associated with the primal regret minimizer (R^{P}) from the regret associated with the dual regret minimizer (R^{D}), while we use $R_T^{\mathsf{P}}(q)$ in place of $R_{1,T}^{\mathsf{P}}(q)$.

Next, we state the main result of this section.

Theorem 5.1. *With probability at least* $1-2\delta$*, by setting* $\eta_t = \left(\overline{\ell}_t C\sqrt{T}\right)^{-1}$*, the* UC-O-GDPS *algorithm satisfies the following for any* $q \in \bigcap_j \Delta(\mathcal{P}_j)$*:*

$$R^{\mathsf{P}}_{t_1,t_2}(q) \leq \overline{\ell}_{t_1,t_2}\mathcal{E}^q_\delta + \overline{\ell}_{t_2} LC\sqrt{T} + \overline{\ell}_{t_1,t_2}\frac{|X||A|}{2}\frac{(t_2-t_1+1)}{C\sqrt{T}},$$

where $\overline{\ell}_{t_1,t_2} := \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$, $\overline{\ell}_t := \overline{\ell}_{1,t}$, $\delta \in [0,1]$.

Proof. Assume Event $E^{\Delta,\widehat{q}}(\delta)$ holds. By Definition 5.1:

$$\begin{aligned} R_{t_1,t_2}(q) &= \sum_{t=t_1}^{t_2} \ell_t^\top (q_t - q) \\ &= \underbrace{\sum_{t=t_1}^{t_2} \ell_t^\top (q_t - \widehat{q}_t)}_{①} + \underbrace{\sum_{t=t_1}^{t_2} \ell_t^\top (\widehat{q}_t - q)}_{②}. \end{aligned}$$

We focus on bounding the first term ① and the second term ②.
We start by the first term, proceeding as follows:

$$\begin{aligned} \sum_{t=t_1}^{t_2} \ell_t^\top (q_t - \widehat{q}_t) &\leq \sum_{t=t_1}^{t_2} ||\ell_t||_\infty ||q_t - \widehat{q}_t||_1 \\ &\leq \overline{\ell}_{t_1,t_2} \sum_{t=t_1}^{t_2} ||q_t - \widehat{q}_t||_1 \\ &\leq \overline{\ell}_{t_1,t_2} \sum_{t=1}^{T} ||q_t - \widehat{q}_t||_1 \\ &\leq \overline{\ell}_{t_1,t_2} \mathcal{E}^q_\delta, \end{aligned} \tag{5.3}$$

with $\overline{\ell}_{t_1,t_2} := \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$ and where Inequality (5.3) holds under the event $E^{\widehat{q}}(\delta)$.
We conclude bounding the second term as follows. By the standard analysis of projected online gradient descent [Lemma 2.12 (Orabona, 2019)], we have, for $q \in \bigcap_j \Delta(\mathcal{P}_j)$:

$$\ell_t^\top (\widehat{q}_t - q) \leq \frac{1}{2\eta_t}||\widehat{q}_t - q||_2^2 - \frac{1}{2\eta_t}||\widehat{q}_{t+1} - q||_2^2 + \frac{\eta_t}{2}||\ell_t||_2^2.$$

Observe that for any two occupancy measures q_1, q_2 it holds:

$$\begin{aligned} ||q_1 - q_2||_2^2 &\leq ||q_1||_2^2 + ||q_2||_2^2 \\ &\leq ||q_1||_1 + ||q_2||_1 \\ &\leq 2L, \end{aligned}$$

where the second Inequality follows from $q(x,a) \in [0,1] \ \ \forall x, a$. Then, summing over the interval $[t_1, \ldots, t_2]$ we get:

$$② \leq \frac{1}{2\eta_{t_1}}||\widehat{q}_{t_1} - q||_2^2 \underbrace{- \frac{1}{2\eta_{t_2}}||\widehat{q}_{t_2+1} - q||_2^2}_{\leq 0}$$

$$+\frac{1}{2}\sum_{t=t_1}^{t_2-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)||\widehat{q}_{t+1}-q||_2^2+\frac{1}{2}\sum_{t=t_1}^{t_2}\eta_t||\ell_t||_2^2$$

$$\leq\frac{L}{\eta_{t_1}}+L\sum_{t=t_1}^{t_2-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)+\frac{1}{2C\sqrt{T}}\sum_{t=t_1}^{t_2}\frac{1}{\overline{\ell}_t}\sum_{x,a}\ell_t(x,a)^2 \tag{5.4}$$

$$\leq\frac{L}{\eta_{t_1}}+L\underbrace{\sum_{t=t_1}^{t_2-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)}_{=\frac{1}{\eta_{t_2}}-\frac{1}{\eta_{t_1}}}+\frac{1}{2C\sqrt{T}}\sum_{t=t_1}^{t_2}\underbrace{\frac{||\ell_t||_\infty}{\max\{||\ell_\tau||_\infty\}_{\tau=1}^{t}}}_{\leq 1}||\ell_t||_\infty\sum_{x,a}1$$

$$\leq L\overline{\ell}_{t_2}C\sqrt{T}+\frac{|X||A|}{2}\overline{\ell}_{t_1,t_2}\frac{(t_2-t_1+1)}{C\sqrt{T}}, \tag{5.5}$$

where Inequality (5.4) follows from the definition of η_t, and from $\eta_t > \eta_{t+1}$, while Inequality (5.5) comes from the telescopic sum over $[t_1,\dots,t_2]$ and from the definition of η_{t_2}. This concludes the proof. □

Furthermore, it follows from Theorem 5.1 that, when $t_1 = 1$ and $t_2 = T$, it holds $R_T^{\mathsf{P}} \leq \widetilde{\mathcal{O}}\left(\overline{\ell}_T\sqrt{T}\right)$.

5.4 Theoretical Results

In this section, we provide the theoretical results attained by Algorithm 5.1 in terms of cumulative regret and cumulative constraint violation. We start providing a fundamental result on the Lagrange multiplier dynamics. Then, we distinguish two cases, which require different treatments. In the first, constraints are stochastic (Section 5.4.1), while in the second case they are adversarial (Section 5.4.2).

The main technical challenge when bounding the cumulative regret and constraint violation concerns bounding the space of the dual variables. We recall that, when employing standard no-regret techniques, an unbounded dual space would lead to an unbounded loss for the primal regret minimizer, resulting in a linear regret. Our choice $\mathcal{D} = [0, T^{1/4}]^m$ of the dual decision space allows us to circumvent such an issue and PDGD-OPS to achieve a cumulative regret bound of $R_T \leq \widetilde{\mathcal{O}}(T^{3/4})$, while keeping the cumulative violation sublinear. Nevertheless, when ρ is large enough (namely, Condition 5.1 holds), we can improve the $\widetilde{\mathcal{O}}(T^{3/4})$ dependency in the upper bounds. In particular, in this case, we can show that the Lagrangian vector never touches the boundaries of $\mathcal{D}$, and this property can be used to show that the regret and violation bounds are $\widetilde{\mathcal{O}}(\sqrt{T})$. In the following, we present our result on how the Lagrange multipliers can be bounded.

Theorem 5.2. *If Condition 5.1 holds and* `PDGD-OPS` *is used, then, when* $\zeta := \frac{20mL^2}{\rho^2}$, *it holds*

$$||\lambda_t||_1 \leq \zeta, \qquad \forall t \in [T+1]$$

with probability at least $1-2\delta$ *in the stochastic constraint setting and with probability at least* $1-\delta$ *in the adversarial constraint setting.*

The proof exploits the fact that both the primal and dual player satisfy the weak no-interval regret property. Precisely, the sum of the values of the Lagrangian function in

$[t_1, \dots, t_2]$ can be lower bounded by using the interval regret of UC-O-GDPS, while the same quantity can be upper bounded with the interval regret of OGD, showing a contradiction concerning the value of Lagrange multipliers can achieve for an opportune choice of constants and learning rates.

Proof. Suppose event $E^{\Delta}(\delta)$ holds. If the constraints are stochastic, suppose event $E^{G}_{q^\diamond}(\delta)$ holds too. Let $M > 1$ be a constant. We prove the statement by absurd. Suppose by absurd that there exists $t_2 \in [T]$ such that:

$$\forall t \leq t_2 \quad ||\lambda_t||_1 \leq \frac{2LM}{\rho^2} \qquad \wedge \qquad ||\lambda_{t_2+1}||_1 > \frac{2LM}{\rho^2}$$

and let $t_1 < t_2$ be such that:

$$||\lambda_{t_1-1}||_1 \leq \frac{2L}{\rho} \qquad \wedge \qquad \forall t : t_1 \leq t \leq t_2 \quad ||\lambda_t||_1 \geq \frac{2L}{\rho}.$$

By construction it holds that $1 < \frac{2L}{\rho} \leq ||\lambda_t||_1 \leq \frac{2LM}{\rho^2}$ for all $t_1 \leq t \leq t_2$. Also notice that by Lemma C.1, for $\eta \leq \frac{1}{mL}$ it holds that:

$$||\lambda_{t_1}||_1 \leq ||\lambda_{t_1-1}||_1 + m\eta L \leq \frac{2L}{\rho} + m\eta L \leq \frac{4L}{\rho}.$$

Focus on the quantity $\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond$: in the stochastic constraint setting we have, under the event $E^{G}_{q^\diamond}(\delta)$:

$$\begin{aligned}
\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond &\geq \sum_{t=t_1}^{t_2} -\lambda_t^\top \overline{G}^\top q^\diamond - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\
&\geq \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} -\lambda_{t,i}\left[\overline{G}^\top q^\diamond\right]_i - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\
&\geq \rho \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} \lambda_{t,i} - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\
&= \rho \sum_{t=t_1}^{t_2} ||\lambda_t||_1 - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\
&\geq \rho \frac{2L}{\rho}(t_2 - t_1 + 1) - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\
&= 2L(t_2 - t_1 + 1) - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2}.
\end{aligned}$$

While in the adversarial setting it holds:

$$\begin{aligned}
\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond &\geq \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} -\lambda_{t,i}\left[G_t^\top q^\diamond\right]_i \\
&\geq \rho \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} \lambda_{t,i}
\end{aligned}$$

$$
\begin{aligned}
&= \rho \sum_{t=t_1}^{t_2} ||\lambda_t||_1 \\
&\geq \rho \frac{2L}{\rho}(t_2 - t_1 + 1) \\
&= 2L(t_2 - t_1 + 1).
\end{aligned}
$$

In particular, we have that

$$
\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond \geq 2L(t_2 - t_1 + 1) - \lambda_{t_1,t_2} \mathcal{E}^G_{t_1,t_2}
$$

is true in both settings under the required events.

We can lower bound the cumulative value of the Lagrangian function, namely ${r_t^\mathcal{L}}^\top \widehat{q}_t$, from t_1 to t_2 by that achievable by the primal minimizer by always playing the feasible occupancy measure $q^\diamond$:

$$
\begin{aligned}
\sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top \widehat{q}_t &= \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top q^\diamond - \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top (q^\diamond - \widehat{q}_t) \\
&= \underbrace{\sum_{t=t_1}^{t_2} r_t^\top q^\diamond}_{\geq 0} + \sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond - \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top (q^\diamond - \widehat{q}_t) \\
&\geq 2L(t_2 - t_1 + 1) - \lambda_{t_1,t_2} \mathcal{E}^G_{t_1,t_2,\delta} - \sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top (q^\diamond - \widehat{q}_t).
\end{aligned}
$$

Applying Lemma C.2 and observing that by construction $1 \leq \lambda_{t_1,t_2} \leq \frac{2LM}{\rho^2}$, we can bound $1 + \lambda_{t_1,t_2} \leq \frac{4LM}{\rho^2}$ and obtain:

$$
\sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top \widehat{q}_t \geq 2L(t_2 - t_1 + 1) - \frac{2LM}{\rho^2} \mathcal{E}^G_{t_1,t_2,\delta} - U_1 \frac{2LM}{\rho^2} C\sqrt{T} - U_2 \frac{2LM}{\rho^2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}},
$$

since under $E^\Delta(\delta)$ we have that $q^\diamond \in \cap_i \Delta(\mathcal{P}_i)$.
We can upper-bound the same quantity with the value achievable by the dual by always playing a vector of zeroes.

$$
\begin{aligned}
\sum_{t=t_1}^{t_2} {r_t^\mathcal{L}}^\top \widehat{q}_t &= \sum_{t=t_1}^{t_2} r_t^\top \widehat{q}_t - \sum_{t=t_1}^{t_2} \lambda_t^\top G_t^\top \widehat{q}_t \\
&\leq \sum_{t=t_1}^{t_2} r_t^\top \widehat{q}_t - \sum_{t=t_1}^{t_2} \underline{0}^\top G_t^\top \widehat{q}_t + R^{\mathsf{D}}_{t_1,t_2}(\underline{0}) \\
&\leq \sum_{t=t_1}^{t_2} L + D_1 \frac{||\lambda_{t_1}||_2^2}{\eta} + D_2 \eta (t_2 - t_1 + 1)
\end{aligned}
$$

$$\leq \sum_{t=t_1}^{t_2} L + D_1 \frac{||\lambda_{t_1}||_1^2}{\eta} + D_2\eta(t_2 - t_1 + 1)$$
$$\leq L(t_2 - t_1 + 1) + D_3 \frac{L^2}{\rho^2\eta} + D_2\eta(t_2 - t_1 + 1),$$

with $D_3 = 4D_1$.
Combining the bounds on the cumulative value of the Lagrangian, we have:

$$2L(t_2 - t_1 + 1) - \frac{2LM}{\rho^2}\mathcal{E}^G_{t_1,t_2,\delta} - U_1\frac{2LM}{\rho^2}C\sqrt{T} - U_2\frac{2LM}{\rho^2}\frac{(t_2 - t_1 + 1)}{C\sqrt{T}}$$
$$\leq$$
$$L(t_2 - t_1 + 1) + D_3\frac{L^2}{\rho^2\eta} + D_2\eta(t_2 - t_1 + 1).$$

Observing that $\mathcal{E}^G_{t_1,t_2,\delta} = 2L\sqrt{2(t_2 - t_1 + 1)\ln\left(\frac{T^2}{\delta}\right)} \leq U_3 l_1\sqrt{t_2 - t_1 + 1}$ with $l_1 = \sqrt{\ln\left(\frac{T^2}{\delta}\right)}$ and $U_3 = 2L\sqrt{2}$ and rearranging the terms we obtain:

$$\begin{aligned} L(t_2 - t_1 + 1) \leq\ & U_3\frac{2LM}{\rho^2}l_1\sqrt{t_2 - t_1 + 1} \\ & + U_1\frac{2LM}{\rho^2}C\sqrt{T} \\ & + U_2\frac{2LM}{\rho^2}\frac{(t_2 - t_1 + 1)}{C\sqrt{T}} \\ & + D_2\eta(t_2 - t_1 + 1) \\ & + D_3\frac{1}{\eta}\frac{L^2}{\rho^2}. \end{aligned}$$

We will make use of the following lemma:

Lemma 5.2. *For $\eta \leq \frac{1}{mL}$ and $\frac{M}{\rho} > 4$ it holds:*

$$(t_2 - t_1 + 1) > \frac{M}{\rho^2 m\eta}.$$

Proof. By Lemma C.1 we have:

$$\sum_{t=t_1}^{t_2}(||\lambda_{t+1}||_1 - ||\lambda_t||_1) \leq \sum_{t=t_1}^{t_2} m\eta L,$$

which, since the sum in the LHS is telescopic, implies:

$$||\lambda_{t_2+1}||_1 - ||\lambda_{t_1}||_1 \leq (t_2 - t_1 + 1)m\eta L.$$

Also note that:

$$\frac{2LM}{\rho^2} - \frac{4L}{\rho} \leq ||\lambda_{t_2+1}||_1 - ||\lambda_{t_1}||_1.$$

Rearranging the terms, we obtain, for $\frac{M}{\rho} > 4$:

$$\frac{M}{\rho^2 m\eta} < \frac{2L(\frac{M}{\rho} - 2)}{\rho m\eta L} \leq (t_2 - t_1 + 1).$$

□

Applying Lemma 5.2 we show that the above leads to a contradiction for some choices of C, M and η, namely, we show that:

$$\begin{aligned} L(t_2 - t_1 + 1) > \; & U_3 \frac{2LM}{\rho^2} l_1 \sqrt{t_2 - t_1 + 1} && (1) \\ & + U_1 \frac{2LM}{\rho^2} C\sqrt{T} && (2) \\ & + U_2 \frac{2LM}{\rho^2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}} && (3) \\ & + D_2 \eta (t_2 - t_1 + 1) && (4) \\ & + D_3 \frac{1}{\eta} \frac{L^2}{\rho^2}. && (5) \end{aligned}$$

In the followings, we prove that each of the terms on the RHS is upper bounded by $\frac{1}{5} L(t_2 - t_1 + 1)$:

1. By trivial computations and applying Lemma 5.2:

$$\begin{aligned} \frac{1}{5} L(t_2 - t_1 + 1) &> U_3 \frac{2LM}{\rho^2} l_1 \sqrt{T} \geq U_3 \frac{2LM}{\rho^2} l_1 \sqrt{t_2 - t_1 + 1} \\ (t_2 - t_1 + 1) &> U_3 \frac{10M}{\rho^2} l_1 \sqrt{T} \\ (t_2 - t_1 + 1) &> \frac{M}{\rho^2 m\eta} \geq U_3 \frac{10M}{\rho^2} l_1 \sqrt{T} \\ \frac{1}{m\eta} &\geq 10 U_3 l_1 \sqrt{T}, \end{aligned}$$

   which is ensured by:

$$\boxed{\eta \leq \frac{1}{10 m U_3 l_1 \sqrt{T}}}$$

2. Then applying again Lemma 5.2:

$$\begin{aligned} \frac{1}{5} L(t_2 - t_1 + 1) &> U_1 \frac{2LM}{\rho^2} C\sqrt{T} \\ (t_2 - t_1 + 1) &> \frac{M}{\rho^2 m\eta} \geq 10 U_1 \frac{M}{\rho^2} C\sqrt{T}, \end{aligned}$$

   which is true for:

$$\boxed{\eta \leq \frac{1}{10 m U_1 C\sqrt{T}}}$$

3. We solve the third term with respect to C.

$$\frac{1}{5}L(t_2 - t_1 + 1) \geq U_2 \frac{2LM}{\rho^2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}},$$

which is ensured by:

$$\boxed{C \geq 10U_2 \frac{M}{\rho^2} \frac{1}{\sqrt{T}}}$$

4.

$$\frac{1}{5}L(t_2 - t_1 + 1) > D_2\eta(t_2 - t_1 + 1)$$
$$\frac{1}{5}L > D_2\eta,$$

which is ensured by:

$$\boxed{\eta < \frac{L}{5D_2}}$$

5. Applying Lemma 5.2, we solve the Inequality with respect to M:

$$\frac{1}{5}L(t_2 - t_1 + 1) > D_3 \frac{1}{\eta} \frac{L^2}{\rho^2}$$
$$(t_2 - t_1 + 1) > \frac{M}{\rho^2 m\eta} \geq 5D_3 \frac{1}{\eta} \frac{L}{\rho^2}$$
$$\frac{M}{m} \geq 5D_3 L,$$

from which:

$$\boxed{M \geq 5mD_3L}$$

We recall all the constants: $D_2 = \frac{mL^2}{2}$, $D_3 = 2$, $U_1 = 2L$, $U_2 = |X||A|$, $U_3 = 2L\sqrt{2}$.
We choose $M = 10mL$ and recall Condition 5.1:

$$\rho \geq T^{-\frac{1}{8}} L\sqrt{20m} \quad \Rightarrow \quad \frac{20mL^2}{\rho^2} \leq T^{\frac{1}{4}} \leq \sqrt{T}.$$

We now focus on the condition on C:

$$C \geq 10U_2 \frac{10mL}{\rho^2} \frac{1}{\sqrt{T}}$$
$$= 5\frac{U_2}{L} \frac{20mL^2}{\rho^2} \frac{1}{\sqrt{T}}$$

is thus always ensured by $C = 5\frac{U_2}{L}$. The conditions on η are satisfied if:

$$\eta \leq \min\left\{ \frac{L}{5D_2}, \frac{1}{10mU_1C\sqrt{T}}, \frac{1}{10mU_3l_1\sqrt{T}} \right\}.$$

Observe that:

$$\min\left\{\frac{L}{5D_2}, \frac{1}{10mU_1C\sqrt{T}}, \frac{1}{10mU_3l_1\sqrt{T}}\right\}$$
$$= \min\left\{\frac{1}{2.5mL}, \frac{1}{10mU_1\left(\frac{5U_2}{L}\right)\sqrt{T}}, \frac{1}{20\sqrt{2}mLl_1\sqrt{T}}\right\}.$$

If we plug in the value of l_1, leads to the choice:

$$\eta = \frac{1}{50m\max\left\{\frac{U_1U_2}{L},\ L\right\}\sqrt{T\ln\left(\frac{T^2}{\delta}\right)}}.$$

The remaining conditions $\frac{M}{\rho} > 4$, $\eta \leq \frac{1}{mL}$ are trivially satisfied. Summing the conditions $(1-5)$ proves the contradiction.
If we plug the values of U_1 and U_2 corresponding to UC-O-GDPS, we have $\max\left\{\frac{U_1U_2}{L},\ L\right\} = \max\left\{2|X||A|,\ L\right\} = 2|X||A|$ and thus obtain:

$$\eta = \frac{1}{100m|X||A|\sqrt{T\ln\left(\frac{T^2}{\delta}\right)}}.$$

This concludes the proof. □

5.4.1 Stochastic Constraints Setting

The peculiarity of this setting is that, at every episode $t \in [T]$ the constraint matrix G is sampled from a fixed distribution, namely $G_t \sim \mathcal{G}$. Instead, rewards r_t can be sampled from a fixed distribution $\mathcal{R}$ or chosen adversarially.

Azuma-Hoeffding Bounds Initially, we bound the error between the realizations of reward vectors and their corresponding mean values when the rewards are chosen stochastically.

Lemma 5.3. *If the rewards are stochastic, then, with probability at least* $1-\delta$*, it holds:*

$$\left|\sum_{t=1}^{T}\left(r_t - \overline{r}\right)^\top q^*\right| \leq \mathcal{E}_\delta^r,$$

where $\mathcal{E}_\delta^r := \frac{L}{\sqrt{2}}\sqrt{T\ln\left(\frac{2}{\delta}\right)}$.

Proof. Observe that:

$$\max_{t\in[t_1,\ldots,t_2]}\left|\left(r_t - \overline{r}\right)^\top q^*\right| \leq \max_{t\in[t_1,\ldots,t_2]}\underbrace{\|r_t - \overline{r}\|_\infty}_{\leq 1}\|q^*\|_1$$
$$\leq L,$$

where the second Inequality holds since since $q^*(x,a) \geq 0$. By the Azuma-Hoeffding

inequality for martingales we have that:

$$\mathbb{P}\left[\left|\sum_{t=t_1}^{t_2}\left(r_t-\overline{r}\right)^\top q^*\right| \geq \frac{L}{\sqrt{2}}\sqrt{T\ln\left(\frac{2}{\delta}\right)}\right] \leq \delta.$$

This concludes the proof. □

Now, we bound the error between the realizations of constraint violations and their corresponding mean values.

Lemma 5.4. *If the constraints are stochastic, given a sequence of occupancy measures $(q_t)_{t=1}^T$, then with probability at least $1-\delta$, for all $[t_1,\dots,t_2] \subseteq [1,\dots,T]$, it holds:*

$$\left|\sum_{t=t_1}^{t_2}\lambda_t^\top\left(G_t^\top-\overline{G}^\top\right)q_t\right| \leq \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2,\delta},$$

where we let $\mathcal{E}^G_{t_1,t_2,\delta} := 2L\sqrt{2(t_2-t_1+1)\ln\left(\frac{T^2}{\delta}\right)}$ and $\lambda_{t_1,t_2} := \max\{\|\lambda_t\|_1\}_{t=t_1}^{t_2}$.

Proof. Observe that:

$$\begin{aligned}\max_{t\in[t_1,\dots,t_2]}\left|\lambda_t^\top(G_t^\top-\overline{G}^\top)q_t\right| &\leq \max_{t\in[t_1,\dots,t_2]}\|\lambda_t\|_1 \underbrace{\left\|G_t^\top-\overline{G}^\top\right\|_\infty}_{\leq 2}\|q_t\|_1\\ &\leq \max_{t\in[t_1,\dots,t_2]} 2||\lambda_t||_1 L\\ &= 2\lambda_{t_1,t_2}L,\end{aligned}$$

where the second Inequality holds since $q_t(x,a) \geq 0$ and $\lambda_{t,i} \geq 0$. By the Azuma-Hoeffding inequality for martingales we have that:

$$\mathbb{P}\left[\left|\sum_{t=t_1}^{t_2}\lambda_t^\top(G_t^\top-\overline{G}^\top)q_t\right| \geq 2\lambda_{t_1,t_2}L\sqrt{2(t_2-t_1+1)\ln\left(\frac{2}{\delta}\frac{T^2}{2}\right)}\right] \leq 2\delta/T^2.$$

A Union Bound over all the t_1, t_2 such that $[t_1,\dots,t_2] \subseteq [1,\dots,T]$ concludes the proof. □

For the sake of notation, we use $\mathcal{E}^G_\delta$ in place of $\mathcal{E}^G_{1,T,\delta}$. Let us remark that $\mathcal{E}^r_\delta, \mathcal{E}^G_\delta \leq \widetilde{\mathcal{O}}(\sqrt{T})$.

Analysis when Condition 5.1 Holds We start by analyzing the case in which Condition 5.1 holds. By Theorem 5.2, we know that the maximum 1-norm of the dual vectors selected by `OGD` during the learning process is upper-bounded by the constant ζ. Since ζ essentially determines the range of the Lagrangian function, we can prove optimal regret and violation bounds of order $\widetilde{O}(\zeta\sqrt{T})$ for `PDGD-OPS`, as stated in the following theorem.

Theorem 5.3. *In the stochastic constraint setting, when Condition 5.1 holds, the cumulative regret and constraint violation incurred by `PDGD-OPS` are upper bounded as follows. If the rewards are adversarial, then with probability at least $1-4\delta$ Algorithm 5.1 provides:*

$$R_T \leq \zeta\mathcal{E}^G_\delta + \zeta\mathcal{E}^q_\delta + R^{\mathsf{D}}_T(\underline{0}) + R^{\mathsf{P}}_T(q^*),$$

$$V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q.$$

If the rewards are stochastic, then with probability at least $1 - 5\delta$ *Algorithm 5.1 provides:*

$$R_T \leq \mathcal{E}_\delta^r + \zeta\mathcal{E}_\delta^G + \zeta\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(q^*),$$
$$V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q.$$

In both cases:

$$R_T \leq \widetilde{\mathcal{O}}\left(\zeta\sqrt{T}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(\zeta\sqrt{T}\right).$$

Proof. Assume events $E_{q^\diamond}^G(\delta)$, $E_{q^*}^G(\delta)$, $E^\Delta(\delta)$ and $E^{\widehat{q}}(\delta)$ hold.
Recall that $\lambda_{1,T} \leq \zeta$ under the events $E^\Delta(\delta)$ and $E_{q^\diamond}^G(\delta)$ since Condition 5.1 holds (see proof of Theorem 5.2).
By Lemma C.5 we have:

$$\begin{aligned} V_T &\leq \widehat{V}_{T,i^*} + \mathcal{E}_\delta^q \\ &\leq \frac{1}{\eta}\lambda_{T+1,i^*} + \mathcal{E}_\delta^q \\ &\leq \frac{1}{\eta}||\lambda_{T+1}||_1 + \mathcal{E}_\delta^q \\ &\leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q, \end{aligned}$$

where the third Inequality holds for Lemma C.4. By the definition of regret of the primal:

$$\begin{aligned} \sum_{t=1}^T r_t^\top q_t &\geq \sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T \lambda_t^\top G_t^\top q^* + \sum_{t=1}^T \lambda_t^\top G_t^\top q_t - R_T^{\mathsf{P}}(q^*) \\ &\geq \sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T \lambda_t^\top G_t^\top q^* - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \qquad (5.6) \\ &\geq \sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T \lambda_t^\top \overline{G}^\top q^* - \lambda_{1,T}\mathcal{E}_\delta^G - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \qquad (5.7) \\ &\geq \sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T \sum_i \lambda_{t,i} \underbrace{(\overline{G})_i q^*}_{\leq 0} - \lambda_{1,T}\mathcal{E}_\delta^G - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*) \qquad (5.8) \\ &\geq \sum_{t=1}^T r_t^\top q^* - \zeta\mathcal{E}_\delta^G - \zeta\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*), \end{aligned}$$

where Inequality (5.6) holds for Lemma C.3, and Inequality (5.7) holds under Event $E_{q^*}^G(\delta)$. We now focus on the case in which the rewards are adversarial. We have:

$$\sum_{t=1}^T r_t^\top q^* = T \cdot \overline{r}^\top q^* = T \cdot \mathrm{OPT}_{\overline{r},\overline{G}},$$

and thus we obtain the stated bound:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - \zeta\mathcal{E}_\delta^G - \zeta\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*).$$

By Union Bound on $E_{q^\diamond}^G(\delta)$, $E_{q^*}^G(\delta)$ and $E^{\Delta,\widehat{q}}(\delta)$, the result holds with probability at least $1 - 4\delta$.

For the stochastic rewards case, we require also event $E_{q^*}^r(\delta)$ to hold. Thus,

$$\sum_{t=1}^{T} r_t^\top q^* \geq \sum_{t=1}^{T} \overline{r}^\top q^* - \mathcal{E}_\delta^r = T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r,$$

and thus we obtain the stated bound:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r - \zeta\mathcal{E}_\delta^G - \zeta\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*).$$

By Union Bound on $E_{q^\diamond}^G(\delta)$, $E_{q^*}^G(\delta)$, $E^{\Delta,\widehat{q}}(\delta)$ and $E_{q^*}^r(\delta)$, the result holds with probability at least $1 - 5\delta$.

Observe that under $E^{\Delta,\widehat{q}}(\delta)$ it holds:

$$R_T^{\mathsf{P}}(q^*) \leq \widetilde{\mathcal{O}}\left((1 + \lambda_{1,T})\sqrt{T}\right) = \widetilde{\mathcal{O}}\left(\zeta\sqrt{T}\right),$$

and

$$R_T^{\mathsf{D}}(\underline{0}) \leq \frac{mL^2}{2} \frac{1}{100m|X||A|\sqrt{\ln\left(\frac{T^2}{\delta}\right)}}\sqrt{T} \leq \mathcal{O}\left(\sqrt{T}\right).$$

This concludes the proof. □

Notice that, if Condition 5.1 does not hold, the bounds stated in Theorem 5.3 can become of order $\widetilde{O}(T^{3/4})$ or even linear. We conclude the analysis of the stochastic constraint setting when Condition 5.1 holds with the following remark.

In Theorem 5.3, the regret bound when the rewards are adversarial is better than the one when the rewards are chosen stochastically. This result may seem counter-intuitive as the adversarial setting is the hardest setting a learner might face. Informally, this is due to the different definition of the optimization baseline used in the stochastic and adversarial settings.

Analysis when Condition 5.1 Does Not Hold We focus on the case in which Condition 5.1 does not hold. As previously observed, in this case the regret and violation bounds given in Theorem 5.3 are not meaningful anymore, as they could become linear in T (in fact, this is exactly the case when $\rho \propto T^{-\frac{1}{4}}$). Nevertheless, by constraining the dual player to the decision space $\mathcal{D} = [0, T^{1/4}]^m$, we are able to prove worst-case regret and violation bounds of the order of $\widetilde{\mathcal{O}}\left(T^{3/4}\right)$. This result is formalized in the following theorem.

Theorem 5.4. *In the stochastic constraint setting, when Condition 5.1 does not hold, the cumulative regret and constraint violations incurred by* `PDGD-OPS` *are upper bounded as follows. If the rewards are adversarial, then with probability at least* $1-4\delta$ *Algorithm 5.1 provides:*

$$R_T \leq mT^{\frac{1}{4}}\mathcal{E}_\delta^G + mT^{\frac{1}{4}}\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(q^*),$$

$$V_T \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} + \mathcal{E}_\delta^q.$$

If the rewards are stochastic, then with probability at least $1-5\delta$ *Algorithm 5.1 provides:*

$$R_T \leq \mathcal{E}_\delta^r + mT^{\frac{1}{4}}\mathcal{E}_\delta^G + mT^{\frac{1}{4}}\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(q^*),$$

$$V_T \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} + \mathcal{E}_\delta^q.$$

In both cases, it holds:

$$R_T \leq \widetilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right).$$

Proof. Assume events $E^{\Delta}(\delta)$, $E^{\widehat{q}}(\delta)$, $E^G_{q^*}(\delta)$, $E^G_{q^\diamond}(\delta)$ hold. We avoid the computations and restart from Equation (5.8), since the previous part of the proofs are identical:

$$\sum_{t=1}^T r_t^\top q_t \geq \sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T \sum_i \lambda_{t,i} \underbrace{(\overline{G})_i q^*}_{\leq 0} - \lambda_{1,T}\mathcal{E}_\delta^G - \lambda_{1,T}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*)$$

$$\geq \sum_{t=1}^T r_t^\top q^* - mT^{\frac{1}{4}}\mathcal{E}_\delta^G - mT^{\frac{1}{4}}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*).$$

By the same reasoning as in the proof of Theorem 5.3, we obtain that if the rewards are adversarial then,

$$\sum_{t=1}^T r_t^\top q_t \geq T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - mT^{\frac{1}{4}}\mathcal{E}_\delta^G - mT^{\frac{1}{4}}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*),$$

with probability at least $1-4\delta$ by union bound on $E^{\Delta,\widehat{q}}(\delta)$, $E^G_{q^*}(\delta)$ and $E^G_{q^\diamond}(\delta)$, while if the rewards are stochastic, under the event $E^r_{q^*}(\delta)$ we have that:

$$\sum_{t=1}^T r_t^\top q_t \geq T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r - mT^{\frac{1}{4}}\mathcal{E}_\delta^G - mT^{\frac{1}{4}}\mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(q^*),$$

with probability at least $1-5\delta$ by Union Bound on $E^{\Delta,\widehat{q}}(\delta)$, $E^G_{q^*}(\delta)$, $E^G_{q^\diamond}(\delta)$ and $E^r_{q^*}(\delta)$.
Observe that:

$$R_T^{\mathsf{P}}(q^*) \leq \widetilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right),$$

and

$$R_T^{\mathsf{D}}(\underline{0}) = \frac{mL^2}{2}\eta T \leq \widetilde{\mathcal{O}}\left(\sqrt{T}\right).$$

In order to bound the violation, we apply Lemma C.6, thus:

$$V_T \leq \widehat{V}_{T,i^*} + \mathcal{E}_\delta^q \leq (2+2L)\frac{1}{\eta}T^{\frac{1}{4}} + \mathcal{E}_\delta^q.$$

This concludes the proof. □

5.4.2 Adversarial Constraints Setting

We recall that in this setting, at every episode $t \in [T]$, the constraint matrix G_t is chosen adversarially. Instead, rewards r_t can be sampled from a fixed distribution $\mathcal{R}$ or chosen adversarially. This case corresponds to the hardest scenario the learner can face. As stated in Section 2.5, the treatment of this case requires a definition of ρ stronger than that used in the stochastic constraint setting. Thanks to such a redefinition, it is possible to achieve guarantees on the cumulative constraint violation of the same order of those attainable in the stochastic setting, while obtaining at least a constant fraction of the optimal reward, that is, sublinear α-regret, where $\alpha = \rho/L+\rho$. Such a result can be achieved when Condition 5.1 holds. Notice that both sublinear cumulative regret and sublinear cumulative constraint violation cannot be achieved in our setting, as shown by Mannor et al. (2009).

The following theorem summarizes our result for the adversarial constraint setting.

Theorem 5.5. *In the adversarial constraint setting, when Condition 5.1 holds, the cumulative regret and constraint violations incurred by* `PDGD-OPS` *are upper bounded as follows. If the rewards are adversarial, then with probability at least* $1-2\delta$ *Algorithm 5.1 provides:*

$$R_T \leq \frac{L}{L+\rho}T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} + \zeta\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(\tilde{q}),$$

$$V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q.$$

If the rewards are stochastic, then with probability at least $1-3\delta$ *Algorithm 5.1 provides:*

$$R_T \leq \frac{L}{L+\rho}T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} + \mathcal{E}_\delta^r + \zeta\mathcal{E}_\delta^q + R_T^{\mathsf{D}}(\underline{0}) + R_T^{\mathsf{P}}(\tilde{q}),$$

$$V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q.$$

In both cases, it holds:

$$\sum_{t=1}^{T} r_t^\top q_t \geq \Omega\left(\frac{\rho}{L+\rho}T \cdot \mathrm{OPT}_{\overline{r},\overline{G}}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(\zeta\sqrt{T}\right).$$

Proof. Assume events $E^{\Delta}(\delta)$ and $E^{\widehat{q}}(\delta)$ hold.
Recall that $\lambda_{1,T} \leq \zeta$ under the event $E^{\Delta}(\delta)$ since Condition 5.1 holds (see the proof of Theorem 5.2).
Following the same steps of the proof of Theorem 5.3, we obtain:

$$V_T \leq \frac{1}{\eta}\zeta + \mathcal{E}_\delta^q.$$

Let $\tilde{q} = \frac{\rho}{L+\rho} q^* + \frac{L}{L+\rho} q^\diamond$, observe that it holds for all t and for all i:

$$[G_t^\top \tilde{q}]_i = \frac{\rho}{L+\rho} \underbrace{[G_t^\top q^*]_i}_{\leq L} + \frac{L}{L+\rho} \underbrace{[G_t^\top q^\diamond]_i}_{\leq -\rho} \leq 0,$$

$$r_t^\top \tilde{q} = \frac{\rho}{L+\rho} r_t^\top q^* + \frac{L}{L+\rho} r_t^\top q^\diamond \geq \frac{\rho}{L+\rho} r_t^\top q^*.$$

By the definition of regret of the primal:

$$\begin{aligned}
\sum_{t=1}^T r_t^\top q_t &\geq \sum_{t=1}^T r_t^\top \tilde{q} - \sum_{t=1}^T \lambda_t^\top G_t^\top \tilde{q} + \sum_{t=1}^T \lambda_t^\top G_t^\top q_t - R_T^{\mathsf{P}}(\tilde{q}) \\
&\geq \frac{\rho}{L+\rho} \sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T \sum_i \lambda_{t,i} \underbrace{[G_t^\top \tilde{q}]_i}_{\leq 0} + \sum_{t=1}^T \lambda_t^\top G_t^\top q_t - R_T^{\mathsf{P}}(\tilde{q}) \\
&\geq \frac{\rho}{L+\rho} \sum_{t=1}^T r_t^\top q^* - \lambda_{1,T} \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}) \\
&\geq \frac{\rho}{L+\rho} \sum_{t=1}^T r_t^\top q^* - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}),
\end{aligned}$$

where the third Inequality holds for Lemma C.3.
By the same reasoning as in the proof of Theorem 5.3, we obtain that if the rewards are adversarial it holds:

$$\begin{aligned}
\sum_{t=1}^T r_t^\top q_t &\geq \frac{\rho}{L+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}) \\
&= T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \frac{L}{L+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}),
\end{aligned}$$

with probability at least $1 - 2\delta$, since we are conditioning on $E^{\Delta,\tilde{q}}(\delta)$.
If the rewards are stochastic, requiring also event $E_{q^*}^r(\delta)$ to hold we obtain:

$$\frac{\rho}{L+\rho} \sum_{t=1}^T r_t^\top q^* \geq \frac{\rho}{L+\rho} \sum_{t=1}^T \overline{r}^\top q^* - \frac{\rho}{L+\rho} \mathcal{E}_\delta^r \geq \frac{\rho}{L+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r.$$

Thus,

$$\sum_{t=1}^T r_t^\top q_t \geq T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \frac{L}{L+\rho} T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_\delta^r - \zeta \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}) - R_T^{\mathsf{P}}(\tilde{q}),$$

with probability at least $1 - 3\delta$. Finally observe that, under Condition 5.1 and event $E^{\Delta,\tilde{q}}(\delta)$, it holds:

$$R_T^{\mathsf{P}}(\tilde{q}) \leq \widetilde{\mathcal{O}}\left((1 + \lambda_{1,T})\sqrt{T}\right) \leq \widetilde{\mathcal{O}}\left(\zeta\sqrt{T}\right)$$

and

$$R_T^{\mathsf{D}}(\underline{0}) \leq \frac{mL^2}{2} \frac{1}{100m|X||A|\sqrt{\ln\left(\frac{T^2}{\delta}\right)}} \sqrt{T} \leq \mathcal{O}\left(\sqrt{T}\right).$$

This concludes the proof. □

CHAPTER 6

A Best-of-Both-Worlds Algorithm for Bandit Feedback

In this chapter, we generalize the results provided in Chapter 5 to the *bandit feedback* case. Specifically, we study *online learning* in episodic CMDPs where both the rewards and the constraints can be either stochastic or adversarial, and we provide the first *best-of-both-worlds* algorithm for online learning in episodic CMDPs with *bandit feedback*. This means that, after each episode, the algorithm only needs to observe the realized rewards and costs along the trajectory traversed during that episode, as it is the case in most of the real-world applications. Moreover, our algorithm is based on a primal-dual policy optimization method, and, thus, it is arguably much more efficient than the one provided in Chapter 5, since it does *not* require solving convex programs.

When the costs are *stochastic*, our algorithm attains $\widetilde{O}(\sqrt{T})$ regret and constraint violation, while, when they are *adversarial*, it achieves $\widetilde{O}(\sqrt{T})$ violation and a fraction of the optimal reward. These results match those of the full-feedback algorithm (Algorithm 5.1) and are provably tight. We also analyze the performances of our algorithm with respect to a parameter ρ measuring by "how much" Slater's condition is satisfied. Specifically, if ρ is arbitrarily small, our algorithm can still guarantee $\widetilde{\mathcal{O}}(T^{3/4})$ regret and violation in the stochastic setting.

Crucially, similarly to Algorithm 5.1, ours *does not require* any knowledge of the Slater's parameter ρ. In order to obtain this result, we show that the Lagrangian multipliers are automatically bounded during the learning dynamics, by employing the *no-interval-regret* property of our primal and dual regret minimizers. Indeed, we develop the first algorithm for unconstrained MDPs with no-interval-regret, under *bandit* feedback. We believe that this result may also be of independent interest.

Finally, we show that our algorithm may achieve sublinear regret and violation in the adversarial setting, by using a *weaker* baseline that has to satisfy the constraints at ev-

ery round. Specifically, when ρ is large enough our algorithm attains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation, while it still achieves $\widetilde{\mathcal{O}}(T^{3/4})$ regret and violation when ρ is arbitrarily small.

6.1 Setting and Additional Notation

The setting and the notation follows exactly the ones of Chapter 5, except for the feedback, which is *bandit* in this case. Similarly, we will make use of the following Slater's like condition, which plays a central role when proving algorithm theoretical guarantees.

Condition 6.1. *It holds that* $\rho \geq T^{-\frac{1}{8}} L\sqrt{112m}$.

We finally refer to Section D.1 for the dictionary of the definition of different quantities which will be employed in the rest of the chapter.

6.2 A Policy Optimization Primal-Dual Approach

In this section, we provide the description of our algorithm. We resort to a primal-dual formulation of the CMDP problem, and we employ different regret minimizers to optimize over the primal space (namely, the policy space) and the dual one (that is, the Lagrangian variables space). Furthermore, our primal algorithm is based on a policy optimization approach. Thus, the learning update is *not* performed over the occupancy measure space, but state-by-state along the MDP structure. This allows us to avoid solving a convex program at each episode (as it is the case of Algorithm 5.1). As concerns the dual, we employ online gradient descent (`OGD`). We remark that our algorithm does *not* require any knowledge of the Slater's parameter ρ. Indeed, as we further discuss in the rest of this work, we can show that the Lagrangian multipliers are automatically bounded given specific no-regret properties of the primal and dual regret minimizers.

6.2.1 Meta-Algorithm

In Algorithm 6.1, we provide the pseudocode of *primal-dual bandit policy search* (`PDB-PS`). Algorithm 6.1 initializes the policy uniformly over the space (see Line 1). Moreover, the Lagrangian variables are initialized as the zero vector, the loss scaling factor to 1, the loss range to 2, and, finally, the dual space is instantiated as $\left[0, T^{1/4}\right]^m$ (see Line 3). We underline that we force the dual space to be bounded in $\left[0, T^{1/4}\right]^m$ only to deal with degenerate cases where Condition 6.1 does *not* hold. When Condition 6.1 holds, our algorithm guarantees that the Lagrangian variables are automatically bounded during learning. Furthermore, the algorithm keeps track of the maximum loss range observed by the primal algorithm Ξ_t, up to episode $t \in [T]$, since the primal regret minimizer needs to dynamically update its belief on the loss range, in order to attain optimal regret bounds. The algorithm plays policy π_t and observes the *bandit* feedback as depicted in Algorithm 2.1 (see Line 5). Given the observed feedback, `PDB-PS` builds a re-scaled Lagrangian loss for each layer k as:

$$\ell_t(x_k, a_k) := \Gamma_t + \sum_{i=1}^{m} \lambda_{t,i} g_{t,i}(x_k, a_k) - r_t(x_k, a_k). \tag{6.1}$$

Notice that the loss built in Equation (6.1) can been seen as the Lagrangian suffered by π_t for state-action pair (x, a), scaled by Γ_t to guarantee that the losses are always positive

Algorithm 6.1 Primal-Dual Bandit Policy Search (`PDB-PS`)

Require: State space X, action space A, number of episodes T, confidence parameter $\delta \in (0,1)$

1: $\pi_1(a|x) \leftarrow \frac{1}{|A|} \quad \forall (x,a) \in X \times A$
2: $\lambda_1 \leftarrow \underline{0}, \Gamma_1 \leftarrow 1, \Xi_1 \leftarrow 2$
3: $\mathcal{K} \leftarrow \left[0, T^{1/4}\right]^m, \eta \leftarrow \frac{1}{D \ln\left(|A||X|^2 T^2/\delta\right)\sqrt{T}}$
4: **for** $t = 1, \dots, T$ **do**
5: Play policy π_t, observe trajectory $\{(x_k, a_k)\}_{k=0}^{L-1}$, rewards $\{r_t(x_k,a_k)\}_{k=0}^{L-1}$ and constraint violations $\{g_{t,i}(x_k,a_k)\}_{k=0}^{L-1}$ for all $i \in [m]$
6: **for** $k = 0, \dots, L-1$ **do**
7: $\ell_t(x_k,a_k) \leftarrow \Gamma_t + \sum_{i=1}^m \lambda_{t,i} g_{t,i}(x_k,a_k) - r_t(x_k,a_k)$
8: **end for**
9: $\pi_{t+1} \leftarrow$ Call `FS-PODB.UPDATE`$(\{(x_k,a_k)\}_{k=0}^{L-1}, \{\ell_t(x_k,a_k)\}_{k=0}^{L-1}, \Xi_t)$
10: $\lambda_{t+1} \leftarrow \Pi_{\mathcal{K}} \left[\lambda_t + \eta \sum_{k=0}^{L-1} G_t(x_k,a_k)\right]$
11: $\Gamma_{t+1} \leftarrow 1 + \|\lambda_{t+1}\|_1$
12: $\Xi_{t+1} \leftarrow \max\{\Xi_t, 2\Gamma_t\}$
13: **end for**

(see Line 7). This loss is properly built to feed the primal policy optimization procedure. Moreover, we underline that the feedback given to the primal algorithm encompasses the trajectory and the maximum loss range observed, besides the loss built in Equation (6.1). Policy π_{t+1} is returned by the primal algorithm (Line 9). We refer the reader to the next section for further discussion on the primal optimization algorithm. Algorithm 6.1 updates the Lagrangian multipliers using an online gradient descent update with loss $-\sum_{k=0}^{L-1} G_t(x_k,a_k)$ in the bounded dual space $[0, T^{1/4}]^m$:

$$\lambda_{t+1} \leftarrow \Pi_{\mathcal{K}} \left[\lambda_t + \eta \sum_{k=0}^{L-1} G_t(x_k,a_k)\right],$$

where $\Pi_{\mathcal{K}}$ is the euclidean projection over the space $\mathcal{K}$ and $G_t(x_k,a_k)$ is the m-dimensional vector composed by the violations of any constraint for the state-action pair (x_k,a_k) (Line 10). Thus, the current loss scaling factor is computed as $\Gamma_{t+1} \leftarrow 1 + \|\lambda_{t+1}\|_1$ (Line 11). Finally, the maximum observed loss range Ξ_{t+1} is updated as follows, $\Xi_{t+1} \leftarrow \max\{\Xi_t, 2\Gamma_{t+1}\}$, since the range of losses observed by the primal depends on the Lagrangian multipliers (Line 12).

6.2.2 Primal Regret Minimizer

In Algorithm 6.2, we provide the pseudocode of *fixed share policy optimization with dilated bonus* (`FS-PODB.UPDATE`), namely, the update performed by the primal regret minimizer employed by Algorithm 6.1. Algorithm 6.2 builds on top of the state-of-the-art policy optimization algorithm for adversarial MDPs (see (Luo et al., 2021)), equipping it with a fixed share update (Cesa-Bianchi et al., 2012). This modification allows us to achieve the no-interval regret property, which, to the best of our knowledge, has never been shown for adversarial MDPs with bandit feedback. Thus, we believe that the theoretical guarantees of Algorithm 6.2 are of independent interest.

Specifically, Algorithm 6.2 requires in input the trajectory traversed during the learner-environment interaction, the incurred loss functions, and the maximum loss range observed

for any $t \in [T]$.[1] During the first episode, the algorithm initializes the estimated transitions space as the set of all possible transition functions (Line 2). Thus, at each episode the algorithm defines a dynamic learning rate $\eta_t \propto \frac{1}{\sqrt{T}\Xi_t}$ (Line 4), where Ξ_t is the upper bound on the range of the loss functions up to t. This is done to control the different scales of the loss, due to the Lagrangian multipliers choice of the dual algorithm. Then, Algorithm 6.2 builds an *optimistic* estimator of the state-action value function as:

$$\widehat{Q}_t(x,a) \coloneqq \frac{L_{t,k}}{\overline{q}_t(x,a)+\gamma}\mathbb{I}_t(x,a),$$

where $\mathbb{I}_t(x,a) \coloneqq \mathbb{I}\{x_{t,k} = x, a_{t,k} = a\}$ and $L_{t,k} \coloneqq \sum_{j=k}^{L-1} \ell_t(x_j,a_j)$ is the loss incurred by the algorithm at episode t starting from layer k. Indeed, since $\overline{q}_t(x,a) \coloneqq \max_{\widehat{P}\in\mathcal{P}_t} q^{\widehat{P},\pi_t}(x,a)$,[2] and γ is a positive quantity, $\widehat{Q}_t(x,a)$ results in an optimistic estimator of the state-action value function (Line 5). The optimistic estimator is employed to control the variance of the loss estimation and, thus, in order to achieve high-probability results. Finally, notice that the state-action value function (as the estimated one) is commonly used in policy optimization as it allows to optimize efficiently state-by-state. In addition to the estimated state-action value function, Algorithm 6.2 defines a dilated bonus similar to the one introduced by Luo et al. (2021), which is then incorporated in the final objective of the optimization update. The bonus is defined as:

$$B_t(x,a) \coloneqq b_t(x) + \left(1+\frac{1}{L}\right)\max_{\widehat{P}\in\mathcal{P}_t} \mathbb{E}_{x'\sim\widehat{P}(\cdot|x,a)}\mathbb{E}_{a\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right],$$

where the term $b_t(x)$ depends on the uncertainty on the transitions estimation and the range of the losses, while the term $\left(1+\frac{1}{L}\right)$ attributes more weight to the deeper layers, so as to incentivize exploration (Line 6). The weights associated to any action are computed employing the so called fixed share update (Cesa-Bianchi et al., 2012); specifically, the weights are computed as the convex combination between the uniform weight and the solution to optimization step $\propto w_t(a|x)e^{-\eta_t\left(\widehat{Q}_t(x,a)-B_t(x,a)\right)}$. The policy is simply computed as a normalization between weights (see Line 7). Notice that the convex combination mentioned above is crucial to bound the regret for each interval (that is, to attain the no-interval regret property). Indeed, it guarantees a lower bound for the value taken by the policy in each available action at each episode, and, thus, for all intervals $[t_1,\dots,t_2] \subset [T]$, it allows to find a nice upper bound for the Bregman divergence $D_\psi(\pi(\cdot|a);\pi_{t_1}(\cdot|a))$, for all policies π. Finally, the estimation of the transitions is updated given the trajectory traversed in the MDP (Line 8). This estimation is standard in the literature. Thus, we refer to (Rosenberg and Mansour, 2019b) for further discussion on the use of counters and epochs to estimate a superset of the transition space $\mathcal{P}_t$.

6.2.3 No-Interval Regret Property

When the Slater's parameter ρ is known, the only necessary requirement for the primal and the dual regret minimizers is to be no-regret. Thus, it is sufficient to bound the Lagrangian

[1] While the input of Algorithm 6.2 may seem different from the standard bandit feedback received in adversarial MDPs, this is *not* the case. Indeed, it is sufficient to set $\Xi_t = 1$ for all $t \in [T]$ to achieve the same guarantees attained by Algorithm 6.2, in the Lagrangian formulation of CMDPs, in standard adversarial MDPs.

[2] As shown in (Jin et al., 2020a), $\overline{q}_t(x,a)$ can be computed efficiently by means of dynamic programming.

Algorithm 6.2 Fixed Share Policy Optimization with Dilated Bonus (`FS-PODB.UPDATE`)

Require: Observed trajectory $\{(x_k, a_k)\}_{k=0}^{L-1}$, observed losses $\{\ell_t(x_k, a_k)\}_{k=0}^{L-1}$, loss range upper bound Ξ_t

1: **if** $t = 1$ **then**

2: $\quad \mathcal{P}_1 \leftarrow$ set of all possible transitions

3: **end if**

4: $\eta_t \leftarrow \frac{1}{2L\Xi_t C\sqrt{T}}$, $\gamma \leftarrow \frac{1}{C\sqrt{T}}$, $\sigma \leftarrow \frac{1}{T}$

5: For all $K = 0, \dots, L-1$ and $(x, a) \in X_h \times A$:

$$L_{t,k} \leftarrow \sum_{j=k}^{L-1} \ell_t(x_j, a_j)$$

$$\widehat{Q}_t(x, a) \leftarrow \frac{L_{t,k}}{\overline{q}_t(x, a) + \gamma} \mathbb{I}_t(x, a),$$

where we let $\overline{q}_t(x, a) \coloneqq \max_{\widehat{P} \in \mathcal{P}_t} q^{\widehat{P}, \pi_t}(x, a)$ and $\mathbb{I}_t(x, a) \coloneqq \mathbb{I}\{x_{t,k} = x, a_{t,k} = a\}$

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

$$b_t(x) \leftarrow \mathop{\mathbb{E}}_{a \sim \pi_t(\cdot|x)} \left[\frac{3\gamma L\Xi_t + L\Xi_t \left(\overline{q}_t(x, a) - \underline{q}_t(x, a) \right)}{\overline{q}_t(x, a) + \gamma} \right]$$

$$B_t(x, a) \leftarrow b_t(x) + \left(1 + \frac{1}{L}\right) \max_{\widehat{P} \in \mathcal{P}_t} \mathbb{E}_{x' \sim \widehat{P}(\cdot|x,a)} \mathbb{E}_{a \sim \pi_t(\cdot|x')} \left[B_t(x', a') \right]$$

where we let $\underline{q}_t(x, a) \coloneqq \min_{\widehat{P} \in \mathcal{P}_t} q^{\widehat{P}, \pi_t}(x, a)$, and $B_t(x_L, a) \coloneqq 0$ for all $a \in A$

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

$$w_{t+1}(a|x) \leftarrow (1 - \sigma) w_t(a|x) e^{-\eta_t \left(\widehat{Q}_t(x,a) - B_t(x,a) \right)} + \frac{\sigma}{|A|} \sum_{a' \in A} w_t(a'|x) e^{-\eta_t \left(\widehat{Q}_t(x,a') - B_t(x,a') \right)}$$

$$\pi_{t+1}(a|x) \leftarrow \frac{w_{t+1}(a|x)}{\sum_{a' \in A} w_{t+1}(a'|x)}$$

8: $\mathcal{P}_{t+1} \leftarrow$`TRANSITION.UPDATE`$(\{(x_k, a_k)\}_{k=0}^{L-1})$

space so that $\|\lambda\|_1 \leq \mathcal{O}(L/\rho)$ to attain sublinear regret and violation. Nevertheless, knowing ρ is generally *not* possible in real-world scenarios. In order to relax the assumption on the knowledge of ρ, we require our primal and dual regret minimizers to have the no-interval regret property, as defined in Definition 5.2, but omitting the dependency on q for convenience.

When *full feedback* is available, as for the dual algorithm, it is sufficient to employ `OGD`-like updates to attain the desired result. This is *not* the case when the feedback is *bandit*. Nevertheless, employing a policy optimization procedure which can be naturally extended to incorporate a fixed share update, we build the first algorithm for adversarial MDPs with no-interval-regret. We state the result in the following theorem.

Theorem 6.1. *For any $\delta \in (0, 1)$, with probability at least $1 - 8\delta$, Algorithm* `FS-PODB` *attains:*

$$R_{t_1,t_2} \leq \widetilde{\mathcal{O}} \left(\Xi_{t_1,t_2} \sqrt{T} + \Xi_{t_1,t_2} \frac{t_2 - t_1}{\sqrt{T}} \right),$$

where the regret can be computed with respect to any policy function $\pi \in \Pi$.

Proof. By means of Theorem D.1 and by Lemma D.7 we have that with probability at least $1 - 4\delta$:

$$R^P_{t_1,t_2} \leq \gamma \frac{L\Xi_{t_1,t_2}}{2} \ln\left(\frac{LT^2}{\delta}\right) + \frac{6L^2\Xi_{t_1,t_2}}{\gamma} \ln\left(\frac{L|A|T^2}{\delta}\right) + 3\sum_{t=t_1}^{t_2} \widehat{V}^{\pi_t}(x_0; b_t),$$

where

$$\widehat{V}^{\pi_t}(x_0; b_t) = \sum_{x,a} q^{\widehat{P}_t,\pi_t}(x,a) \left(\frac{L\Xi_t(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + 3L\Xi_t\gamma}{\overline{q}_t(x,a) + \gamma}\right).$$

We can bound $\sum_{t=t_1}^{t_2} \widehat{V}^{\pi_t}(x_0; b_t)$, with probability at least $1 - 4\delta$, as:

$$\begin{aligned}
&\sum_{t=t_1}^{t_2} \widehat{V}^{\pi_t}(x_0; b_t) \\
&= \sum_{t=t_1}^{t_2} \sum_{x,a} q^{\widehat{P}_t,\pi_t}(x,a) \left(\frac{L\Xi_t(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + 3L\Xi_t\gamma}{\overline{q}_t(x,a) + \gamma}\right) \\
&\leq \sum_{t=t_1}^{t_2} \sum_{x,a} L\Xi_t \left((\overline{q}_t(x,a) - \underline{q}_t(x,a)) + 3\gamma\right) \\
&\leq \sum_{t=t_1}^{t_2} \sum_{x,a} L\Xi_t(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + 3\Xi_{t_1,t_2}\gamma L(t_2 - t_1 + 1)|X||A| \\
&\leq 4L^2\Xi_{t_1,t_2}|X|^2\sqrt{2T\ln\left(\frac{L|X|}{\delta}\right)} + 6\Xi_{t_1,t_2}L^2|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)} \\
&\qquad + 3\Xi_{t_1,t_2}\gamma L|X||A|(t_2 - t_1 + 1),
\end{aligned}$$

where the second inequality holds under the event $q^{\widehat{P}_t,\pi_t}(x,a) \leq \overline{q}_t(x,a)$ for all $(x,a) \in X \times A$ and for all $t \in [T]$ and the last inequality uses Lemma D.10. Thus, with probability at least $1 - 8\delta$, it holds:

$$\begin{aligned}
R^P_{t_1,t_2} &\leq \gamma \frac{L\Xi_{t_1,t_2}}{2} \ln\left(\frac{LT^2}{\delta}\right) + \frac{6L^2\Xi_{t_1,t_2}}{\gamma} \ln\left(\frac{L|A|T^2}{\delta}\right) \\
&\quad + 30\Xi_{t_1,t_2}L^2|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)} \\
&\quad + 9\Xi_{t_1,t_2}\gamma L|X||A|(t_2 - t_1 + 1) \\
&\leq \frac{L\Xi_{t_1,t_2}}{2C\sqrt{T}} \ln\left(\frac{LT^2}{\delta}\right) + 6L^2\Xi_{t_1,t_2}C\sqrt{T}\ln\left(\frac{L|A|T^2}{\delta}\right) \\
&\quad + 30\Xi_{t_1,t_2}L^2|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}
\end{aligned}$$

$$
\begin{aligned}
&+ 9L|X||A|\Xi_{t_1,t_2}\frac{(t_2-t_1+1)}{C\sqrt{T}} \\
&= U_1\Xi_{t_1,t_2}C\sqrt{T} + U_2\Xi_{t_1,t_2}\frac{(t_2-t_1+1)}{C\sqrt{T}} \\
&\quad + U_3\Xi_{t_1,t_2}\frac{1}{C\sqrt{T}} + U_4\Xi_{t_1,t_2}\sqrt{T} \\
&= \mathcal{E}^P_{t_1,t_2},
\end{aligned}
$$

which concludes the proof. □

As it is standard for online learning algorithms, R_{t_1,t_2} scales as the loss range, as shown by the dependence on Ξ_{t_1,t_2}, that is, the maximum possible range of losses in the interval.

6.2.4 Bound on the Lagrangian Multipliers Dynamics

Next, we show that, given the no-interval regret property of the primal and the dual regret minimizers, it is possible to show that the Lagrangian multipliers are automatically bounded during learning. Notice that this bound is necessary since any adversarial regret minimizer needs the loss to be bounded to achieve the no-regret property. Thus, since the rewards $\{r_t\}_{t=1}^T$ and the constraints $\{G_t\}_{t=1}^T$ are assumed to be bounded for all episodes, the problem of bounding the loss suffered by the primal algorithm becomes the problem of bounding the Lagrangian multipliers $\{\lambda_t\}_{t=1}^T$.

Theorem 6.2. *Under Condition 6.1, for any $\delta \in (0,1)$, with probability at least $1 - 11\delta$, it holds:*

$$\|\lambda_t\|_1 \leq \Lambda \quad \forall t \in [T+1],$$

where $\Lambda = \frac{112mL^2}{\rho^2}$.

The general idea behind the proof is to compare, for every interval $[t_1, \dots, t_2] \subset [T]$, the upper bound to $-\sum_{t=t_1}^{t_2} \ell_t^{\mathcal{L},\top} q_t$ obtained through the regret of the dual algorithm with the lower bound to the same quantity obtained through the primal interval regret, where we define the non-scaled Lagrangian loss $\ell_t^{\mathcal{L}}$ as the vector composed by

$$\ell_t^{\mathcal{L}}(x,a) \coloneqq \sum_{i=1}^m \lambda_{t,i} g_{t,i}(x,a) - r_t(x,a)$$

for all $(x,a) \in X \times A$ and for all $t \in [T]$. The resulting inequality leads, by contradiction, to the desired bound. In this sense, a fundamental requirement for the proof is that the regret guarantees for both the primal and the dual algorithm hold for all subsets of episodes.

Proof. Let $M > 1$ be a constant. By absurd suppose $\exists t_2 \in [T]$ s.t.

$$\forall t \leq t_2 \quad \|\lambda_t\|_1 \leq \frac{2LM}{\rho^2} \quad \wedge \quad \|\lambda_{t_2+1}\|_1 > \frac{2LM}{\rho^2} \tag{6.2}$$

and let $t_1 < t_2$ s.t.

$$\|\lambda_{t_1-1}\|_1 \leq \frac{2L}{\rho} \quad \wedge \quad \forall t : t_1 \leq t \leq t_2 \; \|\lambda_t\|_1 \geq \frac{2L}{\rho}.$$

By construction $1 < \frac{2L}{\rho} \leq \|\lambda_t\|_1 \leq \frac{2LM}{\rho^2}$ for all $t_1 \leq t \leq t_2$, and it holds if $\eta \leq \frac{1}{mL}$:

$$\|\lambda_{t_1}\|_1 \leq \|\lambda_{t_1-1}\|_1 + m\eta L \leq \frac{2L}{\rho} + m\eta L \leq \frac{4L}{\rho}. \tag{6.3}$$

Notice also that by construction, calling $\lambda_{t_1,t_2} = \max_{t\in[t_1,\dots t_2]}\|\lambda_t\|_1$, it holds:

$$1 < \lambda_{t_1,t_2} \leq \frac{2LM}{\rho^2} \quad \wedge \quad 1 + \lambda_{t_1,t_2} < \frac{4LM}{\rho^2}. \tag{6.4}$$

In the stochastic setting the following holds by Azuma-Hoeffding inequality with probability at least $1-\delta$:

$$\begin{aligned}\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond &\geq \sum_{t=t_1}^{t_2} -\lambda_t^\top \overline{G}^\top q^\diamond - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\ &\geq \lambda_{t_1,t_2}(t_2 - t_1 + 1)\rho - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} \\ &\geq (t_2 - t_1 + 1)2L - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2},\end{aligned}$$

where $\mathcal{E}^G_{t_1,t_2} = B_1\sqrt{(t_2 - t_1 + 1)} = 2H\sqrt{\ln\left(T^2/\delta\right)}\sqrt{(t_2 - t_1 + 1)}$. Instead, in the adversarial setting, it holds:

$$\begin{aligned}\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond &\geq \sum_{t=t_1}^{t_2}\sum_{i=1}^{m} -\lambda_{t,i}\left[G_t^\top q^\diamond\right]_i \\ &\geq \rho\sum_{t=t_1}^{t_2}\sum_{i=1}^{m}\lambda_{t,i} \\ &= \rho\sum_{t=t_1}^{t_2}\|\lambda_t\|_1 \\ &\geq (t_2 - t_1 + 1)2L.\end{aligned}$$

Generalizing the result, it holds, both for the stochastic and the adversarial setting, the following inequality with probability equal to 1 in the adversarial case and with probability at least $1-\delta$ in the stochastic case:

$$\sum_{t=t_1}^{t_2} -\lambda_t^\top G_t^\top q^\diamond \geq (t_2 - t_1 + 1)2L - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2}.$$

Thank to this result we can find a lower bound for $-\sum_{t=t_1}^{t_2}\ell_t^{\mathcal{L},\top}q_t$ with probability at least $1-9\delta$ in the stochastic setting and with probability at least $1-8\delta$ in the adversarial case, employing Theorem 6.1. Specifically, we proceed as follows:

$$\begin{aligned}-\sum_{t=t_1}^{t_2}\ell_t^{\mathcal{L},\top}q_t &= \sum_{t=t_1}^{t_2}(r_t^\top q^\diamond - \lambda^\top G_t^\top q^\diamond) - \sum_{t=t_1}^{t_2}\ell_t^{\mathcal{L},\top}(q_t - q^\diamond) \\ &\geq \sum_{t=t_1}^{t_2} -\lambda^\top G_t^\top q^\diamond - \sum_{t=t_1}^{t_2}\ell_t^{\mathcal{L},\top}(q_t - q^\diamond)\end{aligned} \tag{6.5}$$

$$\geq 2L(t_2 - t_1 + 1) - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} - \mathcal{E}^P_{t_1,t_2}, \tag{6.6}$$

where Inequality (6.5) holds since $r_t^\top q^\diamond \geq 0$, for all $t \in [T]$, and Inequality (6.6) is derived employing the bound on the primal interval regret given by Theorem 6.1 (that is, $\mathcal{E}^P_{t_1,t_2}$) and by Lemma D.2.
At the same time, it is possible to find the following upper bound for the same quantity $-\sum_{t=t_1}^{t_2} \ell_t^{\mathcal{L},\top} q_t$, with probability at least $1 - 2\delta$:

$$\begin{aligned}
&-\sum_{t=t_1}^{t_2} \ell_t^{\mathcal{L},\top} q_t \\
&\quad= \sum_{t=t_1}^{t_2} (r_t^\top q_t - \lambda_t^\top G_t^\top q_t) \\
&\quad\leq \sum_{t=t_1}^{t_2} L - \sum_{t=t_1}^{t_2} \lambda_t^\top (G_t^\top q_t - \sum_{k=0}^{L-1} G_t(x_k, a_k)) + \sum_{t=t_1}^{t_2} (\underline{0} - \lambda_t) \sum_{k=0}^{L-1} G_t(x_k, a_k) \\
&\quad\leq L(t_2 - t_1 + 1) + \lambda_{t_1,t_2} \sum_{t=t_1}^{t_2} \sum_{x,a} G_t(x,a)(\mathbb{I}_t(x,a) - q_t(x,a)) + \mathcal{E}^D_{t_1,t_2}(\underline{0}) \\
&\quad\leq L(t_2 - t_1 + 1) + \lambda_{t_1,t_2}\mathcal{E}^{\mathbb{I}}_{t_1,t_2} + \mathcal{E}^D_{t_1,t_2}(\underline{0}),
\end{aligned}$$

where $\mathcal{E}^{\mathbb{I}} = F_1\sqrt{(t_2 - t_1 + 1)} = L\sqrt{2\ln\left(T^2/\delta\right)}\sqrt{(t_2 - t_1 + 1)}$ and $\mathcal{E}^D(\underline{0}) = D_1\frac{\|\lambda_{t_1}\|_2^2}{\eta} + D_2\eta(t_2 - t_1 + 1) = \frac{1}{2}\frac{\|\lambda_{t_1}\|_2^2}{\eta} + \frac{mL^2}{2}\eta(t_2 - t_1 + 1)$. Thus, combining the two bounds we get with probability at least $1 - 10\delta$ in the adversarial case and $1 - 11\delta$ in the stochastic case the following bound,

$$2L(t_2 - t_1 + 1) - \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} - \mathcal{E}^P_{t_1,t_2} \leq L(t_2 - t_1 + 1) + \lambda_{t_1,t_2}\mathcal{E}^{\mathbb{I}}_{t_1,t_2} + \mathcal{E}^D_{t_1,t_2}(\underline{0}),$$

which can be reordered as

$$L(t_2 - t_1 + 1) \leq \lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2} + \lambda_{t_1,t_2}\mathcal{E}^{\mathbb{I}}_{t_1,t_2} + \mathcal{E}^D_{t_1,t_2}(\underline{0}) + \mathcal{E}^P_{t_1,t_2}.$$

We recall here the definitions of the bounds $\mathcal{E}^G_{t_1,t_2}, \mathcal{E}^{\mathbb{I}}_{t_1,t_2}, \mathcal{E}^D_{t_1,t_2}(\underline{0})$ and $\mathcal{E}^P_{t_1,t_2}$.

$$\mathcal{E}^G_{t_1,t_2} = B_1\sqrt{(t_2 - t_1 + 1)},$$

where $B_1 = 2L\sqrt{\ln\left(\frac{T^2}{\delta}\right)}$.

$$\mathcal{E}^{\mathbb{I}}_{t_1,t_2} = F_1\sqrt{(t_2 - t_1 + 1)},$$

where $F_1 = H\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}$.

$$\mathcal{E}^D_{t_1,t_2}(\underline{0}) = D_1\frac{\|\lambda_{t_1}\|_2^2}{\eta} + D_2\eta(t_2 - t_1 + 1),$$

where $D_1 = \frac{1}{2}$, $D_2 = \frac{mL^2}{2}$.

$$\mathcal{E}^P_{t_1,t_2} = U_1 \Xi_{t_1,t_2} C\sqrt{T} + U_2 \Xi_{t_1,t_2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}} + U_3 \Xi_{t_1,t_2} \frac{1}{C\sqrt{T}} + U_4 \Xi_{t_1,t_2} \sqrt{T},$$

where $U_1 = 6L^2 \ln\left(\frac{L|A|T^2}{\delta}\right)$, $U_2 = 9L|X||A|$, $U_3 = \frac{L}{2}\ln\left(\frac{LT^2}{\delta}\right)$ and $U_4 = 30L^2|X|^2\sqrt{2|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}$.

Thus, we can write:

$$\begin{aligned} L(t_2 - t_1 + 1) \leq & \underbrace{\lambda_{t_1,t_2}(F_1 + B_1)\sqrt{(t_2 - t_1 + 1)}}_{①} + \underbrace{U_2 \Xi_{t_1,t_2} \frac{t_2 - t_1 + 1}{C\sqrt{T}}}_{②} \\ & + \underbrace{U_3 \Xi_{t_1,t_2} \frac{1}{C\sqrt{T}}}_{③} + \underbrace{U_1 \Xi_{t_1,t_2} C\sqrt{T}}_{④} + \underbrace{D_2 \eta (t_2 - t_1 + 1)}_{⑤} \\ & + \underbrace{D_1 \frac{\|\lambda_{t_1}\|_2^2}{\eta}}_{⑥} + \underbrace{U_4 \Xi_{t_1,t_2} \sqrt{T}}_{⑦}. \end{aligned}$$

To conclude the proof by absurd it is sufficient to prove that all ①, ②, ③, ④, ⑤, ⑥, ⑦ are smaller or equal to $\frac{L(t_2-t_1+1)}{7}$, with at least one being strictly smaller.

$① < \frac{L(t_2-t_1+1)}{7}$ If $\eta \leq \frac{1}{14m(F_1+B_1)\sqrt{T}}$, then $① < \frac{L(t_2-t_1+1)}{7}$ holds. Indeed:

$$\begin{aligned} \frac{L(t_2 - t_1 + 1)}{7} &> \frac{LM}{7\rho^2 m\eta} && \text{(6.7a)} \\ &\geq \frac{2LM}{\rho^2}(F_1 + B_1)\sqrt{T} && \text{(6.7b)} \\ &\geq \lambda_{t_1,t_2}(F_1 + B_1)\sqrt{T} && \text{(6.7c)} \\ &\geq \lambda_{t_1,t_2}(F_1 + B_1)\sqrt{t_2 - t_1 + 1}, \end{aligned}$$

where Inequality (6.7a) holds by Lemma 5.2, Inequality (6.7b) is equivalent to condition $\eta \leq \frac{1}{14m(F_1+B_1)\sqrt{T}}$ and Inequality (6.7c) is true by Assumption (6.2).

$② < \frac{L(t_2-t_1+1)}{7}$ If $C \geq 56\frac{MU_2}{\rho^2\sqrt{T}}$ holds, then $② < \frac{L(t_2-t_1+1)}{7}$ also holds. Indeed:

$$\begin{aligned} \frac{L(t_2 - t_1 + 1)}{7} &\geq 2U_2\left(\frac{4LM}{\rho^2}\right)\frac{t_2 - t_1 + 1}{C\sqrt{T}} && \text{(6.8a)} \\ &> U_2 2(1 + \lambda_{t_1,t_2})\frac{t_2 - t_1 + 1}{C\sqrt{T}} && \text{(6.8b)} \\ &\geq U_2 \Xi_{t_1,t_2}\frac{t_2 - t_1 + 1}{C\sqrt{T}}, && \text{(6.8c)} \end{aligned}$$

where Inequality (6.8a) is equivalent to the condition $C \geq 56\frac{MU_2}{\rho^2\sqrt{T}}$, Inequality (6.8b) holds by Inequality (6.4) and Inequality (6.8c) is true since $\Xi_{t_1,t_2} \leq 2(1 + \lambda_{t_1,t_2})$.

$\boxed{③ < \frac{L(t_2-t_1+1)}{7}}$ If $\eta \leq \frac{C\sqrt{T}}{56mU_3}$ holds then also $③ < \frac{L(t_2-t_1+1)}{7}$ holds. Indeed:

$$\begin{aligned}
\frac{L(t_2-t_1+1)}{7} &> \frac{LM}{7\rho^2 m\eta} && \text{(6.9a)}\\
&\geq U_3 2\frac{4LM}{\rho^2}\frac{1}{C\sqrt{T}} && \text{(6.9b)}\\
&\geq U_3 2(1+\lambda_{t_1,t_2})\frac{1}{C\sqrt{T}} && \text{(6.9c)}\\
&\geq U_3 \Xi_{t_1,t_2}\frac{1}{C\sqrt{T}}, && \text{(6.9d)}
\end{aligned}$$

where Inequality (6.9a) hold by Lemma 5.2, Inequality (6.9b) holds if condition $\eta \leq \frac{C\sqrt{T}}{56mU_3}$ holds, and Inequality (6.9c) and Inequality (6.9d) follow the same reasoning as Inequality (6.8b) and Inequality (6.8c).

$\boxed{④ < \frac{L(t_2-t_1+1)}{7}}$ If $\eta \leq \frac{1}{56mU_1C\sqrt{T}}$ holds then also $④ < \frac{L(t_2-t_1+1)}{7}$ holds. Indeed:

$$\begin{aligned}
\frac{L(t_2-t_1+1)}{7} &> \frac{LM}{7\rho^2 m\eta}\\
&\geq U_1 2\frac{4LM}{\rho^2}C\sqrt{T} && \text{(6.10)}\\
&\geq U_1 2(1+\lambda_{t_1,t_2})C\sqrt{T}\\
&\geq U_1 \Xi_{t_1,t_2}C\sqrt{T},
\end{aligned}$$

where Inequality (6.10) holds when condition $\eta \leq \frac{1}{56mU_1C\sqrt{T}}$ also holds, and the rest of the inequalities follow a similar reasoning to the one used to bound ③.

$\boxed{⑤ \leq \frac{L(t_2-t_1+1)}{7}}$ It is immediate to see that if $\eta \leq \frac{L}{7D_2}$ holds, then it holds also that:

$$⑤ = D_2\eta(t_2-t_1+1) \leq \frac{L(t_2-t_1+1)}{7}.$$

$\boxed{⑥ < \frac{L(t_2-t_1+1)}{7}}$ If the condition $M \geq 112D_1Lm$ is satisfied than the inequality $⑥ < \frac{L(t_2-t_1+1)}{7}$ holds too. Indeed:

$$\begin{aligned}
\frac{L(t_2-t_1+1)}{7} &> \frac{LM}{7\rho^2 m\eta} && \text{(6.11a)}\\
&\geq D_1\frac{16L^2}{\rho^2}\frac{1}{\eta} && \text{(6.11b)}\\
&\geq D_1\frac{\|\lambda_{t_1}\|_1^2}{\eta} && \text{(6.11c)}\\
&\geq D_1\frac{\|\lambda_{t_1}\|_2^2}{\eta},
\end{aligned}$$

where Inequality (6.11a) holds by Lemma 5.2, Inequality (6.11b) holds when the condition $M \geq 112D_1Lm$ is satisfied and Inequality (6.11c) holds by Inequality (6.3).

$\boxed{⑦ < \frac{L(t_2-t_1+1)}{7}}$ If the condition $\eta \leq \frac{1}{56mU_4\sqrt{T}}$ is satisfied then $⑦ < \frac{L(t_2-t_1+1)}{7}$ also holds. In fact

$$\begin{aligned}\frac{L(t_2-t_1+1)}{7} &> \frac{LM}{7\rho^2 m\eta} && \text{(6.12a)}\\ &\geq U_4 2\frac{4LM}{\rho^2}\sqrt{T} && \text{(6.12b)}\\ &\geq U_4 2(1+\lambda_{t_1,t_2})\sqrt{T}\\ &\geq U_4 \Xi_{t_1,t_2}\sqrt{T}.\end{aligned}$$

where Inequality (6.12a) holds by Lemma 5.2 and inequality (6.12b) holds if condition $\eta \leq \frac{1}{56mU_4\sqrt{T}}$ also holds.

Conclusion of the proof Thus, we have the following 3 conditions:

- First condition:

$$\begin{aligned}M &\geq 112D_1 Lm\\ &= 112\frac{1}{2}Lm\\ &= 56Lm.\end{aligned}$$

- Second condition:

$$\begin{aligned}C &\geq 56\frac{MU_2}{\rho^2\sqrt{T}}\\ &= 56\frac{M9L|X||A|}{\rho^2\sqrt{T}}.\end{aligned}$$

- Third condition:

$$\eta \leq \min\left\{\frac{1}{14m(F_1+B_1)\sqrt{T}}, \frac{C\sqrt{T}}{56mU_3}, \frac{1}{56mU_1C\sqrt{T}}, \frac{L}{7D_2}, \frac{1}{56mU_4\sqrt{T}}\right\}.$$

We set M as $M = 56Lm$, and consequently, under Condition 6.1 we set $C = 252|X||A|L$ since

$$\begin{aligned}C &= 252|X||A|L\\ &\geq 252|X||A|\\ &\geq 252|X||A|\frac{112mL^2}{\rho^2}\frac{1}{\sqrt{T}}\\ &= 56\frac{(56Lm)9L|X||A|}{\rho^2\sqrt{T}}\\ &= 56\frac{9ML|X||A|}{\rho^2\sqrt{T}}.\end{aligned}$$

Notice that the inequality is deduced directly by Condition 6.1. In fact if $\rho \geq T^{-\frac{1}{8}} L\sqrt{112m}$ then it is also true that

$$\frac{112mL^2}{\rho^2} \leq T^{\frac{1}{4}} \leq \sqrt{T}.$$

As a final remark, we choose $252|X||A|L$ as value of C instead of the smaller value $252|X||A|$, which is useful for Lemma D.6. Finally we study the condition on η.

$$\begin{aligned}
&\min\left\{\frac{1}{14m(F_1+B_1)\sqrt{T}}, \frac{C\sqrt{T}}{56mU_3}, \frac{1}{56mU_1C\sqrt{T}}, \frac{L}{7D_2}, \frac{1}{56mU_4\sqrt{T}}\right\} \\
&\quad \geq \min\left\{\frac{1}{14m\left(4L\sqrt{\ln\left(\frac{T^2}{\delta}\right)}\right)\sqrt{T}}, \frac{252|X||A|L\sqrt{T}}{56m\left(\frac{L}{2}\ln\left(\frac{LT^2}{\delta}\right)\right)}, \right. \\
&\qquad \frac{1}{56m\left(6L^2\ln\left(\frac{L|A|T^2}{\delta}\right)\right)(252|X||A|L)\sqrt{T}}, \frac{L}{7\left(\frac{mL^2}{2}\right)}, \\
&\qquad \left.\frac{1}{56m\left(30L^2|X|^2\sqrt{2|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}\right)\sqrt{T}}\right\} \\
&\quad \geq \frac{1}{84672mL^2|X|^2|A|\ln\left(\frac{|A||X|^2T^2}{\delta}\right)\sqrt{T}}.
\end{aligned}$$

Thus, the proof is concluded taking $\eta = \frac{1}{84672mL^2|X|^2|A|\ln\left(\frac{|A||X|^2T^2}{\delta}\right)\sqrt{T}}$. □

6.3 Theoretical Analysis

In this section, we prove the best-of-both-world guarantees attained by Algorithm 6.1.

6.3.1 Stochastic Setting

We first study the performance of Algorithm 6.1 when the constraints are stochastic.

In such a setting, our algorithm can handle two scenarios. In both of them, employing a primal-dual analysis shows that both the regret and the violations are bounded with order $\widetilde{\mathcal{O}}(\sqrt{T})$ times the maximum value taken over all episodes of the Lagrangian multipliers, *i.e.* $\max_{t\in[T]}\|\lambda_t\|_1$. In the first scenario, Condition 6.1 holds and thus we can apply Theorem 6.2 to show that the Lagrangian multipliers are bounded. In such a case, $\max_{t\in[T]}\|\lambda_t\|_1$ can be easily bounded by Λ. When Conditions 6.1 does *not* hold, we need to resort to the bound of Lagrangian multipliers derived by the instantiation of OGD decision space, leading to $\widetilde{\mathcal{O}}(T^{3/4})$ regret and violations bounds.

Specifically, when Condition 6.1 holds, the Lagrangian multipliers are nicely bounded by Λ.

Theorem 6.3. *Suppose that Condition 6.1 holds and the constraints are generated stochas-*

tically. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains:

$$R_T \leq \widetilde{\mathcal{O}}\left(\Lambda\sqrt{T}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(\Lambda\sqrt{T}\right),$$

with probability at least $1-14\delta$ when the rewards are stochastic, and with probability at least $1-13\delta$ when the rewards are adversarial.

Proof. With probability at least $1-12\delta$ it holds:

$$V_T \leq \widehat{V}_{T,i^*} + \mathcal{E}^{\mathbb{I}} \tag{6.13a}$$

$$\leq \frac{1}{\eta}\lambda_{T+1,i^*} + \mathcal{E}^{\mathbb{I}} \tag{6.13b}$$

$$\leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}}, \tag{6.13c}$$

where Inequality (6.13a) holds by Lemma D.4, Inequality (6.13b) holds by Lemma D.3 and Inequality (6.13c) holds by Theorem 6.2. Then, with probability at least $1-12\delta$ we observe that:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t} \\
&\quad\leq \sum_{t=1}^{T}\left(r_t^\top q^* - \lambda_t^\top G_t^\top q^*\right) - \sum_{t=1}^{T}\left(r_t^\top q_t - \lambda_t^\top G_t^\top q_t\right) + \sum_{t=1}^{T}\lambda_t^\top G_t^\top (q^* - q_t) \\
&\quad\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\lambda_t^\top G_t^\top q^* && (6.14a) \\
&\quad= \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\lambda_t^\top (G_t - \overline{G})^\top q^* + \sum_{t=1}^{T}\lambda_t^\top \overline{G}^\top q^* \\
&\quad\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \lambda_{1,T}\mathcal{E}^G && (6.14b) \\
&\quad\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + \Lambda\mathcal{E}^G, && (6.14c)
\end{aligned}$$

where Inequality (6.14a) holds by Theorem 6.1 and by Theorem D.2, Inequality (6.14b) holds since in the stochastic constraint case $\sum_{t=1}^{T}(G_t - \overline{G})^\top q^* \leq \mathcal{E}^G$ with probability at least $1-\delta$ by definition of $\mathcal{E}^G$, and finally Inequality (6.14c) holds by Theorem 6.2. Finally we observe that in the stochastic case with probability at least $1-\delta$:

$$\left(T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \sum_{t=1}^{T} r_t^\top q_t\right) - \sum_{t=1}^{T} r_t^\top (q^* - q_t) \leq \mathcal{E}^r.$$

Thus, if the rewards are stochastic with probability at least $1-14\delta$ it holds:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + \Lambda\mathcal{E}^G + \mathcal{E}^r, \quad V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

and if the rewards are adversarial with probability at least $1-13\delta$ it holds:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + \Lambda\mathcal{E}^G, \quad V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

which concludes the proof. □

When Condition 6.1 does *not* hold, we can still use the bound forced by Algorithm 6.1 on the dual space. Therefore, the Lagrangian multipliers are bounded by $mT^{1/4}$, leading to the following result.

Theorem 6.4. *Suppose that Condition 6.1 does not hold and the constraints are generated stochastically. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains:*

$$R_T \leq \widetilde{\mathcal{O}}\left(T^{3/4}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(T^{3/4}\right),$$

with probability at least $1-11\delta$ when the rewards are stochastic, and with probability at least $1-10\delta$ when the rewards are adversarial.

Proof. Similar to the proof of Lemma 6.3 it holds with probability at least $1-10\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t} \\
&\quad \leq \sum_{t=1}^{T} \left(r_t^\top q^* - \lambda_t^\top G_t^\top q^*\right) - \sum_{t=1}^{T} \left(r_t^\top q_t - \lambda_t^\top G_t^\top q_t\right) + \sum_{t=1}^{T} \lambda_t^\top G_t^\top (q^* - q_t) \\
&\quad \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^* \\
&\quad = \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T} \lambda_t^\top (G_t - \overline{G})^\top q^* + \sum_{t=1}^{T} \lambda_t^\top \overline{G}^\top q^* \\
&\quad \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \lambda_{1,T}\mathcal{E}^G.
\end{aligned}$$

Therefore, with probability at least $1-10\delta$, following the reasoning of Lemma D.5 to bound the dual decision space, it holds when the rewards are adversarial:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - mT^{1/4}\mathcal{E}^G + mT^{1/4}\mathcal{E}^{\mathbb{I}} - \mathcal{E}^D(\underline{0}) - \mathcal{E}^P,$$

and when the rewards are stochastic, with probability ta least $1-11\delta$, it holds:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - mT^{1/4}\mathcal{E}^G + mT^{1/4}\mathcal{E}^{\mathbb{I}} - \mathcal{E}^D(\underline{0}) - \mathcal{E}^P - \mathcal{E}^r.$$

Applying Lemma D.5 to bound the constraints violation concludes the proof. □

6.3.2 Adversarial Setting

We then study the performance of Algorithm 6.1 when the constraints are adversarial.

Notice that, in such a setting, there exists an impossibility result from (Mannor et al., 2009) that prevents any algorithm from attaining both sulinear regret and sublinear violations. Thus, best-of-both-worlds algorithms in constrained settings focus on attaining sublinear violations and a fraction of the optimal rewards (see *e.g.*, (Castiglioni et al.,

2022b)).[3]

In such a setting, we can show the following result.

Theorem 6.5. *Suppose Condition 6.1 holds and the constraints are adversarial. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains:*

$$\sum_{t=1}^{T} r_t^\top q_t \geq \Omega\left(\frac{\rho}{\rho + L} \cdot \mathrm{OPT}_{\overline{r},\overline{G}}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(\Lambda\sqrt{T}\right),$$

with probability at least $1 - 14\delta$ when the rewards are stochastic, and with probability at least $1 - 13\delta$ when the rewards are adversarial.

Proof. Thanks to Theorem 6.1 , Theorem D.2 and Theorem 6.2 with probability at least $1 - 11\delta$ it holds for all $q \in \Delta(M)$:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q - \sum_{t=1}^{T} r_t^\top q_t \\
&\quad\leq -\sum_{t=1}^{T} \ell_t^{\mathcal{L},\top} q + \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top} q_t + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t \\
&\quad\leq \mathcal{E}_T^P + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q + \sum_{t=1}^{T} (\underline{0} - \lambda_t)^\top \sum_{k=0}^{L-1} G_t(x_k, a_k) \\
&\qquad\qquad + \sum_{t=1}^{T} \lambda_t^\top \left(\sum_{k=0}^{L-1} G_t(x_k, a_k) - G_t^\top q_t\right) \\
&\quad\leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\sum_{i=1}^{m} \lambda_{t,i} g_{t,i}^\top q \\
&\quad\leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\sum_{i=1}^{m} \lambda_{t,i} g_{t,i}^\top q \\
&\quad\leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\sum_{i=1}^{m} \lambda_{t,i} g_{t,i}^\top q.
\end{aligned}$$

Consider now the occupancy measure $\widetilde{q} = \frac{\rho}{L+\rho} q^* + \frac{L}{L+\rho} q^\diamond$. For all $i \in [m]$ and for all $t \in [T]$:

$$\begin{aligned}
g_{t,i}^\top \widetilde{q} &\leq \left(\frac{\rho}{L+\rho} g_{t,i}^\top q^* + \frac{L}{L+\rho} g_{t,i}^\top q^\diamond\right) \\
&\leq \left(\frac{L\rho}{L+\rho} - \frac{L\rho}{L+\rho}\right) \\
&= 0,
\end{aligned}$$

given that $g_{t,i}^\top q^* \leq \|q^*\|_1 \leq L$, and $g_{t,i}^\top q^\diamond \leq -\rho$ by definition of $q^\diamond$ and by definition

[3] Attaining the no-α-regret property, that is, being no-regret w.r.t. a fraction of the optimum, achieving a competitive ratio, and guaranteeing a fraction of the optimal rewards are used as synonyms in the literature, since any of the aforementioned guarantees can be derived by the others.

of ρ. Moreover, it holds:

$$\sum_{t=1}^{T} r_t^\top \widetilde{q} = \sum_{t=1}^{T} \left(\frac{\rho}{L+\rho} r_t^\top q^* + \frac{L}{L+\rho} r_t^\top q^\diamond \right)$$
$$\geq \frac{\rho}{L+\rho} \sum_{t=1}^{T} r_t^\top q^*,$$

since $r_t^\top q^\diamond \geq 0$. Notice also that with adversarial rewards $\sum_{t=1}^{T} r_t^\top q^* = T \cdot \text{OPT}_{\overline{r},\overline{G}}$, while with stochastic rewards with probability at least $1-\delta$ it holds $\sum_{t=1}^{T} r_t^\top q^* \geq T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}^r$, by definition of $\mathcal{E}^r$ and $\text{OPT}_{\overline{r},\overline{G}}$ for stochastic rewards. By reordering the terms we get that with probability at least $1-11\delta$:

$$\sum_{t=1}^{T} r_t^\top q_t \geq \frac{\rho}{L+\rho} \sum_{t=1}^{T} r_t^\top q^* - \mathcal{E}_T^P - \mathcal{E}_T^D(\underline{0}) - \Lambda \mathcal{E}^{\mathbb{I}}.$$

We can proceed to bound the regret in both cases: adversarial rewards and stochastic rewards.
With probability at least $1-11\delta$ with adversarial rewards it holds:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q_t \\ &\leq \sum_{t=1}^{T} r_t^\top q^* - \left(\frac{\rho}{L+\rho} \sum_{t=1}^{T} r_t^\top q^* - \mathcal{E}_T^P - \mathcal{E}_T^D(\underline{0}) - \Lambda \mathcal{E}^{\mathbb{I}} \right) \\ &\leq \frac{L}{L+\rho} \sum_{t=1}^{T} r_t^\top q^* + \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \Lambda \mathcal{E}^{\mathbb{I}} \\ &\leq \frac{L}{L+\rho} T \cdot \text{OPT}_{\overline{r},\overline{G}} + \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \Lambda \mathcal{E}^{\mathbb{I}}. \end{aligned}$$

With stochastic rewards it holds with probability at least $1-11\delta$:

$$\sum_{t=1}^{T} r_t^\top q_t \geq \frac{\rho}{L+\rho} \sum_{t=1}^{T} r_t^\top q^* - \mathcal{E}_T^P - \mathcal{E}_T^D(\underline{0}) - \Lambda \mathcal{E}^{\mathbb{I}},$$

and with probability at least $1-12\delta$:

$$\sum_{t=1}^{T} r_t^\top q_t \geq \frac{\rho}{L+\rho} T \cdot \text{OPT}_{\overline{r},\overline{G}} - \mathcal{E}_T^P - \mathcal{E}_T^D(\underline{0}) - \Lambda \mathcal{E}^{\mathbb{I}} - \mathcal{E}^r.$$

To conclude the proof we observe that following the analogous reasoning to Theorem 6.3 in case of adversarial constraints it also holds with probability at least $1-12\delta$:

$$V_T \leq \frac{1}{\eta} \Lambda + \mathcal{E}^{\mathbb{I}}.$$

This concludes the proof. □

A Weaker Baseline

In this section, we show that the impossibility result by Mannor et al. (2009) can be circumvented by adopting a different baseline in the regret definition. Precisely, we compute the weaker baseline as the solution to the following linear program:

$$\text{OPT}^{\mathcal{W}} := \begin{cases} \max_{q \in \Delta(M)} & \overline{r}^\top q \\ \text{s.t.} & G_t^\top q \leq \underline{0} \quad \forall t \in [T]. \end{cases}$$

Notice that, in the previous sections, we allow the optimal policy q^* to satisfy the constraints on average, *i.e.*, $\sum_{t=1}^T G_t^\top q^* \leq 0$. In such a case, the set of feasible policies is much smaller than the one associated with the weaker baseline, that is, when a feasible policy must satisfy the constraints *at each episode*. Given the new baseline, we can rewrite the regret as $R_T := T\,\text{OPT}^{\mathcal{W}} - \sum_{t=1}^T r_t^\top q_t$.

When the regret is computed w.r.t. the weaker baseline, we can recover the same theoretical results of the stochastic setting. Precisely, when Condition 6.1 holds we have the following result.

Theorem 6.6. *Suppose that Condition 6.1 holds and the constraints are generated adversarially. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains:*

$$R_T \leq \widetilde{\mathcal{O}}\left(\Lambda\sqrt{T}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(\Lambda\sqrt{T}\right),$$

with probability at least $1 - 13\delta$ when the rewards are stochastic, and with probability at least $1 - 12\delta$ when the rewards are adversarial.

Proof. The violation can be bounded as in Theorem 6.3. Moreover, similarly to Theorem 6.3, it holds, with probability $1 - 12\delta$:

$$\begin{aligned}
&\sum_{t=1}^T r_t^\top q^* - \sum_{t=1}^T r_t^\top q_t \\
&\quad \leq -\sum_{t=1}^T \ell_t^{\mathcal{L},\top} q^* + \sum_{t=1}^T \ell_t^{\mathcal{L},\top} q_t + \sum_{t=1}^T \lambda_t^\top G_t^\top q^* - \sum_{t=1}^T \lambda_t^\top G_t^\top q_t \\
&\quad \leq \mathcal{E}_T^P + \sum_{t=1}^T (\underline{0} - \lambda_t)^\top \sum_{k=0}^{L-1} G_t(x_k, a_k) + \sum_{t=1}^T \lambda_t^\top \left(\sum_{k=0}^{L-1} G_t(x_k, a_k) - G_t^\top q_t \right) \\
&\quad \leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \lambda_{1,T} \mathcal{E}^{\mathbb{I}} \\
&\quad \leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \Lambda \mathcal{E}^{\mathbb{I}}.
\end{aligned}$$

Finally, when the rewards are stochastic, it holds, with probability at least $1 - \delta$:

$$T \cdot \text{OPT}_{\overline{r},\overline{G}} - \sum_{t=1}^T r_t^\top q^* \leq \mathcal{E}^r.$$

Thus, when the rewards are adversarial it holds with probability at least $1 - 12\delta$:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda \mathcal{E}^{\mathbb{I}}, \quad V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

while when the rewards are stochastic it holds, with probability at least $1 - 13\delta$:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + \mathcal{E}^r, \quad V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

which concludes the proof. □

We conclude the section by analyzing the scenario in which Condition 6.1 does *not* hold.

Theorem 6.7. *Suppose that Condition 6.1 does not hold and the constraints are generated adversarially. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains:*

$$R_T \leq \widetilde{\mathcal{O}}\left(T^{3/4}\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(T^{3/4}\right),$$

with probability at least $1 - 12\delta$ when the rewards are stochastic, and with probability at least $1 - 11\delta$ when the rewards are adversarial.

Intuitively, Theorems 6.6 and 6.7 can be proved by the fact that playing the optimal policy guarantees small violations independently on the episode the optimum is chosen. This is *not* the case for the stronger baseline, since playing the optimum in some episodes may lead to arbitrarily large constraint violations.

Proof. The violation can be bounded thanks to Lemma D.3, as in Theorem 6.4. To bound the regret, notice that it holds, with probability $1 - 9\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q_t \\
&\quad \leq -\sum_{t=1}^{T} \ell_t^{\mathcal{L},\top} q^* + \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top} q_t + \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^* - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t \\
&\quad \leq \mathcal{E}_T^P + \sum_{t=1}^{T} (\underline{0} - \lambda_t)^\top \sum_{k=0}^{L-1} G_t(x_k, a_k) + \sum_{t=1}^{T} \lambda_t^\top \left(\sum_{k=0}^{L-1} G_t(x_k, a_k) - G_t^\top q_t \right) \\
&\quad \leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} \\
&\quad \leq \mathcal{E}_T^P + \mathcal{E}_T^D(\underline{0}) + mT^{1/4}\mathcal{E}^{\mathbb{I}}.
\end{aligned}$$

Furthermore, when the rewards are stochastic, it holds, with probability at least $1 - \delta$:

$$T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \sum_{t=1}^{T} r_t^\top q^* \leq \mathcal{E}^r.$$

Therefore, when the rewards are adversarial it holds, with probability at least $1 - 11\delta$:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + mT^{1/4}\mathcal{E}^{\mathbb{I}}, \quad V_T \leq \frac{4T^{1/4}}{\eta},$$

and when the rewards are stochastic, it holds with probability at least $1 - 12\delta$:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + mT^{1/4}\mathcal{E}^{\mathbb{I}} + \mathcal{E}^r, \quad V_T \leq \frac{4T^{1/4}}{\eta},$$

which concludes the proof. □

CHAPTER 7

Beyond Slater's Condition

In this chapter, we study *online learning* in episodic CMDPs where the constraints may be either stochastic or adversarial, under *bandit feedback*. We propose a novel algorithm that greatly improves the best-of-both-worlds results provided in Chapter 6. Specifically, in the stochastic setting, our algorithm attains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret R_T and violation V_T without Slater's condition, *i.e.*, even when a strictly feasible solution does not exist. Furthermore, our algorithm attains $\widetilde{\mathcal{O}}(\sqrt{T})$ *strong* constraint violation $\mathcal{V}_T$, which does not allow for cancellations between episodes. In the adversarial setting, our algorithm attains sublinear violation without Slater's condition. Furthermore, by employing a slightly stronger notion of Slater's parameter, our algorithm attains sublinear α-regret with respect to the *unconstrained* optimum, instead of the constrained one.

Crucially, the algorithm proposed in this chapter is not *primal-dual* but it is *projection based*, with the per-episode projection performed over a moving decision space that adapts to the violation attained by the algorithm. This allows to avoid the use of no-interval regret algorithms as in Chapter 5 - 6, since no Lagrangian variables are involved.

In Table 7.1, we summarize the comparison with the results provided in Chapter 6.

7.1 Setting and Additional Notation

We study episodic constrained MDPs, when *bandit* feedback is available. We consider the case where the rewards are adversarial. Similarly to Chapters 5 - 6, $\{G_t\}_{t=1}^T$ is a sequence of constraint matrices describing the m *constraint* costs at each episode $t \in [T]$, namely $G_t \in [-1,1]^{|X\times A|\times m}$, where non-strictly positive cost values stand for satisfaction of the constraints. For $i \in [m]$, we refer to the cost of the i-th constraint for a specific state-action pair $x \in X, a \in A$ at episode $t \in [T]$ as $g_{t,i}(x,a)$. Constraint costs may be *stochastic* (we

will refer to this case as stochastic setting), in that case G_t is a random variable distributed according to a probability distribution $\mathcal{G}$ for every $t \in [T]$, or chosen by an *adversary* (we will refer to this case as adversarial setting).

Table 7.1: *Comparison between the performance of Algorithm 7.1, that is, the algorithm presented in this chapter, and the one of Algorithm 6.1. † The result assumes the existence of a stronger notion of the Slater's parameter ρ.*

	Algorithm 6.1	Algorithm 7.1
R_T *Stoc. Constraints*	$\widetilde{\mathcal{O}}\left(\min\left\{\frac{1}{\rho^2}\sqrt{T}, T^{\frac{3}{4}}\right\}\right)$	$\widetilde{\mathcal{O}}(\sqrt{T})$
V_T *Stoc. Constraints*	$\widetilde{\mathcal{O}}\left(\min\left\{\frac{1}{\rho^2}\sqrt{T}, T^{\frac{3}{4}}\right\}\right)$	$\widetilde{\mathcal{O}}(\sqrt{T})$
$\mathcal{V}_T$ *Stoc. Constraints*	✗	$\widetilde{\mathcal{O}}(\sqrt{T})$
α-R_T *Adv. Constraints*	$\widetilde{\mathcal{O}}\left(\frac{1}{\rho^2}\sqrt{T}\right)$	$\widetilde{\mathcal{O}}(\sqrt{T})^\dagger$
V_T *Adv. Constraints*	$\widetilde{\mathcal{O}}\left(\frac{1}{\rho^2}\sqrt{T}\right)$	$\widetilde{\mathcal{O}}(\sqrt{T})$

Remark 7.1 (On the stochastic rewards setting)**.** *In this chapter, we focus exclusively on the adversarial reward setting, unlike for the constraints, where both stochastic and adversarial scenarios are analyzed. This is because the stochastic reward setting follows directly from the adversarial reward one by a straightforward application of the Azuma–Hoeffding inequality.*

7.1.1 Baseline for the Stochastic Setting

We define the safe optimum for the stochastic constraints setting as follows:

$$\mathrm{OPT}_{\overline{G}} := \begin{cases} \max_{q \in \Delta(M)} & \frac{1}{T}\sum_{t=1}^{T} r_t^\top q \\ \text{s.t.} & \overline{G}^\top q \leq \underline{0}, \end{cases} \tag{7.1}$$

where $q \in [0,1]^{|X \times A|}$ is the occupancy measure vector, $\Delta(M)$ is the set of valid occupancy measures, and $\overline{G}$ is the expected value of $\mathcal{G}$. Thus, we restate the notion of *cumulative regret* as:

$$R_T := T \cdot \mathrm{OPT}_{\overline{G}} - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t}.$$

We refer to an optimal safe occupancy measure (*i.e.*, a feasible one achieving value $\mathrm{OPT}_{\overline{G}}$) as q^*. Thus, the regret reduces to $R_T = \sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t}$.

7.1.2 Baseline for the Adversarial Setting

In the adversarial case, we define the *cumulative α-regret* as follows:

$$\alpha\text{-}R_T = \alpha T \cdot \mathrm{OPT} - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t},$$

where the *unconstrained* optimal value is defined as $\text{OPT} := \max_{q \in \Delta(M)} \frac{1}{T} \sum_{t=1}^{T} r_t^\top q$. In order to quantify α, we introduce a *new* notion of the problem-specific parameter $\rho \in [0, 1]$, which is defined as:

$$\rho := \max_{q \in \Delta(M)} \min_{(x,a) \in \mathcal{Q}(q)} \min_{t \in [T]} \min_{i \in [m]} -g_{t,i}(x, a),$$

where $\mathcal{Q}(q) := \{(x, a) \in X \times A : q(x, a) > 0\}$. Thus, we define $\alpha := \rho/1+\rho$. We denote the occupancy measure leading to the value of ρ as $q^\diamond$. Intuitively, ρ represents the "margin" by which the "most feasible" strictly feasible occupancy satisfies the constraints, in the "worst" state-action pair. Our definition of ρ is slightly stronger than the one employed in Chapters 5 - 6, where the problem-specific parameter is not computed with respect to the "worst" state-action pair. Nonetheless, we underline that the baseline employed for the adversarial setting is the unconstrained optimum, while in Chapters 5 - 6, we provide no-α regret guarantees with respect to the *constrained* optimum, only.

7.2 Algorithm

In this section, we describe the key components of *Weighted Constrained Optimistic Policy Search* (`WC-OPS`, for short), which is the main algorithmic contribution of this chapter. In Algorithm 7.1, we provide the pseudocode of `WC-OPS`.

Algorithm 7.1 Weighted Constrained Optimistic Policy Search (`WC-OPS`)

Require: T, X, A, η, γ, δ

Initialize occupancy $\widehat{q}_1 \leftarrow \frac{1}{|X_k||A||X_{k+1}|}$, the estimated transitions space $\mathcal{P}_0$ as the set of all the possible transition functions, and counters $N_0(x, a) = M_0(x'|x, a) = 0$ for all $k \in [0, ..., L-1]$ and $(x, a, x') \in X_k \times A \times X_{k+1}$

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

  Play policy $\pi_t \leftarrow \pi^{\widehat{q}_t}$

  Observe *bandit* feedback as in Algorithm 2.1

  Set $\ell_t(x, a) \leftarrow 1 - r_t(x, a)\mathbb{I}_t(x, a)$ for all $x \in X, a \in A$

  Compute $\widehat{\ell}_t(x, a) = \frac{\ell_t(x,a)}{u_t(x,a)+\gamma}\mathbb{I}_t(x, a)$

  Update counters and compute weights as shown in Equation (7.2)

  Compute $\widehat{g}_{t,i}(x, a) = \sum_{\tau \in \mathcal{T}_{t,x,a}} w_{t,x,a,i}(\tau) g_{\tau,i}(x, a)$ for all $x \in X, a \in A, i \in [m]$

  Update confidence set $\mathcal{P}_t$ and bonus b_t as prescribed in Equations (7.3)–(7.4)

  $\widehat{\Delta}_t(\mathcal{P}_t) \leftarrow \left\{q \in \Delta(\mathcal{P}_t) : (\widehat{g}_{t,i} - b_t)^\top q \leq 0 \quad \forall i \in [m]\right\}$

  Update $\widehat{q}_{t+1} \leftarrow \arg\min_{q \in \widehat{\Delta}_t(\mathcal{P}_t)} \widehat{\ell}_t^\top q + \frac{1}{\eta} B(q||\widehat{q}_t)$

end for

7.2.1 Initialization and Loss Estimation

Algorithm 7.1 receives as input the time horizon T, the set of states X, the set of actions A, the learning rate η, the implicit exploration factor γ, and the confidence $\delta \in (0, 1)$. The occupancy measure $\widehat{q}_1$ is initialized uniformly over all tuples $(x_k, a, x_{k+1}) \in X_k \times A \times X_{k+1}$ for each layer $k \in [0, \ldots, L-1]$. The transition function confidence set $\mathcal{P}_0$ is initialized as the set of all the possible transition functions. The counters $N_t(x, a)$ and $M_t(x'|x, a)$, which are respectively defined as $N_t(x, a) = \sum_{\tau=1}^{t} \mathbb{I}_\tau(x, a)$ for all $(x, a) \in X \times A$, $M_t(x'|x, a) = \sum_{\tau=1}^{t} \mathbb{I}_\tau(x, a, x')$ for all $(x, a, x') \in X_k \times A \times X_{k+1}, k \in$

$[0, ..., L-1]$, are initialized to 0 (see Line 1). We denote by $\mathbb{I}_t(x,a)$ and $\mathbb{I}_t(x,a,x')$ the indicator functions for the state-action(-state) visit at episode t.

At the beginning of each episode t, the algorithm executes the policy π_t induced by the occupancy measure $\widehat{q}_t$ computed at the previous episode (Line 3). After selecting the policy, the learner interacts with the environment and receives the feedback (Line 4). The loss vector ℓ_t is built from the observed reward vector r_t (Line 5). Then, the algorithm builds a *biased* loss estimator $\widehat{\ell}_t$ for episode t, following the optimistic approach originally proposed in (Neu, 2015; Jin et al., 2020a). Specifically, given the transition function confidence set $\mathcal{P}_t$—refer to Equation (7.3) for additional details—, which contains the true transition function with high probability, the algorithm builds an optimistic estimator of ℓ_t. This is done by employing an upper bound on the occupancy u_t, in place of the unknown true occupancy q_t, defined as $u_t(x,a) = \max_{P_t \in \mathcal{P}_t} q^{P_t,\pi_t}(x,a)$ for all $(x,a) \in X \times A$. This upper bound represents the maximum probability of visiting (x,a) under any transition function within the set $\mathcal{P}_t$. Thus, the estimator is computed as $\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{u_t(x,a)+\gamma}\mathbb{I}_t(x,a)$, where γ is the implicit exploration factor given as input (Line 6).

7.2.2 Weights Estimation

At each episode, the counters are updated given the trajectory observed as feedback, namely, $N_t(x,a)$ and $M_t(x'|x,a)$ are updated by incrementing by 1 the entries of the tuples visited during the current episode. Then, the algorithm sets the weights that will be used to build the constraint estimates (Line 7). Specifically, given a pair $(x,a) \in X \times A$, $i \in [m]$, and $t \in [T]$, the weights $w_{t,x,a,i}$ are defined as follows:

$$w_{t,x,a,i}(\tau) \coloneqq \beta_{\tau,i}(x,a) \prod_{h \in \mathcal{T}_{t,x,a} : h > \tau} (1 - \beta_{h,i}(x,a)) \quad \forall \tau \in \mathcal{T}_{t,x,a}, \tag{7.2}$$

where $\mathcal{T}_{t,x,a}$ is the set of episodes where the pair (x,a) has been visited up to episode t, that is:

$$\mathcal{T}_{t,x,a} \coloneqq \{\tau \leq t : \mathbb{I}_\tau(x,a) = 1\}.$$

Moreover, the constraints learning rates $\beta_{t,i}$ are defined as:

$$\beta_{t,i}(x,a) \coloneqq \frac{1}{N_t(x,a)}\left(1 + \Gamma_{t,i}\right),$$

where $\Gamma_{t,i}$ is an adaptive term that depends on the constraint vectors observed and is defined as:

$$\Gamma_{t,i} \coloneqq \left[\sum_{\tau \in [t]} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) - \mathcal{C}_t\right]_0^{\mathcal{C}_t},$$

$\mathcal{C}_t \coloneqq 21L|X|\sqrt{2t|A| \ln \frac{2mT^2|X||A|}{\delta}}$ and $[\cdot]_a^b \coloneqq \min(\max(\cdot, a), b)$. Finally, the weights are employed to build the estimates $\widehat{g}_{t,i}$ for each constraint $i \in [m]$ and each (x,a) as the weighted mean of the values observed during the rounds in $\mathcal{T}_{t,x,a}$ (Line 8). Intuitively, the Γ_t parameter allows the learning rates to meet the requirements of both the stochastic and the adversarial setting, as we point out in the following. In order to better understand it, we first introduce the following result.

Proposition 7.1. *If $\beta_{\tau,i}(x,a) = \frac{1}{N_\tau(x,a)}$ for every $\tau \in \mathcal{T}_{t,x,a}$, then the following holds:*

$$w_{t,x,a,i}(\tau) = \frac{1}{N_t(x,a)},$$

and we recover the empirical mean estimator:

$$\widehat{g}_{t,i}(x,a) = \frac{1}{N_t(x,a)} \sum_{\tau \in \mathcal{T}_{t,x,a}} g_{\tau,i}(x,a).$$

Proof. Consider a pair $(x,a) \in X \times A$, an index $i \in [m]$, and $t \in [T]$. By definition of the weights, it holds:

$$\begin{aligned} w_{t,x,a,i}(\tau) &= \beta_{\tau,i}(x,a) \prod_{h \in \mathcal{T}_{t,x,a} : h > \tau} (1 - \beta_{h,i}(x,a)) \\ &= \frac{1}{N_\tau(x,a)} \prod_{h \in \mathcal{T}_{t,x,a} : h > \tau} \left(1 - \frac{1}{N_h(x,a)}\right), \quad \forall \tau \in \mathcal{T}_{t,x,a}. \end{aligned}$$

We now focus on the term $\prod_{h \in \mathcal{T}_{t,x,a} : h > \tau} \left(1 - \frac{1}{N_h(x,a)}\right)$:

$$\begin{aligned} \prod_{h \in \mathcal{T}_{t,x,a} : h > \tau} \left(1 - \frac{1}{N_h(x,a)}\right) &= \prod_{h \in \mathcal{T}_{t,x,a} : h > \tau} \frac{N_h(x,a) - 1}{N_h(x,a)} \\ &= \prod_{j = N_\tau(x,a)+1}^{N_t(x,a)} \frac{j-1}{j} \\ &= \frac{N_\tau(x,a)}{N_t(x,a)}. \end{aligned}$$

Thus:

$$w_{t,x,a,i}(\tau) = \frac{1}{N_\tau(x,a)} \frac{N_\tau(x,a)}{N_t(x,a)} = \frac{1}{N_t(x,a)}.$$

This concludes the proof. □

Proposition 7.1 simply states that, when $\Gamma_{t,i} = 0$, the weighted approach is equivalent to the empirical mean estimator. Indeed, as we will show in Section 7.3, this is exactly the case when the constraints are stochastic and the empirical mean estimator is sufficient to estimate the constraints. Differently, in the adversarial case, the learning rate is proportional to the violation attained by the algorithm, thus allowing $\widehat{g}_{t,i}$ to move accordingly to the attained performance.

7.2.3 Decision Space Definition and Optimization Update

Given the constraints estimates, Algorithm 7.1 has to properly build the decision space at each episode. Indeed, the algorithm has to ensure that such a decision space includes the true transition function and the true constraint functions, with high probability. In order to do that, Algorithm 7.1 updates its model (Line 9) accordingly.

For the transitions, we follow the approach of Rosenberg and Mansour (2019b). Specifically, the transition function confidence set $\mathcal{P}_t$ is updated as follows:

$$\mathcal{P}_t = \left\{ \widehat{P} : \|\widehat{P}(\cdot|x,a) - \bar{P}_t(\cdot|x,a)\|_1 \leq \epsilon_t(x,a) \right\}, \tag{7.3}$$

where the confidence width $\epsilon_t(x,a)$ is defined as:

$$\epsilon_t(x,a) = \sqrt{\frac{2|X_{k(x)+1}| \ln \frac{T|X||A|}{\delta}}{\max\{1, N_t(x,a)\}}}, \ \forall (x,a) \in X \times A,$$

and the empiric transition $\bar{P}_t$ is defined as:

$$\bar{P}_t(x'|x,a) = \frac{M_t(x'|x,a)}{\max\{1, N_t(x,a)\}} \quad \forall (x,a,x') \in X_k \times A \times X_{k+1}, k \in [0, \dots, L-1].$$

Given $\mathcal{P}_t$, it is possible to build $\Delta(\mathcal{P}_t)$ as the set of all possible occupancy measures.

For the constraints, we build optimistic bonuses $b_t(x,a)$ that are computed as:

$$b_t(x,a) = \sqrt{\frac{2 \ln \frac{2m|X||A|T}{\delta}}{N_t(x,a)}} \quad \forall (x,a) \in X \times A. \tag{7.4}$$

At each episode, the algorithm estimates the per-episode decision space $\widehat{\Delta}_t(\mathcal{P}_t)$ taking the intersection between $\Delta(\mathcal{P}_t)$ and the space of optimistically safe occupancy measures such that $(\widehat{g}_{t,i} - b_t)^\top q \leq 0$ for all $i \in [m]$ (Line 10). We underline that the bonus quantity b_t is necessary for the stochastic setting only, that is, when the empirical mean estimation is employed for the constraints. In the adversarial setting, the constraints estimator $\widehat{g}_{t,i}$ is sufficient to attain the desired theoretical guarantees.

Finally, the algorithm employs an online mirror descent (OMD) (Orabona, 2019) update step on the estimated feasible set $\widehat{\Delta}_t(\mathcal{P}_t)$ (Line 11) employing the unnormalized Kullback-Leibler divergence as the Bregman divergence. Formally:

$$B(q\|\widehat{q}_t) = \sum_{x,a,x'} q(x,a,x') \ln \frac{q(x,a,x')}{\widehat{q}_t(x,a,x')} - \sum_{x,a,x'} (q(x,a,x') - \widehat{q}_t(x,a,x')).$$

Remark 7.2 (Algorithmic comparison with Algorithm 6.1)**.** *Algorithm 7.1 employs a completely different approach with respect to the best-of-both-worlds algorithm for CMDPs presented in Chapter 6. Specifically, Algorithm 6.1 is a primal-dual method, where a primal no-regret algorithm optimizes the Lagrange function of the CMDP, while a dual no-regret algorithm selects the most violated constraint. This approach is substantially different since we do not make any use of the Lagrangian formulation of the CMDP. Differently, we resort to a "moving" decision space approach, where we employ a no-regret optimization update over a decision space that adaptively follows the constraints estimation. As we show in Section 7.3, this technique allows us to be particularly effective when the constraints are stochastic. In this case, we have no need for any Slater's like condition, as the constraints are estimated using a UCB-like approach. Crucially, the "moving" decision space still allows us to recover sublinear violation and sublinear α-regret in the adversarial setting.*

7.3 Theoretical Results

In this section, we prove the theoretical guarantees attained by Algorithm 7.1. Specifically, we first discuss the stochastic setting. Then, we show the performance of our algorithm when the constraints are adversarial.

7.3.1 Stochastic Setting

In this section, we focus on the stochastic setting, that is, the constraints are sampled from fixed distributions. The first fundamental result is to show that, in this setting, the bonus terms $b_t(x,a)$ encompass the distance between the constraints estimator and the true constraint function. This is done by means of the following lemma.

Lemma 7.1. *Let $\delta \in (0,1)$. In the stochastic setting, with probability at least $1-11\delta$, it holds that:*

$$|\widehat{g}_{t,i}(x,a) - \bar{g}_i(x,a)| \leq b_t(x,a) \quad \forall (x,a) \in X \times A, i \in [m], t \in [T].$$

Intuitively, the result is proved as follows. We proceed by induction. In the first episodes, $\Gamma_{t,i} = 0$ for all $i \in [m]$. Thus, by Proposition 7.1, the constraint estimator is computed as $\widehat{g}_{t,i} = \frac{1}{N_t(x,a)} \sum_{\tau \in \mathcal{T}_{t,x,a}} g_{\tau,i}(x,a)$, that is, the sample mean of the observed constraints values. Employing a Hoeffding bound, it is easy to see that Lemma 7.1 holds for those specific episodes. The induction step consists in showing that, assuming $\sum_{\tau \in [t-1]} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) \leq \mathcal{C}_{t-1}$ at episode $t-1$, the same holds for the violation observed at t, too. This is done by showing that the empirical mean estimator and the bonus term are sufficient to keep the violation small when the constraints are stochastic. Again, since we proved that $\sum_{\tau \in [t]} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) \leq \mathcal{C}_t$, we have $\Gamma_{t,i} = 0$, which concludes the proof after a simple application of the Hoeffding inequality.

Proof. To get the final result, it is sufficient to prove that for each $t \in [T]$ and $i \in [m]$, it holds:

$$\sum_{\tau \in [t]} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) \leq 21L|X|\sqrt{2t|A| \ln \frac{2mT|X||A|}{\delta}}.$$

Our proof works by induction on t. It is trivial to show the inequality holds for $t = 1$. Indeed,

$$\sum_{x,a} g_{1,i}(x,a)\mathbb{I}_1(x,a) \leq L \leq 21L|X|\sqrt{2|A| \ln \frac{2mT|X||A|}{\delta}}.$$

Assuming that the inequality holds for all $\tau \leq t-1$, we now show that it also holds for t. By definition of $\Gamma_{\tau,i}$, the induction assumption implies that for $\tau \leq t-1$, we have $\beta_{\tau,i}(x,a) = \frac{1}{N_\tau(x,a)}$ for all $(x,a) \in X \times A$, $i \in [m]$. Then, by Proposition 7.1, we have that:

$$\widehat{g}_{\tau,i}(x,a) = \frac{1}{N_\tau(x,a)} \sum_{\widehat{t} \in \mathcal{T}_{\tau,a}} g_{\widehat{t},i}(x,a).$$

Hence, by Lemma E.5, it holds, with probability at least $1-\delta$:

$$|\widehat{g}_{\tau,i}(x,a) - \bar{g}_i(x,a)| \leq \sqrt{\frac{2 \ln \frac{2m|X||A|T}{\delta}}{N_t(x,a)}}, \quad \forall (x,a) \in X \times A, \tau \leq t-1.$$

Assuming that the event above holds, we consider the following inequalities:

$$
\begin{aligned}
&\sum_{\tau\in[t]}\sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) \\
&\qquad \le V_{t,i} + 2L\sqrt{2t\ln\frac{1}{\delta}} && (7.5)\\
&\qquad = V_{t-1,i} + g_{t,i}^\top q_t + 2L\sqrt{2t\ln\frac{1}{\delta}} \\
&\qquad \le \sum_{\tau=1}^{t-1} g_{\tau,i}^\top \widehat{q}_\tau + g_{t,i}^\top q_t + 2L\sqrt{2t\ln\frac{1}{\delta}} \\
&\qquad\quad + 5L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} && (7.6)\\
&\qquad \le \sum_{\tau=1}^{t-1} (g_{\tau,i} - \widehat{g}_{\tau-1,i})^\top \widehat{q}_\tau + \sum_{\tau=1}^{t-1} b_{\tau-1}^\top \widehat{q}_\tau + g_{t,i}^\top q_t + 2L\sqrt{2t\ln\frac{1}{\delta}} \\
&\qquad\quad + 5L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} && (7.7)\\
&\qquad \le \sum_{\tau=1}^{t-1} (g_{\tau,i} - \widehat{g}_{\tau-1,i})^\top \widehat{q}_\tau + 2\sqrt{2|X||A|Lt\ln\frac{2T|X||A|}{\delta}} + g_{t,i}^\top q_t \\
&\qquad\quad + 2L\sqrt{2t\ln\frac{1}{\delta}} + 12L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} && (7.8)\\
&\qquad \le \sum_{\tau=1}^{t-1} (g_{\tau,i} - \widehat{g}_{\tau-1,i})^\top \widehat{q}_\tau + 2\sqrt{2|X||A|Lt\ln\frac{2T|X||A|}{\delta}} + L \\
&\qquad\quad + 2L\sqrt{2t\ln\frac{1}{\delta}} + 12L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} \\
&\qquad \le \sum_{\tau=1}^{t-1} (\bar{g}_i - \widehat{g}_{\tau-1,i})^\top \widehat{q}_\tau + 2\sqrt{2|X||A|Lt\ln\frac{2T|X||A|}{\delta}} + L \\
&\qquad\quad + 12L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} + 4L\sqrt{2t\ln\frac{1}{\delta}} && (7.9)\\
&\qquad \le \sum_{\tau=1}^{t-1} \sum_{x\in X, a\in A} \left(\bar{g}_i(x,a) - \widehat{g}_{\tau-1,i}(x,a)\right)\mathbb{I}_\tau(x,a) \\
&\qquad\quad + 2\sqrt{2|X||A|Lt\ln\frac{2T|X||A|}{\delta}} + L \\
&\qquad\quad + 12L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} + 4L\sqrt{2t\ln\frac{1}{\delta}} && (7.10)\\
&\qquad \le \sqrt{2\ln\frac{2m|X||A|T}{\delta}} \sum_{x\in X, a\in A} \sum_{\tau=1}^{t-1} \frac{1}{\sqrt{N_{\tau-1}(x,a)}}\mathbb{I}_\tau(x,a)
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\sqrt{2|X||A|Lt \ln \frac{2T|X||A|}{\delta}} + L \\
&+ 12L|X|\sqrt{2t|A| \ln \frac{2mT|X||A|}{\delta}} + 4L\sqrt{2t \ln \frac{1}{\delta}} \\
&\leq 2\sqrt{2|X||A|t \ln \frac{2m|X||A|T}{\delta}} + 2\sqrt{2|X||A|Lt \ln \frac{2T|X||A|}{\delta}} + 2L \\
&+ 12L|X|\sqrt{2t|A| \ln \frac{2mT|X||A|}{\delta}} + 4L\sqrt{2t \ln \frac{1}{\delta}} \\
&\leq (4+1+12+4)L|X|\sqrt{2t|A| \ln \frac{2mT|X||A|}{\delta}},
\end{aligned}
$$

where Inequality (7.5) holds by Lemma E.6 with probability $1-\delta$, Inequality (7.6) holds by Lemma E.3 with probability at least $1-2\delta$, Inequality (7.7) holds because $\widehat{q}_{\tau+1} \in \widehat{\Delta}_\tau(\mathcal{P}_\tau)$, Inequality (7.8) holds by Lemma E.2 with probability at least $1-3\delta$, taking $\alpha = \frac{1}{2}$ and $c = \sqrt{2 \ln \left(\frac{2|X||A|T}{\delta}\right)}$, Inequality (7.9) holds by Lemma E.7 with probability at least $1-\delta$, Inequality (7.10) holds by Lemma E.4 with probability at least $1-3\delta$.

Thus $\sum_{\tau\in[t]} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) \leq 21L|X|\sqrt{2t|A| \ln \frac{2mT|X||A|}{\delta}}$, $\Gamma_{t,i} = 0$ and $\widehat{g}_{t,i}(x,a)$ is the empirical mean of past observations. Therefore, by Lemma E.5, we have with probability at least $1-\delta$:

$$
|\widehat{g}_{t,i}(x,a) - \bar{g}_i(x,a)| \leq \sqrt{\frac{2 \ln \left(\frac{2|X||A|mT}{\delta}\right)}{N_t(x,a)}} \quad \forall (x,a) \in X \times A, i \in [m], t \in [T].
$$

A final Union Bound concludes the proof. □

Given Lemma 7.1, the following corollary holds.

Corollary 7.1. *In the stochastic setting, let $\delta \in (0,1)$ and*

$$
\Delta^\star = \left\{ q \in \Delta(M) : \bar{g}_i^\top q \leq 0 \ \forall i \in [m] \right\}.
$$

Then, with probability at least $1-11\delta$, it holds:

$$
\Delta^\star \subseteq \widehat{\Delta}_t(\mathcal{P}_t) \quad \forall t \in [T].
$$

Corollary 7.1 simply states that the true safe decision space is included in the per-episode decision space. This is intuitive, since, by Lemma 7.1, subtracting the bonus term to the constraints estimator allows, with high probability, to be optimistic in the constraints definition. A similar reasoning holds for the transitions. We are now ready to show the main result of the section, that is, the final regret and violation bound. This is done in the following theorem.

Theorem 7.1. *Let $\delta \in (0,1)$. In the stochastic setting, Algorithm 7.1, with $\eta = \gamma =$*

$\sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$, *guarantees that with probability at least* $1-30\delta$*:*

$$R_T \leq 14L|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)},$$

and

$$V_t \leq 61L|X|\sqrt{2t|A|\ln\left(\frac{2mT^2|X||A|}{\delta}\right)}, \quad \forall t \in [T].$$

Theorem 7.1 follows from the following reasoning. As concerns the regret bound, by Corollary 7.1, it holds that the safe optimum is included in the per-episode decision space, with high probability. Thus, following a standard no-regret argument of OMD with implicit exploration shows that Algorithm 7.1 attains sublinear regret with respect to any occupancy that is included in the algorithm decision space at each episode. Differently, to prove the violation, we proceed by contradiction, that is, we show that the weights definition does not allow the violation to exceed the threshold defined by the bound of Theorem 7.1. We remark that the proof for the violation is equivalent to the one for the adversarial setting, since the definition of V_t is equivalent between the two settings. Indeed, in this case, we do not have to exploit Corollary 7.1, since even when $\Gamma_{t,i} = \mathcal{C}_t$, the violations are not allowed to exceed the aforementioned value.

Proof. By Corollary 7.1 with probability at least $1-11\delta$, it holds $\Delta^\star \subseteq \cap_{t\in[T]}\widehat{\Delta}_t(\mathcal{P}_t)$. By Theorem E.1, we have that for any $q \in \bigcap_{t\in[T]}\widehat{\Delta}_t(\mathcal{P}_t)$, with probability at least $1-15\delta$, it holds:

$$\sum_{t\in[T]} r_t^\top(q-q_t) \leq 14L|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}.$$

Let $q^* = \arg\max_{q\in\Delta^\star}\sum_{t=1}^T r_t^\top q$. Then, by Union Bound, we have that with probability at least $1-26\delta$ it holds:

$$\sum_{t\in[T]} r_t^\top(q^*-q_t) \leq 14L|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}.$$

Similarly, with probability at least $1-4\delta$ by Theorem E.4:

$$V_t \leq 61L|X|\sqrt{2t|A|\ln\left(\frac{2mT^2|X||A|}{\delta}\right)}$$

By a Union Bound on all the events, this holds with probability at least $1-30\delta$. This concludes the proof. □

We conclude the section by providing the strong violation bound attained by Algorithm 7.1.

Theorem 7.2. *Let $\delta \in (0,1)$. In the stochastic setting, Algorithm 7.1 guarantees with probability at least $1-16\delta$:*

$$\mathcal{V}_t \leq 18L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}}, \quad \forall t \in [T].$$

Intuitively, Theorem 7.2 is proved by showing that the strong violation attained by Algorithm 7.1 is proportional to the bonus b_t term employed in the decision space definition. Showing that the term concentrates at a $1/\sqrt{T}$ rate concludes the proof.

Proof. Define for each $i \in [m]$ and $t \in [T]$ the following quantity:

$$\mathcal{V}_{t,i} := \sum_{\tau=1}^{t} \left[\bar{g}_i^\top q_\tau\right]^+.$$

Given an $i \in [m]$ and a $t \in [T]$ we have:

$$\begin{aligned}
\mathcal{V}_{t,i} &= \sum_{\tau=1}^{t} \left[\bar{g}_i^\top q_\tau\right]^+ \\
&= \sum_{\tau=1}^{t} \left[(\bar{g}_i - \widehat{g}_{\tau-1,i} + \widehat{g}_{\tau-1,i})^\top q_\tau\right]^+ \\
&= \sum_{\tau=1}^{t} \left[(\bar{g}_i - \widehat{g}_{\tau-1,i})^\top q_\tau + \widehat{g}_{\tau-1,i}^\top q_\tau\right]^+ \\
&= \sum_{\tau=1}^{t} \left[(\bar{g}_i - \widehat{g}_{\tau-1,i})^\top q_\tau + \widehat{g}_{\tau-1,i}^\top q_\tau - \widehat{g}_{\tau-1,i}^\top \widehat{q}_\tau + \widehat{g}_{\tau-1,i}^\top \widehat{q}_\tau\right]^+ \\
&\leq \sum_{\tau=1}^{t} \left[(\bar{g}_i - \widehat{g}_{\tau-1,i})^\top q_\tau + \widehat{g}_{\tau-1,i}^\top q_\tau - \widehat{g}_{\tau-1,i}^\top \widehat{q}_\tau + b_{\tau-1}^\top \widehat{q}_\tau\right]^+ \qquad (7.11) \\
&\leq \sum_{\tau=1}^{t} \left[(\bar{g}_i - \widehat{g}_{\tau-1,i})^\top q_\tau + b_{\tau-1}^\top \widehat{q}_\tau\right]^+ + \|q_\tau - \widehat{q}_\tau\|_1 \\
&\leq \sum_{\tau=1}^{t} \left[(\bar{g}_i - \widehat{g}_{\tau-1,i})^\top q_\tau + b_{\tau-1}^\top \widehat{q}_\tau\right]^+ + 2L|X|\sqrt{2t\ln\frac{2L}{\delta}} \\
&\quad + 3L|X|\sqrt{2t|A|\ln\frac{2T|X||A|}{\delta}} \qquad (7.12) \\
&\leq \sum_{\tau=1}^{t} \left[b_{\tau-1}^\top q_\tau + b_{\tau-1}^\top \widehat{q}_\tau\right]^+ + 5L|X|\sqrt{2t|A|\ln\frac{2T|X||A|}{\delta}}, \qquad (7.13)
\end{aligned}$$

where Inequality (7.11) holds since $\widehat{q}_{\tau+1} \in \widehat{\Delta}_t(\mathcal{P}_t)$, Inequality (7.12) follows from Lemma B.3 of (Rosenberg and Mansour, 2019b) with probability at least $1-2\delta$ and Inequality (7.13) holds by Lemma 7.1 with probability $1-11\delta$ jointly for each i and t. Since $b_t = \sqrt{\frac{2\ln\left(\frac{2m|X||A|T}{\delta}\right)}{N_t(x,a)}}$ by Lemma E.2 with probability at least $1-3\delta$, employing

a Union Bound we have, with probability at least $1 - 16\delta$:

$$\begin{aligned}\mathcal{V}_{t,i} &\leq 2L\sqrt{2T\ln\frac{1}{\delta}} + 4\sqrt{2|X||A|Lt\ln\left(\frac{2mT|X||A|}{\delta}\right)} \\ &\quad + 12L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} \\ &\leq (2+4+12)L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}} \\ &= 18L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}},\end{aligned}$$

for all $i \in [m], t \in [T]$. This concludes the proof. □

We finally remark that the results provided in this section strongly improve the ones provided in Chapter 6 for the stochastic setting, as we highlight in the following. First, Algorithm 7.1 does not rely on any Slater's like condition to attain the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation bounds. Second, Algorithm 7.1 attains the optimal rate for the *strong* constraints violation metric.

7.3.2 Adversarial Setting

In this section, we focus on the adversarial setting, that is, the constraints are allowed to change arbitrarily over episodes. In such a setting, Mannor et al. (2009) showed the impossibility of attaining sublinear regret and violation, simultaneously. Thus, as is standard in the constrained online learning literature (Castiglioni et al., 2022a), we focus on attaining sublinear violation and sublinear α-regret. Similarly to the stochastic setting, we show that the per-episode decision space is well defined. This is done by means of the following theorem.

Theorem 7.3. *In the adversarial setting, let $\delta \in (0,1)$ and $\Delta^\diamond$ be the interpolation of any point $q \in \Delta(M)$ and $q^\diamond$ and let $\rho' = L \cdot \rho$. Formally,*

$$\Delta^\diamond := \frac{L}{L+\rho'}\{q^\diamond\} + \frac{\rho'}{L+\rho'}\Delta(M).$$

Then, with probability at least $1-\delta$, it holds that $\Delta^\diamond \subseteq \widehat{\Delta}_t(\mathcal{P}_t)$ for all $t \in [T]$.

Proof. For each $t \in [T]$, $i \in [m]$, and $(x,a) \in X \times A$, it holds:

$$\widehat{g}_{t,i}(x,a) = \sum_{\tau \in \mathcal{T}_{t,x,a}} w_{t,x,a}(\tau) g_{\tau,i}(x,a),$$

and by the weights definition,

$$\sum_{\tau \in \mathcal{T}_{t,x,a}} w_{t,x,a}(\tau) = 1.$$

Thus notice that, for all $t \in [T]$ and constraint $i \in [m]$, we have:

$$\max_{(x,a)\in\mathcal{Q}(q^\diamond)} \widehat{g}_{t,i}(x,a) q^\diamond(x,a) \leq -\rho,$$

which implies:

$$\widehat{g}_{t,i}^{\top} q^{\diamond} \leq -L \cdot \rho = -\rho'.$$

Moreover, notice that:

$$\widehat{g}_{t,i}^{\top} q \leq L, \quad \forall q \in \Delta(M).$$

Thus, for any $\tilde{q} \in \Delta^{\diamond}$ and $q \in \Delta(M)$, we obtain:

$$\begin{aligned}
\widehat{g}_{t,i}^{\top} \tilde{q} &= \frac{L}{L+\rho'} \widehat{g}_{t,i}^{\top} q^{\diamond} + \frac{\rho'}{L+\rho'} \widehat{g}_{t,i}^{\top} q \\
&\leq \frac{L}{L+\rho'}(-\rho') + \frac{\rho'}{L+\rho'} L \\
&\leq 0,
\end{aligned}$$

that is, $\tilde{q} \in \widehat{\Delta}_t(P)$. As in the stochastic case, the final result follows from noticing that $\Delta(M) \subseteq \Delta(\mathcal{P}_t)$ since, with probability at least $1-\delta$, $P \in \mathcal{P}_t$, by Lemma 4.1 of (Rosenberg and Mansour, 2019b). □

Intuitively, Theorem 7.3 shows that any α-optimum is included in the per-episode decision space, with high probability. The result is proved employing the definition of the weights and the one of the problem specific parameter ρ. We remark that the quantity $\frac{\rho}{1+\rho}$ is equivalent to $\frac{\rho'}{L+\rho'}$, by definition.

We conclude by providing the final result of the paper.

Theorem 7.4. *Let $\delta \in (0,1)$. In the adversarial setting, Algorithm 7.1, with $\eta = \gamma = \sqrt{\frac{L \ln(L|X||A|/\delta)}{T|X||A|}}$, guarantees that with probability at least $1 - 19\delta$:*

$$\alpha\text{-}R_T \leq 14L|X|^2\sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)},$$

and

$$V_t \leq 61L|X|\sqrt{2t|A| \ln\left(\frac{2mT^2|X||A|}{\delta}\right)},$$

for all $t \in [T]$, where $\alpha = \frac{\rho}{1+\rho}$.

Theorem 7.4 is proved employing a similar approach to the one of Theorem 7.1. Specifically, the α-regret follows from noticing that, by Theorem 7.3, the α-optimum is contained in the per-episode decision space. Thus, employing the OMD with implicit exploration theoretical guarantees gives the result. For the violation, the analysis is equivalent to the one of Theorem 7.1.

Proof. It is sufficient to combine Theorem E.1 and Theorem 7.3. Specifically, with probability at least $1 - 15\delta$, for all $\tilde{q} \in \Delta^{\diamond} \subseteq \widehat{\Delta}_t(\mathcal{P}_t)$, we have:

$$\sum_{t \in [T]} r_t^{\top}(\tilde{q} - q_t) \leq 14L|X|^2\sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)}.$$

Let $q^\dagger = \arg\max_{q \in \Delta(M)} \sum_{t=1}^{T} r_t^\top q$. We observe that:

$$\bar{q} = \frac{L}{L+\rho'} q^\diamond + \frac{\rho'}{L+\rho'} q^\dagger \in \Delta^\diamond.$$

Thus, it holds:

$$\sum_{t=1}^{T} r_t^\top \bar{q} = \sum_{t=1}^{T} r_t^\top \left(\frac{L}{L+\rho'} q^\diamond + \frac{\rho'}{L+\rho'} q^\dagger \right) \geq \frac{\rho'}{L+\rho'} \sum_{t=0}^{T} r_t^\top q^\dagger.$$

This proves that with probability at least $1 - 15\delta$:

$$\left(\frac{\rho'}{L+\rho'} \right) \text{-}R_T \leq 14L|X|^2 \sqrt{2T|A| \ln \left(\frac{T|X|^2|A|}{\delta} \right)}.$$

Similarly to the stochastic case, employing Theorem E.4, with probability at least $1-4\delta$, we have:

$$V_t \leq 61L|X| \sqrt{2t|A| \ln \left(\frac{2mT^2|X||A|}{\delta} \right)}.$$

By Union Bound, this holds with probability $1 - 19\delta$. Noticing that $\frac{\rho}{1+\rho'} = \frac{\rho'}{L+\rho}$concludes the proof. □

Comparing the theoretical guarantees of Algorithm 7.1 and the ones provided in Chapter 6, the following remarks are in order. First, the violation bound provided by Algorithm 7.1 neither relies on the Slater's condition nor has any dependence on the Slater's parameter. Second, while in this case we employ a slightly stronger ρ definition, our α-regret is computed with respect to the *unconstrained* optimum, rather than the constrained one. Moreover, our bound does not rely on the Slater's parameter of the problem, whereas only the definition of α-regret does.

Part IV

Non-Stationary Rewards and Constraints

CHAPTER 8

A Primal-Dual Approach

In this chapter, we make a preliminary step towards the fourth part of this dissertation, extending the theoretical results provided in Chapter 6 to a non-stationary setting. Specifically, we show how Algorithm 6.1 handles adversarial constraints—attaining both sublinear regret and sublinear violation—when the adversariality of the *constraints* is bounded.

8.1 Setting and Additional Notation

The setting and the notation follow exactly the ones of Chapter 6. In this chapter, we will assume that the adversariality of the *constraints* is limited. Specifically, we parametrize the adversariality of the constraints given the following measure of non-stationarity, *i.e.*, the distance between the actual constraints selected by the adversary and the closest fixed constraints matrix. This is defined as follows.

Definition 8.1. *We define a measure of adversariality as:*

$$\mathcal{B} \coloneqq \min_{G \in [-1,1]^{|\mathcal{X}\times A|\times m}} \sum_{t=1}^{T} \|G_t - G\|_1.$$

Similarly to Chapter 6, we refer to Section D.1 for the dictionary of the definition of different quantities which will be employed in the rest of the chapter.

8.2 Theoretical Results

In this section, we present the theoretical guarantees of Algorithm 6.1 when the adversariality of the constraints is limited.

We first show the performance of our algorithm when Condition 6.1 holds.

Theorem 8.1. *Suppose that Condition 6.1 holds, the constraints are generated adversarially and are parameterized given $\mathcal{B}$. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains the following bounds:*

$$R_T \leq \widetilde{\mathcal{O}}\left(\Lambda(\sqrt{T}+\mathcal{B})\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(\Lambda\sqrt{T}\right),$$

with probability at least $1-14\delta$ when the rewards are stochastic, and with probability at least $1-13\delta$ when the rewards are adversarial, where $\Lambda = \frac{112mL^2}{\rho^2}$.

Proof. In the following, we will refer as $\widehat{G}$ to the constraint matrix that minimize the definition of $\mathcal{B}$, that is $\widehat{G} := \arg\min_{G\in[-1,1]^{|X\times A|\times m}} \sum_{t=1}^{T}\|G_t - G\|_1$.
Analogously to Theorem 6.3, it holds with probability at least $1-12\delta$:

$$V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

Then, with probability at least $1-12\delta$ we observe that:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t} \\
&= \sum_{t=1}^{T}\left(r_t^\top q^* - \lambda_t^\top G_t^\top q^*\right) - \sum_{t=1}^{T}\left(r_t^\top q_t - \lambda_t^\top G_t^\top q_t\right) + \sum_{t=1}^{T}\lambda_t^\top G_t^\top (q^* - q_t) \\
&\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\lambda_t^\top G_t^\top q^* \qquad (8.1a)\\
&= \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + \sum_{t=1}^{T}\lambda_t^\top (G_t - \overline{G})^\top q^* + \sum_{t=1}^{T}\lambda_t^\top \overline{G}^\top q^* \\
&\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + 2\lambda_{1,T}\mathcal{B} \qquad (8.1b)\\
&\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + 2\Lambda\mathcal{B}, \qquad (8.1c)
\end{aligned}$$

where Inequality (8.1a) holds by Theorem 6.1 and by Theorem D.2, Inequality (8.1b) holds since in the non-stationary constraint case $\sum_{t=1}^{T}(G_t - \overline{G})^\top q^* \leq \sum_{t=1}^{T}\|G_t - \overline{G}\|_1 = \sum_{t=1}^{T}\|G_t - \widehat{G}\|_1 + \sum_{t=1}^{T}\|\widehat{G} - \frac{1}{T}\sum_{t=1}^{T} G_t\|_1 \leq 2\mathcal{B}$ and finally Inequality (8.1c) holds by Theorem 6.2. Finally, we observe that in the stochastic rewards case, it holds, with probability at least $1-\delta$:

$$\left(T\cdot \mathrm{OPT}_{\overline{r},\overline{G}} - \sum_{t=1}^{T} r_t^\top q_t\right) - \sum_{t=1}^{T} r_t^\top (q^* - q_t) \leq \mathcal{E}^r.$$

Thus, if the rewards are stochastic, with probability at least $1-14\delta$, it holds:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + 2\Lambda\mathcal{B} + \mathcal{E}^r, \quad V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

and if the rewards are adversarial, with probability at least $1-13\delta$ it holds:

$$R_T \leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \Lambda\mathcal{E}^{\mathbb{I}} + 2\Lambda\mathcal{B}, \quad V_T \leq \frac{1}{\eta}\Lambda + \mathcal{E}^{\mathbb{I}},$$

which concludes the proof. □

We conclude by showing the performance of our algorithm when ρ can be arbitrarily small.

Theorem 8.2. *Suppose that Condition 6.1 does not hold, the constraints are generated adversarially and are parameterized given $\mathcal{B}$. Then, for any $\delta \in (0,1)$, Algorithm 6.1 attains:*

$$R_T \leq \widetilde{\mathcal{O}}\left(T^{1/4}(\sqrt{T}+\mathcal{B})\right), \quad V_T \leq \widetilde{\mathcal{O}}\left(T^{3/4}\right),$$

with probability at least $1-11\delta$ when the rewards are stochastic, and with probability at least $1-10\delta$ when the rewards are adversarial.

Proof. Similar to the proof of Theorem 8.1, it holds with probability at least $1-10\delta$:

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q^* - \sum_{t=1}^{T} r_t^\top q^{P,\pi_t} \\
&= \sum_{t=1}^{T} \left(r_t^\top q^* - \lambda_t^\top G_t^\top q^*\right) - \sum_{t=1}^{T} \left(r_t^\top q_t - \lambda_t^\top G_t^\top q_t\right) + \sum_{t=1}^{T} \lambda_t^\top G_t^\top (q^* - q_t) \\
&\leq \mathcal{E}^P + \mathcal{E}^D(\underline{0}) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} + 2\lambda_{1,T}\mathcal{B}.
\end{aligned}$$

Therefore, with probability at least $1-10\delta$ by following the reasoning of Lemma D.5 to bound the dual decision space, it holds, when the rewards are adversarial:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - mT^{1/4}\mathcal{E}^{\mathbb{I}} - 2mT^{1/4}\mathcal{B} - \mathcal{E}^D(\underline{0}) - \mathcal{E}^P,$$

and, when the rewards are stochastic, with probability ta least $1-11\delta$:

$$\sum_{t=1}^{T} r_t^\top q_t \geq T \cdot \mathrm{OPT}_{\overline{r},\overline{G}} - mT^{1/4}\mathcal{E}^{\mathbb{I}} - 2mT^{1/4}\mathcal{B} - \mathcal{E}^D(\underline{0}) - \mathcal{E}^P - \mathcal{E}^r.$$

Applying Lemma D.5 to bound the constraints violation concludes the proof. □

CHAPTER 9

A Meta Procedure to Handle Strong Violation

In this chapter, we study *online learning* problems in *episodic* CMDPs, under *bandit feedback*. As pointed out throughout this dissertation, a crucial feature distinguishing online learning problems in CMDPs is whether rewards and constraints are selected *stochastically* or *adversarially*. Indeed, most of the works in the literature focus on the case in which constraints are stochastic (see, *e.g.*, (Wei et al., 2018; Zheng and Ratliff, 2020; Efroni et al., 2020; Qiu et al., 2020; Liu et al., 2021; Bai et al., 2023)), while the main exception is provided by this dissertation (Chapters 5 - 6 - 7). This is primarily motivated by the well-known impossibility result by Mannor et al. (2009), which prevents any learning algorithm from attaining both sublinear regret and sublinear constraint violation, when competing against a best-in-hindsight policy that satisfies the constraints *on average*.

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

We propose algorithms that attain $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *strong* constraint violation, where C denotes the corruption of the setting. We remark that $C = \Theta(T)$ in the worst case, and, thus, our bounds are coherent with the impossibility result by Mannor et al. (2009). Moreover, in stochastic CMDPs, our bounds reduce to state-of-the-art $\widetilde{\mathcal{O}}(\sqrt{T})$

bounds (Efroni et al., 2020). Notably, our algorithms work under *bandit* feedback, namely by only observing rewards and constraint costs of the state-action pairs visited during episodes. Moreover, they are able to manage *strong* constraint violation—these guarantees cannot be attained employing the technique provided in Chapter 8—. This means that they do *not* allow for a negative violation (*i.e.*, a constraint satisfaction) to cancel out a positive one across different episodes. This is a crucial for most of the practical applications. For instance, in autonomous driving, avoiding a collision does *not* "repair" a previous crash.

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

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

9.1 Additional Comparison with Related Works

The work provided in this chapter is also closely related to *corruption-robust* online learning, which, while well-established in different settings, such as unconstrained MDPs with corrupted transitions, *remains largely unexplored for CMDPs*. Specifically, Lykouris et al. (2021) are the first to establish sublinear regret guarantees for MDPs with corrupted rewards and transitions under bandit feedback, achieving $\widetilde{O}(C\sqrt{T})$ regret without requiring prior knowledge of C. Chen et al. (2021) are the first to provide a regret bound that additively depends on C, namely of the order of $\widetilde{O}(\sqrt{T}+C^2)$. This result is improved by Wei et al. (2022a), who show a regret bound of order $\widetilde{O}(\sqrt{T}+C)$ under the same conditions. Finally, very recently Jin et al. (2023) study MDPs with adversarial rewards and corrupted transitions, under bandit feedback and unknown corruption value, attaining regret $\widetilde{O}(\sqrt{T}+C^P)$, where C^P is the corruption associated with the transitions. While the techniques employed in these works share some similarities with ours, they *cannot* be easily extended to CMDPs. This is because CMDPs involve a dual objective: minimizing the regret while ensuring low constraint violation. This cannot be achieved through *standard* corralling techniques that are commonly used in the corruption-robust online learning (Agarwal et al., 2017), as these are designed to deal with single-objective settings.

Finally, there is also a related literature that focuses on studying dynamic regret in *non-stationary* CMDPs. While in such settings the learner-environment interaction closely resembles ours, the performance metrics are different from ours and *not* easily comparable. Specifically, Ding and Lavaei (2023) and Wei et al. (2023) consider the case in which rewards and constraints are non-stationary, assuming that their variation is bounded. Our

work differs from theirs in multiple aspects. First, we consider *strong* constraint violation, while they allow for cancellations. As concerns the definition of regret, ours and that by Ding and Lavaei (2023) and Wei et al. (2023) are *not* comparable. Indeed, they employ a dynamic regret baseline, which, in general, is harder than the static regret employed in our work. However, they compare learner's performances against a dynamic policy that satisfies the constraints at every round. Instead, we consider a policy that satisfies the constraints *on average*, which can perform arbitrarily better than a policy satisfying the constraints at every round. Furthermore, the dependence on T in their regret bound is much worse than ours, even when the non-stationarity is small, namely when it is a constant independent of T and does *not* affect our regret bound. Finally, we do *not* make any assumption on T, while the bounds in (Wei et al., 2023) only hold for large T.

9.2 Setting and Additional Notation

In this chapter, we study *online learning* in CMDPs, under *bandit feedback*, where the rewards and constraints are sampled from *non-stationary* distributions at each episode. Specifically, we consider a setting in which the sequences of probability distributions $\{\mathcal{R}_t\}_{t=1}^T$ and $\{\mathcal{G}_t\}_{t=1}^T$ are selected *adversarially*. Thus, reward vectors r_t and constraint cost matrices G_t are random variables whose distributions are allowed to change arbitrarily from episode to episode. In other terms, they exhibit non-stationarity.

To measure how much such probability distributions change over the episodes, we introduce the notion of *(adversarial) corruption*. In particular, we define the adversarial corruption C_r for the rewards as:

$$C_r := \min_{r \in [0,1]^{|X \times A|}} \sum_{t \in [T]} \|\mathbb{E}[r_t] - r\|_1 . \tag{9.1}$$

Intuitively, the corruption C_r encodes the sum over all episodes of the distances between the means $\mathbb{E}[r_t]$ of the adversarial distributions $\mathcal{R}_t$ and a "fictitious" non-corrupted reward vector r. Notice that a similar notion of corruption has been employed in unconstrained MDPs to measure the non-stationarity of transition probabilities; see (Jin et al., 2023). 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:

$$C_G := \min_{G \in [0,1]^{|X \times A| \times m}} \sum_{t \in [T]} \max_{i \in [m]} \|\mathbb{E}[g_{t,i}] - g_i\|_1, \tag{9.2}$$

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

Now, we restate the notion of *cumulative regret* and *cumulative strong constraint violation*, which are the performance metrics used to evaluate algorithms, in the *non-stationary* setting. Specifically, the regret over T episodes is defined as:

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

where $\overline{r} := \frac{1}{T} \sum_{t=1}^T \mathbb{E}[r_t]$ and $\overline{G} := \frac{1}{T} \sum_{t=1}^T \mathbb{E}[G_t]$. In the following, we denote by q^* an occupancy measure solving Program (2.3) instantiated with $\overline{r}$, $\overline{G}$, and θ, while its

corresponding policy is π^*. Thus, $\text{OPT}_{\overline{r},\overline{G},\theta} = \overline{r}^\top q^*$ and the regret can be written as $R_T := \sum_{t=1}^T \mathbb{E}[r_t]^\top (q^* - q^{P,\pi_t})$. Furthermore, the cumulative *strong* constraint violation over T episodes is defined as follows:

$$\mathcal{V}_T := \max_{i \in [m]} \sum_{t \in [T]} \left[\mathbb{E}[G_t]^\top q^{P,\pi_t} - \theta \right]_i^+, \text{ where we let } [\cdot]^+ := \max\{0, \cdot\}.$$

Remark 9.1 (Relation with adversarial/stochastic CMDPs)**.** *Our setting naturally encompasses both stochastic and adversarial CMDPs. Indeed, if the distributions $\mathcal{R}_t$ and $\mathcal{G}_t$ do not change over the episodes, then we recover a CMDP with stochastic rewards and constraints. Moreover, when the supports of $\mathcal{R}_t$ and $\mathcal{G}_t$ are singletons (and, thus, mean values are fully revealed), our setting reduces to a CMDP with adversarial rewards and constraints, since $\mathcal{R}_t$ and $\mathcal{G}_t$ are selected adversarially.*

Remark 9.2 (Impossibility results carrying over from adversarial CMDPs)**.** *Mannor et al. (2009) show that, in online learning problems with constraints selected adversarially, it is impossible to achieve both regret and constraint violation growing sublinearly in T. This result holds for a regret definition that corresponds to ours. Thus, it carries over to our setting. This is why we look for algorithms whose regret and strong constraint violation scale as $\widetilde{\mathcal{O}}(\sqrt{T} + C)$, with a linear dependency on the adversarial corruption C. Notice that the impossibility result by Mannor et al. (2009) does not rule out the possibility of achieving such a guarantee, since regret and strong constraint violation are not sublinear when C grows linearly in T, as it could be the case in a classical adversarial setting.*

We refer to Appendix F.1 for the definition of the events employed in the rest of the chapter. Similarly, we refer to Appendix F.5.1 for the additional notation employed in the second part of the chapter.

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

We start studying the case in which the learner *knows* the adversarial corruption C. We propose an algorithm (called `NS-SOPS`, see Algorithm 9.1), which adopts a suitably-designed UCB-like approach encompassing the adversarial corruption C in the confidence bounds of rewards and constraint costs. This effectively results in "boosting" the *optimism* of the algorithm, and it allows to achieve regret and strong constraint violation of the order of $\widetilde{\mathcal{O}}(\sqrt{T} + C)$. The `NS-SOPS` algorithm is a crucial building block to deal with the case in which C is *not* known, as we show in the following section.

9.3.1 `NS-SOPS`: Non-Stationary Safe Optimistic Policy Search

Algorithm 9.1 provides the pseudocode of the *non-stationary safe optimistic policy search* (`NS-SOPS`) algorithm. The algorithm keeps track of suitably-defined confidence bounds for transitions, rewards, and constraint costs. At each episode $t \in [T]$, the algorithm builds a confidence set $\mathcal{P}_t$ for the transition function P by following the same approach as Jin et al. (2020a) (see Appendix A.5 for its definition). Instead, for rewards and constraint costs, the algorithm adopts novel *enlarged* confidence bounds, which are suitably designed to tackle non-stationarity.

Given any confidence parameter $\delta \in (0, 1)$, by letting $N_t(x, a)$ be the total number of visits to the state-action pair $(x, a) \in X \times A$ up to episode t, the confidence bound for the

reward $r_t(x,a)$ is:

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

while the bound for the constraint cost $g_{t,i}(x,a)$ is

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

Intuitively, the first term in the expressions above is derived from Azuma-Hoeffding inequality, the second term allows to deal with the non-stationarity of rewards and constraint costs, while the third term is needed to bound how much the averages $\overline{r}$ and $[\overline{G}]_i$ differ from their "fictitious" non-corrupted counterparts r° and $[G^\circ]_i$, respectively. Algorithm 9.1 also computes empirical rewards and constraint costs. At each episode $t \in [T]$, for any state-action pair $(x,a) \in X \times A$ and constraint $i \in [m]$, such estimates are defined as $\widehat{r}_t(x,a) \coloneqq \frac{\sum_{\tau\in[t]} \mathbb{I}_\tau(x,a) r_\tau(x,a)}{\max\{N_t(x,a),1\}}$ and $\widehat{g}_{t,i}(x,a) \coloneqq \frac{\sum_{\tau\in[t]} \mathbb{I}_\tau(x,a) g_{\tau,i}(x,a)}{\max\{N_t(x,a),1\}}$, where $\mathbb{I}_\tau(x,a) = 1$ if and only if the pair (x,a) is visited during episode τ, while $\mathbb{I}_\tau(x,a) = 0$ otherwise. For ease of notation, we let $\widehat{G}_t \in [0,1]^{|X\times A|\times m}$ be the matrix with components $\widehat{g}_{t,i}(x,a)$. We refer to Appendix F.2 for a detailed treatment of all the results related to confidence bounds.

Algorithm 9.1 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 2.1 and observe *bandit* feedback from interaction
  Compute $\mathcal{P}_t$, $\overline{r}_t$, and $\underline{G}_t$
  $q \leftarrow$ solution to $\texttt{OPT-CB}_{\Delta(\mathcal{P}_t),\overline{r}_t,\underline{G}_t,\theta}$
  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 9.1 selects policies with an UCB-like approach encompassing *optimism* in both rewards and constraints satisfaction, following an approach similar to that employed by Efroni et al. (2020). Specifically, at each episode $t \in [T]$ and for any state-action pair $(x,a) \in X \times A$, the algorithm employs an *upper* confidence bound for the reward $r_t(x,a)$, defined as $\overline{r}_t(x,a) \coloneqq \widehat{r}_t(x,a) + \phi_t(x,a)$, while it uses *lower* confidence bounds for the constraint costs $g_{t,i}(x,a)$, defined as $\underline{g}_{t,i}(x,a) \coloneqq \widehat{g}_{t,i}(x,a) - \xi_t(x,a)$ for every constraint $i \in [m]$. Then, by letting $\overline{r}_t \in [0,1]^{|X\times A|}$ be the vector with components $\overline{r}_t(x,a)$ and $\underline{G}_t$ be the matrix with entries $\underline{g}_{t,i}(x,a)$, Algorithm 9.1 chooses the policy to be employed in

the next episode $t+1$ by solving the following linear program:

$$\texttt{OPT-CB}_{\Delta(\mathcal{P}_t),\overline{r}_t,\underline{G}_t,\theta} \coloneqq \begin{cases} \arg\max_{q\in\Delta(\mathcal{P}_t)} & \overline{r}_t^\top q \quad \text{s.t.} \\ & \underline{G}_t^\top q \leq \theta, \end{cases} \tag{9.3}$$

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

9.3.2 Theoretical Guarantees of NS-SOPS

Next, we prove the theoretical guarantees attained by Algorithm 9.1 (see Appendix F.3 for complete proofs of the theorems and associated lemmas). First, we analyze the strong cumulative violation incurred by the algorithm. Formally, we can state the following result.

Theorem 9.1. *Given any $\delta \in (0,1)$, with probability at least $1-8\delta$, Algorithm 9.1 attains the following strong violation bound:*

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

Intuitively, Theorem 9.1 is proved by showing that every constraint-satisfying occupancy measure is also feasible for Program (9.3) with high probability. This holds since Program (9.3) employs lower confidence bounds for constraint costs. Thus, in order to bound $\mathcal{V}_T$, it is sufficient to analyze at which rate the feasible region of Program (9.3) concentrates to the *true* one (*i.e.*, the one defined by $\overline{G}$ in Program (2.3)). Since by definition of $\xi_t(x,a)$ the feasibility region of Program (9.3) concentrates as $1/\sqrt{t} + C/t$, the resulting bound for the strong violation $\mathcal{V}_T$ is of the order of $\widetilde{\mathcal{O}}(\sqrt{T}+C)$.

Proof. In the following, we will refer as $\mathcal{E}_{\widehat{q}}$ to the event described in Lemma A.9, which holds with probability at least $1-6\delta$. Thus, under $\mathcal{E}_G \cap \mathcal{E}_{\widehat{q}}$, the linear program solved by Algorithm 9.1 has a feasible solution (see Lemma F.8) and it holds:

$$\begin{aligned}
\mathcal{V}_T &= \max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[G_t]^\top q_t - \theta\right]_i^+ \\
&= \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + g_i^{\circ\top} q_t - \theta_i\right]^+ \\
&\leq \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + \left(\underline{g}_{t-1,i} + 2\xi_{t-1}\right)^\top q_t - \theta_i\right]^+ && \text{(9.4a)} \\
&= \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + \underline{g}_{t-1,i}^\top \widehat{q}_t + 2\xi_{t-1}^\top q_t - \theta_i\right]^+ \\
&\leq \max_{i\in[m]} \sum_{t\in[T]} \left[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + 2\xi_{t-1}^\top q_t\right]^+ && \text{(9.4b)} \\
&\leq \max_{i\in[m]} \sum_{t\in[T]} \left|\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t\right| + 2\max_{i\in[m]} \sum_{t\in[T]} \left|\xi_{t-1}^\top q_t\right|
\end{aligned}$$

$$+ \max_{i\in[m]} \sum_{t\in[T]} \left| \underline{g}_{t-1,i}^{\top} (q_t - \widehat{q}_t) \right| \tag{9.4c}$$

$$\leq \max_{i\in[m]} \sum_{t\in[T]} \|\mathbb{E}[g_{t,i}] - g_i^{\circ}\|_1 + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^{\top} q_t + \max_{i\in[m]} \sum_{t\in[T]} \|q_t - \widehat{q}_t\|_1 \tag{9.4d}$$

$$\leq C_G + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^{\top} q_t + \sum_{t\in[T]} \|q_t - \widehat{q}_t\|_1, \tag{9.4e}$$

where Inequality (9.4a) follows from Corollary F.2, Inequality (9.4b) holds since Algorithm 9.1 ensures, for all $t \in [T]$ and for all $i \in [m]$, that $\underline{g}_{t,i}^{\top} \widehat{q}_t \leq \theta_i$, Inequality (9.4c) holds since $[a+b]^+ \leq |a| + |b|$, for all $a, b \in \mathbb{R}$, Inequality (9.4d) follows from Hölder inequality since $||\underline{g}_{t,i}(x,a)||_\infty \leq 1$ and $||q_t(x,a)||_\infty \leq 1$, and finally Equation (9.4e) holds for the definition of C_G.
To bound the last term of Equation (9.4e), we notice that, under $\mathcal{E}_{\widehat{q}}$, by Lemma A.9, it holds:

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

To bound the second term of Equation (9.4e) we proceed as follows. Under $\mathcal{E}_{\widehat{q}}$,with probability at least $1-\delta$, it holds:

$$\sum_{t\in[T]} \xi_{t-1}^{\top} q_t \leq \sum_{t\in[T]} \sum_{x,a} \xi_{t-1}(x,a) \mathbb{I}_t(x,a) + L\sqrt{2T \ln \frac{1}{\delta}} \tag{9.5a}$$

$$\leq \sum_{x,a} \sum_{t\in[T]} \mathbb{I}_t(x,a) \left(\sqrt{\frac{1}{2\max\{N_{t-1}(x,a),1\}} \ln\left(\frac{2mT|X||A|}{\delta}\right)} + \frac{C_G}{\max\{N_{t-1}(x,a),1\}} + \frac{C_G}{T} \right) + L\sqrt{2T \ln \frac{1}{\delta}} \tag{9.5b}$$

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

$$\leq 3\sqrt{\frac{1}{2}|X||A|LT \ln\left(\frac{2mT|X||A|}{\delta}\right)} + |X||A|(2+\ln(T))C_G + |X||A|C_G + L\sqrt{2T \ln \frac{1}{\delta}} \tag{9.5c}$$

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

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

where Inequality (9.5a) follows from the Azuma-Hoeffding inequality and noticing that $\sum_{x,a}\xi_{t-1}(x,a)q_t(x,a) \leq L$, Equality (9.5b) follows from the definition of ξ_t and finally, Inequality (9.5c) holds since $1 + \sum_{t=1}^{N_T(x,a)}\sqrt{\frac{1}{t}} \leq 1 + 2\sqrt{N_T(x,a)} \leq 3\sqrt{N_T(x,a)}$, since $1 + \sum_{t=1}^{N_T(x,a)}\frac{1}{t} \leq 2 + \ln(T)$ and by Cauchy-Schwarz inequality. Finally, we notice that the intersection event $\mathcal{E}_G \cap \mathcal{E}_{\widehat{q}} \cap \mathcal{E}_{\text{Azuma}}$ holds with the following probability,

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

Noticing that, by Corollary F.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. □

The regret guaranteed by Algorithm 9.1 is formalized by the following theorem.

Theorem 9.2. *Given any $\delta \in (0,1)$, with probability at least $1 - 9\delta$, Algorithm 9.1 attains the following regret bound:*

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

Theorem 9.2 is proved similarly to Theorem 9.1. Indeed, since every constraint-satisfying occupancy measure is feasible for Program (9.3) with high probability, this also holds for q^*, as it satisfies constraints by definition. Thus, since by definition of $\phi_t(x,a)$ the upper confidence bound for the rewards maximized by Program (9.3) concentrates as $1/\sqrt{t} + C/t$, the regret bound follows.

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

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

Let's observe that, by Lemma F.8, under the event $\mathcal{E}_G \cap \mathcal{E}_P$, $\widehat{q}_t$ is optimal solution for $\overline{r}_{t-1}$ in $\left\{q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \leq \theta\right\}$. Thus, under $\mathcal{E}_G \cap \mathcal{E}_P$ the optimal feasible solution q^* is such that:

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

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

$$\begin{aligned}\frac{1}{T}\sum_{t\in[T]}\mathbb{E}[r_t]^\top q^* &\leq \overline{r}_{t-1}^\top q^* \\ &\leq \overline{r}_{t-1}^\top \widehat{q}_t\end{aligned}$$

$$\leq \left(\frac{1}{T} \sum_{t \in [T]} \mathbb{E}[r_t] + 2\phi_{t-1} \right)^\top \widehat{q}_t.$$

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

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

By Lemma A.9 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:

$$\sum_{t \in [T]} \phi_{t-1}^\top q_t \leq \sum_{t \in [T]} \sum_{x,a} \phi_{t-1}(x,a) \mathbb{I}_t(x,a) + L\sqrt{2T \ln \frac{1}{\delta}} \tag{9.6a}$$

$$\begin{aligned}
&\leq \sum_{x,a} \sum_{t \in [T]} \mathbb{I}_t(x,a) \left(\sqrt{\frac{1}{2 \max\{N_{t-1}(x,a), 1\}} \ln\left(\frac{2T|X||A|}{\delta} \right)} \right. \\
&\quad \left. + \frac{C_r}{\max\{N_{t-1}(x,a), 1\}} + \frac{C_r}{T} \right) + L\sqrt{2T \ln \frac{1}{\delta}}
\end{aligned} \tag{9.6b}$$

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

$$
\begin{aligned}
&\leq 3\sqrt{\frac{1}{2}|X||A|LT\ln\left(\frac{2T|X||A|}{\delta}\right)} + |X||A|(2+\ln(T))C_r \\
&\quad + |X||A|C_r + L\sqrt{2T\ln\frac{1}{\delta}} \qquad (9.6c)\\
&\leq 3\sqrt{\frac{1}{2}|X||A|LT\ln\left(\frac{2T|X||A|}{\delta}\right)} + (3+\ln(T))|X||A|C_r \\
&\quad + L\sqrt{2T\ln\frac{1}{\delta}} \\
&= \mathcal{O}\left(\sqrt{|X||A|LT\ln\left(\frac{T|X||A|}{\delta}\right)} + \ln(T)|X||A|C_r\right),
\end{aligned}
$$

where Inequality (9.6a) follows from Azuma-Hoeffding inequality, Equality (9.6b) holds for the definition of ϕ_t, and Inequality (9.6c) holds since $1+\sum_{t=1}^{N_T(x,a)}\sqrt{\frac{1}{t}} \leq 1+2\sqrt{N_T(x,a)} \leq 3\sqrt{N_T(x,a)}$, $1+\sum_{t=1}^{N_T(x,a)}\frac{1}{t} \leq 2+\ln(T)$ and by Cauchy-Schwarz inequality. Thus, we observe that with probability at least $1-9\delta$ it holds:

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

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

Remark 9.3 (What if some under/overestimate of C is available)**.** *We also study what happens if the learner runs Algorithm 9.1 with an under/overestimate on the adversarial corruption as input. We defer to Appendix F.4 all the technical results related to this analysis. In particular, it is possible to show that any underestimate on C does not detriment the bound on $\mathcal{V}_T$, which remains the one in Theorem 9.1. On the other hand, an overestimate on C, say $\widehat{C} > C$, results in a bound on $\mathcal{V}_T$ of the order of $\widetilde{\mathcal{O}}(\sqrt{T}+\widehat{C})$, which is worse than the one in Theorem 9.1. Intuitively, this is because using an overestimate makes Algorithm 9.1 too conservative. As a result, one could be tempted to conclude that running Algorithm 9.1 with an underestimate of C as input is satisfactory when the true value of C is unknown. However, this would lead to a regret R_T growing linearly in T, since, intuitively, a regret-minimizing policy could be cut off from the algorithm decision space. This motivates the introduction of additional tools to deal with the case in which C is unknown, as we do in Section 9.4.*

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

In this section, we go beyond Section 9.3 by studying the more relevant case in which the learner does *not* know the value of the adversarial corruption C. In order to tackle this challenging scenario, we develop a *meta-procedure* (called `Lag-FTRL`, see Algorithm 9.2) that instantiates multiple instances of an algorithm working for the case in which C is known, with each instance taking care of a different "guess" on the value

Algorithm 9.2 Lagrangified Follow-The-Regularized-Leader (`Lag-FTRL`)

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

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

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

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

  $\texttt{Alg}^j \leftarrow$ stabilized Algorithm 9.1 with $C = 2^j$

end for

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

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

  Sample index $j_t \sim w_t$

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

  Choose policy $\pi_t^{j_t}$ in Algorithm 2.1 and observe *bandit* feedback from interaction

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

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

    Build $\ell_{t,j}$ as in Equation (9.7)

    Build $b_{t,j}$ as in Equation (9.8)

  end for

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

end for

of C. The `Lag-FTRL` algorithm is inspired by the work of Agarwal et al. (2017) in the context of classical (unconstrained) multi-armed bandit problems. Let us remark that standard "coralling" techniques, such as the one proposed by Agarwal et al. (2017), cannot be easily generalized to our setting, since our objective is twofold: minimizing the regret while simultaneously ensuring small constraint violation. In this section, to deal with our non-stationary CMDP setting, we let `Lag-FTRL` instantiate multiple instances of the `NS-SOPS` algorithm in Section 9.3.

9.4.1 `Lag-FTRL`: Lagrangified FTRL

At a high level, the *Lagrangified follow-the-regularized-leader* (`Lag-FTRL` for short) algorithm works by instantiating several instances of Algorithm 9.1, suitably stabilized (see section F.7), with each instance $\texttt{Alg}^j$ being run for a different "guess" of the (unknown) adversarial corruption value C. The algorithm plays the role of a *master* by choosing which instance $\texttt{Alg}^j$ to use at each episode. The selection is done by employing an FTRL approach with a suitable log-barrier regularization. In particular, at each episode $t \in [T]$, by letting $\texttt{Alg}^{j_t}$ be the selected instance, the `Lag-FTRL` algorithm employs the policy $\pi_t^{j_i}$ prescribed by $\texttt{Alg}^{j_t}$ and provides feedback to instance $\texttt{Alg}^{j_t}$ only.

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

the FTRL procedure with losses constructed starting from the Lagrangian of the offline optimization problem in Program (2.3), and suitably modified to manage *strong* violation.

The pseudocode of the `Lag-FTRL` algorithm is provided in Algorithm 9.2. At Line 4, it instantiates $M := \lceil \log_2 T \rceil$ instances of Algorithm 9.1, with each instance $\texttt{Alg}^j$, for $j \in [M]$, receiving as input a "guess" on the adversarial corruption $C = 2^j$. Notice that, to every instance of Algorithm 9.1, a standard doubling trick and a stabilization procedure is applied (see Algorithm F.1 for additional details). This modification to Algorithm 9.1 is necessary to guarantee that each instance j attains a regret and strong cumulative constraints violation which linearly degrade in $\nu_{T,j} = {}^1/_{\min_{t \in [T]} w_{t,j}}$ and C, when employed by the master algorithm. The algorithm assigns weights defining a probability distribution to instances $\texttt{Alg}^j$, with $w_{t,j} \in [0,1]$ denoting the weight of instance $\texttt{Alg}^j$ at episode $t \in [T]$. We denote by $w_t \in \Delta_M$ the weight vector at episode t, with Δ_M being the $M-1$-dimensional simplex. At the first episode, all the weights $w_{1,j}$ are initialized to the value ${}^1/_M$ (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 2.1 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 to be fed into each instance $\texttt{Alg}^j$. In particular, at episode $t \in [T]$ and for every $j \in [M]$, the optimistic loss estimator is defined as:

$$\ell_{t,j} := \frac{\mathbb{I}(j_t = j)}{w_{t,j} + \gamma} \left(L - \sum_{k \in [0,\dots,L-1]} r_t(x_k^t, a_k^t) + \Lambda \sum_{i \in [m]} \left[\left(\widehat{G}_t^j \right)^\top \widehat{q}_t^j - \theta \right]_i^+ \right), \tag{9.7}$$

where γ is a suitably-defined implicit exploration factor, (x_k^t, a_k^t) is the state-action pair visited at layer k during episode t, Λ is a suitably-defined upper bound on the optimal values of Lagrangian multipliers,[1] $\widehat{G}_t^j$ is the matrix of empirical constraint costs built by the instance $\texttt{Alg}^j$ of Algorithm 9.1 at episode t, while $\widehat{q}_t^j$ is the occupancy measure computed by instance $\texttt{Alg}^j$ of Algorithm 9.1 at t. Finally, the algorithm updates the weight vector according to an FTRL update on a cut decision space with a suitable log-barrier regularization and a bonus term b_t defined as:

$$b_{t,j} := \left((m\Lambda\beta_5 + \beta_2) + \left(\sqrt{\beta_1} + m\Lambda\sqrt{\beta_4} \right) \sqrt{T} \right) \cdot (\nu_{t,j} - \nu_{t-1,j}), \tag{9.8}$$

where $\nu_{t,j} = \max_{\tau \le t} \frac{1}{w_{\tau,j}}$ and the parameters β are linked to the performance of Algorithm 9.1 (see Line 14 and Section F.5.1 for additional details). See Line 16 for the complete definition of the update. The bonus term purpose is to balance out the term related to the difference between the performance of Algorithm 9.1 updated at each episode and the performance of its stabilized version, which works under the condition imposed by the master algorithm.

9.4.2 Theoretical Guarantees of `Lag-FTRL`

Next, we prove the theoretical guarantees attained by Algorithm 9.2 (see Appendix F.5 for complete proofs of the theorems and associated lemmas). We start by extending standard

[1] Notice that, in the definition of Λ, ρ is the feasibility parameter of Program (2.3) for the reward vector $\overline{r}$, the constraint cost matrix $\overline{G}$, and the threshold vector θ. In order to compute Λ, Algorithm 9.2 needs knowledge of ρ. Nevertheless, our results continue to hold even if Algorithm 9.2 is only given access to a lower bound on ρ.

strong-duality results for CMDPs (see Lemma A.2) to the case of a Lagrangian function suitably-modified to encompass *strong* violations. We call it *positive Lagrangian* of Program (2.3), defined as follows.

Definition 9.1 (Positive Lagrangian)**.** *Given a CMDP with a transitions P, for every reward vector $r \in [0,1]^{|X\times A|}$, constraint cost matrix $G \in [0,1]^{|X\times A|\times m}$, and threshold vector $\theta \in [0,L]^m$, the* positive Lagrangian *of Program (2.3) is defined as a function $\mathcal{Q} : \mathbb{R}_{\geq 0} \times \Delta(M) \to \mathbb{R}$ such that $\mathcal{Q}(\beta, q) \coloneqq r^\top q - \beta \sum_{i\in[m]} \left[G^\top q - \theta\right]_i^+$ for every $\beta \geq 0$ and $q \in \Delta(M)$.*

The positive Lagrangian is related to the Lagrangian of a variation of Program (2.3) in which the $[\cdot]^+$ operator is applied to the constraints. Notice that such a problem does *not* meet Slater's condition, since, by definition of $[\cdot]^+$, it does *not* exist an occupancy measure $q^\diamond$ such that $\left[G^\top q^\diamond - \theta\right]_i^+ < 0$ for every $i \in [m]$. Nevertheless, we show that some sort of strong duality result still holds for $\mathcal{Q}(L/\rho, q)$, when Slater's condition is met by Program (2.3). This is made formal by the following theorem.

Theorem 9.3. *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 $\theta \in [0,L]^m$, if Program (2.3) satisfies Slater's condition (Condition 2.1), then the following holds:*

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

where ρ is the feasibility parameter of Program (2.3).

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

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

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 A.2. Noticing that for all q belonging to $\left\{q \in \Delta(P) : G^\top q \leq \theta\right\}$, we have $\mathcal{Q}(1/\rho, q) = r^\top q$, which implies that $\max_{q\in\Delta(M)} \mathcal{Q}(1/\rho, q) \geq \mathrm{OPT}_{r,G,\theta}$, concludes the proof. □

Theorem 9.3 intuitively shows that a L/ρ multiplicative factor on the positive constraint violation is enough to compensate the large rewards that non-feasible policies would attain when employed by the learner. This result is crucial since, without properly defining the Lagrangian function optimized by Algorithm 9.2, the FTRL optimization proce-

dure would choose instances with both large rewards and large constraint violation, thus preventing the violation bound from being sublinear.

By means of Theorem 9.3, it is possible to provide the following result.

Theorem 9.4. *If Program (2.3) instantiated with $\overline{r}$, $\overline{G}$ and θ satisfies Slater's condition (Condition 2.1), then, given any $\delta \in (0,1)$, with probability at least $1-34\delta$, Algorithm 9.2 attains strong constraint violation:*

$$\begin{aligned}\mathcal{V}_T = \mathcal{O}\Big(&m^2L^2|X|\sqrt{|A|T\ln\left(mT|X||A|/\delta\right)}\ln(T)^2\\&+m^2L|X|^2|A|^2\ln(T)^3\ln\left(\ln(T)/\delta\right)\\&+m^2L\ln(T)^2|X||A|C\Big).\end{aligned}$$

Intuitively, to prove Theorem 9.4, it is necessary to bound the negative regret attained by the algorithm, *i.e.*, how better Algorithm 9.2 can perform in terms of rewards with respect to an optimal occupancy in hindsight q^*. Notice that this is equivalent to showing that the FTRL procedure cannot gain more than $\text{OPT}_{\overline{r},\overline{G},\theta}$ by playing policies that are *not* feasible, or, equivalently, by choosing instances $\texttt{Alg}^j$ with a large corruption guess, which, by definition of the confidence sets employed by Algorithm 9.1, may play non-feasible policies attaining large rewards. This is done by employing Theorem 9.3, which shows that the positive Lagrangian does *not* allow the algorithm to achieve too large rewards with respect to q^*. Thus, the violations are still upper bounded by $\widetilde{\mathcal{O}}(\sqrt{T}+C)$.

Proof. In order to obtain the final violation bound, it is necessary to find an upper bound to the negative regret $-R_T$. We proceed as follows,

$$\begin{aligned}\overline{r}^\top q^* &= \text{OPT}_{\overline{r},\overline{G},\theta} && \text{(9.9a)}\\&= \max_{q\in\Delta(M)}\left(\overline{r}^\top q - \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q-\theta_i\right]^+\right) && \text{(9.9b)}\\&\geq \overline{r}^\top q_t - \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t-\theta_i\right]^+,\end{aligned}$$

where Equality (9.9a) holds since q^* is the feasible occupancy that maximizes the reward vector $\overline{r}$ and Equality (9.9b) holds by Theorem 9.3 . This implies $\overline{r}^\top q_t - \overline{r}^\top q^* \leq \frac{L}{\rho}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t-\theta_i\right]^+$. Moreover, it holds:

$$\begin{aligned}&\sum_{t\in[T]}\sum_{i\in[m]}\left[\overline{G}_i^\top q_t-\theta_i\right]^+\\&\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t-\theta_i\right]^+ + \sum_{i\in[m]}\left[(\overline{G}_i-\mathbb{E}[g_{t,i}])^\top q_t\right]^+\right) && \text{(9.10a)}\\&\leq \sum_{t\in[T]}\left(\sum_{i\in[m]}\left[\mathbb{E}[g_{t,i}]^\top q_t-\theta_i\right]^+ + \sum_{i\in[m]}\left\|\overline{G}_i-\mathbb{E}[g_{t,i}]\right\|_1\right) && \text{(9.10b)}\end{aligned}$$

$$\leq \sum_{t\in[T]} \left(\sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t - \theta_i\right]^+ + \sum_{i\in[m]} \left(\left\|\overline{G}_i - g_i^\circ\right\|_1 + \|g_i^\circ - \mathbb{E}[g_{t,i}]\|_1 \right) \right)$$

$$\leq m\mathcal{V}_T + 2mC, \tag{9.10c}$$

where Inequality (9.10a) holds since $[a+b]^+ \leq [a]^+ + [b]^+, a \in \mathbb{R}, b \in \mathbb{R}$, Inequality (9.10b) holds since $q_t(x,a) \leq 1, \forall t \in [T], \forall (x,a) \in X \times A$, and finally Inequality (9.10c) holds by definition of C and $\mathcal{V}_T$ 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:

$$\begin{aligned}
R_T + \frac{Lm+1}{\rho}\mathcal{V}_T &= \sum_{t\in[T]} \mathbb{E}[r_t]^\top (q^* - q_t) + \frac{Lm+1}{\rho}\mathcal{V}_T \\
&= \sum_{t\in[T]} (\mathbb{E}[r_t] - \overline{r})^\top (q^* - q_t) + \sum_{t\in[T]} \overline{r}^\top (q^* - q_t) + \frac{Lm+1}{\rho}\mathcal{V}_T \\
&\geq -\sum_{t\in[T]} \|\mathbb{E}[r_t] - \overline{r}\|_1 + \sum_{t\in[T]} \overline{r}^\top (q^* - q_t) + \frac{Lm+1}{\rho}\mathcal{V}_T \qquad (9.11a) \\
&\geq -2C - \frac{L}{\rho}(m\mathcal{V}_T + 2mC) + \frac{Lm+1}{\rho}\mathcal{V}_T \qquad (9.11b) \\
&= -2C - \frac{2LmC}{\rho} + \mathcal{V}_T \left(\frac{Lm+1}{\rho} - \frac{Lm}{\rho} \right) \\
&= \frac{1}{\rho}\mathcal{V}_T - \left(2C + \frac{2LmC}{\rho} \right), \qquad (9.11c)
\end{aligned}$$

where Inequality (9.11a) 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 (9.11b) holds since $\overline{r}^\top (q^* - q_t) \geq -\frac{L}{\rho} \sum_{i\in[m]} \left[\overline{G}_i^\top q_t - \theta_i\right]^+ \geq -(m\mathcal{V}_T + 2mC)$ and by definition of C. Thus, rearranging Inequality (9.11c), we finally bound the cumulative violation as follows:

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

where the last inequality holds by Lemma F.13, with probability at least $1 - 4\delta$ under $\mathcal{E}_{\widehat{q}}$. Employing Equation (9.14) and a Union Bound, setting $\gamma = \sqrt{\frac{\ln(M/\delta)}{TM}}$ and $\eta \leq \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T} + \beta_2 + \beta_5 + \sqrt{\beta_4 T}\right)}$ concludes the proof. □

Finally, we prove the regret bound attained by Algorithm 9.2.

Theorem 9.5. *If Program (2.3) instantiated with $\overline{r}$, $\overline{G}$ and θ satisfies Slater's condition (Condition 2.1), then, given any $\delta \in (0,1)$, with probability at least $1-30\delta$, Algorithm 9.2 attains regret:*

$$\begin{aligned} R_T = \mathcal{O}\Big(& m^2 L^2 |X| \sqrt{|A| T \ln\left(mT|X||A|/\delta\right)} \ln(T)^2 \\ & + m^2 L |X|^2 |A|^2 \ln(T)^3 \ln\left(\ln(T)/\delta\right) \\ & + m^2 L \ln(T)^2 |X||A|C \Big). \end{aligned}$$

Bounding the regret attained by Algorithm 9.2 requires different techniques with respect to bounding constraint violation. Indeed, strong duality is *not* needed, since, even if Λ is set to a too small value and thus the algorithm plays non-feasible policies, then the regret would still be sublinear. The regret bound is strongly related to the optimal value of the problem associated with the positive Lagrangian, which, by definition of $[\cdot]^+$ cannot perform worse than the optimum of Program (2.3), in terms of rewards gained. Thus, by letting j^* be the index of the instance associated with true corruption value C, proving Theorem 9.5 reduces to bounding the regret and the constraint violation of instance $\texttt{Alg}^{j^*}$, with the additional challenge of bounding the estimation error of the optimistic loss estimator. Finally, by means of the results for the *known* C case derived in Section 9.3, we are able to show that the regret is at most $\widetilde{\mathcal{O}}(\sqrt{T}+C)$, which is the desired bound.

Proof. Employing Algorithm 9.2, with probability at least $1-14\delta$, it holds:

$$\begin{aligned} R_T &= \sum_{t\in[T]} \overline{r}^\top q^* - \sum_{t\in[T]} \overline{r}^\top q_t \\ &= \sum_{t\in[T]} \overline{r}^\top (q^* - q_t^{j^*}) + \sum_{t\in[T]} \overline{r}^\top (q_t^{j^*} - q_t) \\ &= \sqrt{\beta_1 T}\nu_{T,j^*} + \beta_2 \nu_{T,j^*} + 2\beta_3 C + \sum_{t\in[T]} \overline{r}^\top (q_t^{j^*} - q_t) \qquad (9.12a) \\ &\le \sqrt{\beta_1 T}\nu_{T,j^*} + \beta_2 \nu_{T,j^*} + 2\beta_3 C + 2C - \frac{Lm+1}{\rho}\widehat{V}_T + \frac{Lm+1}{\rho}\widehat{V}_{T,j^*} \\ &\quad - \left(\sqrt{\beta_1} + \frac{m(Lm+1)}{\rho}\sqrt{\beta_4}\right)\sqrt{T}\nu_{T,j*} - \left(\beta_2 + \frac{m(mL+1)}{\rho}\beta_5\right)\nu_{T,j^*} \\ &\quad + \mathcal{O}\Bigg(\frac{M\ln T}{\eta} + \eta\, m^4 L^4 T M + \eta\, M \ln(T) m^4 L^2 \left(\beta_2^2 + \beta_5^2\right) \\ &\quad + \eta T(\beta_1 + L^2 m^4 \beta_4) M \ln(T) + \gamma T L M \\ &\quad + L\sqrt{T\ln\left(1/\delta\right)} + \frac{Lm}{\gamma}\ln\left(1/\delta\right) \Bigg). \qquad (9.12b) \end{aligned}$$

where Inequality (9.12a) hold with probability at least $1-11\delta$ by Corollary F.7, Inequality (9.12b) holds with probability at least $1-3\delta$ thanks to Lemma F.12 and to the following reasoning, which holds with probability at least $1-\delta$:

$$\sum_{t\in[T]} \overline{r}^\top (q_t^{j*} - q_t)$$

$$
\begin{aligned}
&= \sum_{t\in[T]} (\overline{r} - \mathbb{E}[r_t])^\top (q_t^{j*} - q_t) + \sum_{t\in[T]} \mathbb{E}[r_t]^\top (q_t^{j*} - q_t) \\
&\leq \sum_{t\in[T]} \|\overline{r} - \mathbb{E}[r_t]\|_1 + \sum_{t\in[T]} \mathbb{E}[r_t]^\top \left(q_t^{j*} - q_t\right) && \text{(9.13a)} \\
&\leq 2C + \sum_{t\in[T]} \mathbb{E}[r_t]^\top \left(q_t^{j*} - q_t\right) && \text{(9.13b)} \\
&\leq 2C + \sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{j*} - \sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} \mathbb{E}[r_t]^\top q_t^j + L\sqrt{2T\ln(1/\delta)}, && \text{(9.13c)}
\end{aligned}
$$

where Inequality (9.13a) holds since $|q_t(x,a) - q_t^{j^*}(x,a)| \leq 1,\ \forall (x,a) \in X \times A$, where Inequality (9.13b) holds by definition of C, and where Inequality (9.13c) use Azuma-Hoeffding inequality.
We can apply Lemma F.14 to bound $\widehat{V}_{T,j^*}$ with high probability. In fact we observe that with probability at least $1 - 16\delta$, it holds:

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

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

$$
\begin{aligned}
& R_T + \frac{Lm+1}{\rho}\widehat{V}_T \\
& \leq \mathcal{O}\Bigg(\frac{M\ln T}{\eta} + \eta\, m^4L^4TM + \eta\, M\ln(T)m^4L^2(\beta_2^2+\beta_5^2) \\
& \quad + \eta T(\beta_1 + L^2m^4\beta_4)M\ln(T) \\
& \quad + \gamma TLM + L\sqrt{T\ln(1/\delta)} + \frac{Lm}{\gamma}\ln(1/\delta) \\
& \quad + m^2L^2|X|\sqrt{|A|T\ln\left(\frac{mMT|X||A|}{\delta}\right)} + mL\beta_6 C \\
& \quad + \beta_3 C + m^2L|X||A|\ln(T)C\Bigg) && \text{(9.14)}
\end{aligned}
$$

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

CHAPTER 10

Conclusions and Discussion

The aim of this dissertation has been to significantly advance the theoretical understanding of *online learning* in *constrained Markov decision processes* (Altman, 1999). We studied CMDPs where rewards and constraints can be stochastic, adversarial, or non-stationary, and we provided algorithms with provable guarantees in terms of *regret* and *violation*. Our contributions cover a wide range of settings, under both *full* and *bandit feedback*.

In the first part of the dissertation, we considered CMDPs with *stochastic rewards* and *stochastic constraints*. In such a case, we showed that it is possible to design an efficient primal–dual *policy optimization* algorithm that achieves the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds on both *strong regret* and *strong violation*. This resolves an open question left by (Efroni et al., 2020; Müller et al., 2024), and shows that efficient algorithms with optimal guarantees exist without relying on occupancy measures.

In the second part, we addressed CMDPs with *adversarial rewards* and *stochastic constraints*, focusing on *hard* constraints (Pacchiano et al., 2021). This work introduced the first algorithms capable of operating in this setting. Our contributions are threefold: (i) we designed an algorithm that guarantees sublinear regret together with sublinear cumulative strong violation; (ii) we developed a *safe* algorithm that satisfies the constraints at every episode with high probability; and (iii) we proposed an algorithm that achieves sublinear regret while limiting the cumulative strong violation to a constant, provided that Slater's condition holds. In addition, we established a lower bound demonstrating that any algorithm with $o(\sqrt{T})$ strong violation must incur a regret term that depends on the Slater parameter, thereby formalizing the intrinsic trade-off between feasibility and performance in this class of problems.

In the third part, we introduced algorithms for CMDPs that can handle both stochastic and adversarial constraints, providing *best-of-both-worlds* guarantees (Castiglioni et al.,

2022b). Under full feedback, we designed a primal–dual method that achieves optimal regret and violation when the environment is stochastic, and sublinear violation with no-α-regret guarantees when the constraints are adversarial. We then extended these results to the bandit setting through efficient policy optimization methods. Finally, we proposed an improved algorithm that strengthens the guarantees, removing the need for Slater's condition in the stochastic case and ensuring strong violation bounds.

In the fourth part, we studied CMDPs with *non-stationary rewards and constraints*. We proposed algorithms whose performance adapts to the level of non-stationarity, expressed through a corruption parameter C. Our methods achieve $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and strong violation, matching impossibility results in the worst case and recovering optimal rates in the stochastic case.

In summary, this dissertation establishes a comprehensive theoretical framework for online CMDPs across stochastic, adversarial, and non-stationary regimes. We provided the first efficient policy optimization algorithms with optimal strong guarantees, introduced new methods for hard constraints under adversarial rewards, developed best-of-both-worlds algorithms for stochastic and adversarial constraints, and extended the analysis to non-stationary settings. Together, these results significantly broaden our understanding of constrained reinforcement learning.

10.1 Future Directions

The results presented in this dissertation open several avenues for further research.

A natural extension concerns CMDPs defined over large or *continuous state and action spaces*. In such scenarios, function approximation techniques and structural assumptions, such as linear MDP models (Jin et al., 2020b), become essential to design efficient algorithms with theoretical guarantees. Bridging our results with this line of research could provide scalable methods that preserve strong feasibility properties even in high-dimensional settings.

Another promising direction is the study of CMDPs with *infinite horizon* and discounted objectives (see (Sutton and Barto, 1998) for the unconstrained definition). Our results are derived in the finite-horizon episodic framework; extending the analysis to the discounted setting would require new techniques to handle the long-term accumulation of rewards and constraints while ensuring sublinear regret and violation. This step is fundamental to bring constrained reinforcement learning theory closer to classical MDP formulations.

Finally, constrained settings under partial observability remain largely unexplored. In many practical applications, the learner has access only to partial or noisy observations of the underlying state. Extending our framework to *partially observable* CMDPs would represent an important step towards making constrained learning algorithms applicable in more realistic environments where uncertainty and limited feedback are involved.

APPENDIX $\mathcal{A}$

Omitted Lemmas and Proofs of Chapter 3

A.1 Confidence Intervals

In this section, we report the omitted proof related to the confidence interval employed by our algorithm to deal with the uncertainty on the environments.

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

$$\left|\widehat{r}_t(x,a) - \overline{r}(x,a)\right| \leq \iota_t(x,a).$$

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

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

where $\iota_t(x,a) := \sqrt{\frac{\ln(2/\delta)}{2N_t(x,a)}}$.

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

$$\mathbb{P}\left[\left|\widehat{r}_t(x,a) - \overline{r}(x,a)\right| \geq \frac{c}{N_t(x,a)}\right] \leq 2\exp\left(-\frac{2c^2}{N_t(x,a)}\right)$$

Setting $\delta = 2\exp\left(-\frac{2c^2}{N_t(x,a)}\right)$ and solving to find a proper value of c gives the result for the reward function.
Focusing on specifics $i \in [m]$, $t \in [T]$ and $(x,a) \in X \times A$ and following the previous reasoning applied to the constraints functions concludes the proof. □

A.2 Strong Duality

We start by proving that, when Slater's condition holds (Condition 2.1), in an optimal solution the vector of Lagrange multipliers is bounded.

Lemma A.2. *Given a CMDP with reward vector $r \in [0,1]^{|X\times A|}$ and cost matrix $G \in [0,1]^{|X\times A|\times m}$, it holds:*

$$\min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) = \max_{\pi\in\Pi} \min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \mathcal{L}_{r,G,\theta}(\pi,\lambda) = \mathrm{OPT}_{r,G,\theta}.$$

Proof. We start by proving the following result:

$$\min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) = \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda).$$

To do so, let us first notice that, for every $\lambda \in \mathbb{R}^m_{\geq 0}$ such that $\|\lambda\|_1 > L/\rho$:

$$\max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \geq \mathcal{L}_{r,G,\theta}(\pi^\diamond,\lambda) \geq -\sum_{i\in[m]} \lambda_i \left(V^{\pi^\diamond}(g_i) - \theta_i\right) \geq \|\lambda\|_1 \rho > L,$$

where we recall that $\pi^\diamond := \arg\max_{\pi\in\Pi} \min_{i\in[m]} (\theta_i - V^\pi(g_i))$. Moreover:

$$\min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \leq \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\underline{0}) = \max_{\pi\in\Pi} V^\pi(r) \leq L.$$

This implies that:

$$\begin{aligned}
&\min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \\
&= \min\left\{ \min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda), \min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1>L/\rho} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \right\} \\
&= \min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda).
\end{aligned}$$

In conclusion,

$$\begin{aligned}
\mathrm{OPT}_{r,G,\theta} &= \max_{\pi\in\Pi} \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \\
&\leq \max_{\pi\in\Pi} \min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \\
&\leq \min_{\lambda\in\mathbb{R}^m_{\geq 0}:\|\lambda\|_1\in[0,L/\rho]} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \\
&= \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \max_{\pi\in\Pi} \mathcal{L}_{r,G,\theta}(\pi,\lambda) \\
&= \mathrm{OPT}_{r,G,\theta},
\end{aligned}$$

where the second inequality above holds by the *max-min* inequality and the last equality by strong duality in CMDPs (refer to (Altman, 1999)). This concludes the proof. □

Now we extend the result showing that, depending on whether an occupancy is safe or not, there exist values of λ such that the optimal value of the Lagrangian is upper bounded by the optimum of Program (2.3).

Lemma A.3. *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 $\theta \in [0,L]^m$ and $\pi \in \Pi$ s.t. $V^\pi(g_i) \le \theta_i \ \forall i \in [m]$, it holds:*

$$\mathcal{L}_{r,G,\theta}(\pi, \underline{0}) \le OPT_{r,G,\theta}.$$

Proof. The proof directly follows from the definition of $\text{OPT}_{r,G,\theta}$ induced by Program (2.3). □

We conclude the section with the following result.

Lemma A.4. *With probability at least $1-\delta$, Algorithm 3.1 guarantees, for all $t \in [T]$:*

$$\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) \ge \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) - \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) - \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1.$$

Proof. Similarly to Lemma 3.4, we notice that $\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*, \lambda_t) \ge \text{OPT}_{\overline{r},\overline{G},\theta}$. Thus, following the same steps as in Lemma 3.4, we proceed as follows:

$$\begin{aligned}
&\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) \\
&\le V^{\pi_t}(\overline{r}) - \max_{\lambda \in \left\{0, \frac{L+1}{\rho}\right\}^m} \sum_{i\in[m]} \frac{\lambda_i L}{L+1}\left(V^{\pi_t}(\overline{g}_i) - \theta_i\right) + \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\qquad + \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \\
&= V^{\pi_t}(\overline{r}) - \max_{\lambda \in \left[0, \frac{L+1}{\rho}\right]^m} \sum_{i\in[m]} \frac{\lambda_i L}{L+1}\left(V^{\pi_t}(\overline{g}_i) - \theta_i\right) + \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\qquad + \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \\
&= V^{\pi_t}(\overline{r}) - \max_{\lambda \in \left[0, \frac{L}{\rho}\right]^m} \sum_{\in[m]} \lambda_i\left(V^{\pi_t}(\overline{g_i}) - \theta_i\right) + \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\qquad + \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \\
&\le \max_{\pi\in\Pi}\left[V^{\pi}(\overline{r}) - \max_{\lambda \in \left[0, \frac{L}{\rho}\right]^m} \sum_{i\in[m]} \lambda_i\left(V^{\pi}(\overline{g_i}) - \theta_i\right)\right] + \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\qquad + \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \\
&\le \max_{\pi\in\Pi}\left[V^{\pi}(\overline{r}) - \max_{\|\lambda\|_1 \in \left[0, \frac{L}{\rho}\right]} \sum_{i\in[m]} \lambda_i\left(V^{\pi}(\overline{g}_i) - \theta_i\right)\right] + \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) \\
&\qquad + \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1 \\
&\le \text{OPT}_{\overline{r},\overline{G},\theta} + \frac{2Lm}{\rho} V^{\pi_t}(\xi_t) + \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1, \qquad \text{(A.1)}
\end{aligned}$$

where Inequality (A.1) holds by Lemma A.2.

Thus, it holds:

$$\begin{aligned}\mathcal{L}_{\overline{r},\overline{G},\theta}(\pi^*,\lambda_t) &\geq \mathrm{OPT}_{\overline{r},\overline{G},\theta} \\ &\geq \mathcal{L}_{\overline{r},\overline{G},\theta}(\pi_t, {}^{\lambda_t L}/_{L+1}) - \frac{2Lm}{\rho}V^{\pi_t}(\xi_t) - \frac{4Lm}{\rho}\|q^{\widehat{P},\pi_t} - q^{P,\pi_t}\|_1,\end{aligned}$$

which concludes the proof. □

A.3 Regret

We show the concentration rate of the confidence intervals.

Lemma A.5. *With probability at least* $1-\delta$*, it holds:*

$$\sum_{t=1}^{T} V^{\pi_t}(\phi_t) \leq 4\sqrt{L|X||A|T\ln\left(\frac{T|X||A|}{\delta}\right)} + L\sqrt{2T\ln\frac{1}{\delta}}$$

Proof. We first notice the following bound,

$$V^{\pi_t}(\phi_t) \leq L.$$

Thus, we can employ the Azuma inequality to bound the following Martingale difference sequence as

$$\sum_{t=1}^{T} V^{\pi_t}(\phi_t) - \sum_{t=1}^{T}\sum_{x,a}\phi_t(x,a)\mathbb{I}_t(x,a) \leq L\sqrt{2T\ln\frac{1}{\delta}},$$

which holds with probability $1-\delta$. Thus we can bound the quantity of interest as follows,

$$\begin{aligned}\sum_{t=1}^{T} V^{\pi_t}(\phi_t) &\leq \sum_{t=1}^{T}\sum_{x,a}\phi_t(x,a)\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} \\ &= \sqrt{4\ln\left(\frac{T|X||A|}{\delta}\right)}\sum_{t=1}^{T}\sum_{x,a}\sqrt{\frac{1}{\max\{1,N_t(x,a)\}}}\mathbb{I}_t(x,a) + L\sqrt{2T\ln\frac{1}{\delta}} \\ &\leq 2\sqrt{4\ln\left(\frac{T|X||A|}{\delta}\right)}\sum_{x,a}\sqrt{N_T(x,a)} + L\sqrt{2T\ln\frac{1}{\delta}} \qquad \text{(A.2)} \\ &\leq 4\sqrt{L|X||A|T\ln\left(\frac{T|X||A|}{\delta}\right)} + L\sqrt{2T\ln\frac{1}{\delta}}, \qquad \text{(A.3)}\end{aligned}$$

where Inequality (A.2) holds since $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ and Inequality (A.3) follows from Cauchy-Schwarz inequality and noticing that $\sqrt{\sum_{x,a}N_T(x,a)} \leq \sqrt{LT}$. This concludes the proof. □

Lemma A.6. *With probability at least* $1-\delta$*, it holds:*

$$\sum_{t=1}^{T} V^{\pi_t}(\xi_t) \leq 4\sqrt{L|X||A|T\ln\left(\frac{T|X||A|m}{\delta}\right)} + L\sqrt{2T\ln\frac{1}{\delta}}$$

Proof. The proof follows the one of Lemma A.5, replacing ϕ_t with ξ_t. □

A.4 Policy Optimization with Dilated Bonuses

In this section, we present the regret bound attained by `PO-DB`.

Lemma A.7 (Luo et al. (2021))**.** *For any sequence of losses* ℓ_t *such that* $\ell_t \in [0,1]^{|X\times A|}$ *and any valid occupancy measure* $\pi \in \Pi$*,* `PO-DB` *attains:*

$$\sum_{t=1}^{T} V^{\pi_t}(\ell_t) - \sum_{t=1}^{T} V^{\pi}(\ell_t) \leq \widetilde{\mathcal{O}}\left(L^2|X|\sqrt{|A|T} + L^4\right),$$

with probability at least $1-\mathcal{O}(\delta)$*.*

A.5 Transition Estimations

In the following section, we show how the estimated occupancy measure concentrates to the true one.

To do so, we first provide some discussion on the transitions confidence set.

A.5.1 Confidence Set

We introduce *confidence sets* for the transition function of a CMDP, by exploiting suitable concentration bounds for estimated transition probabilities. By letting $M_t(x,a,x')$ be the total number of episodes up to $t \in [T]$ in which the state-action pair $(x,a) \in X \times A$ is visited and the environment evolves to the new state $x' \in X$, we define the estimated transition probability at t for the triplet (x,a,x') as $\widehat{P}_t(x' \mid x,a) = \frac{M_t(x,a,x')}{\max\{1,N_t(x,a)\}}$. Then, the confidence set at $t \in [T]$ is $\mathcal{P}_t := \bigcap_{(x,a,x')\in X\times A\times X} \mathcal{P}_t^{x,a,x'}$, where:

$$\mathcal{P}_t^{x,a,x'} := \left\{\overline{P} : \left|\overline{P}(x'|x,a) - \widehat{P}_t(x'|x,a)\right| \leq \epsilon_t(x,a,x')\right\},$$

with $\epsilon_t(x,a,x')$ equal to:

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

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

The next lemma establishes $\mathcal{P}_t$ is a proper confidence set.

Lemma A.8 (Jin et al. (2020a))**.** *Given a confidence parameter* $\delta \in (0,1)$*, with probability at least* $1-4\delta$*, it holds that the transition function* P *belongs to* $\mathcal{P}_t$ *for all* $t \in [T]$*.*

A.5.2 Concentration Results

Given the confidence set of the transition, it is possible to derive the following lemma.

Lemma A.9 (Lemma 4, Jin et al. (2020a)). *With probability at least* $1 - 6\delta$*, for any collection of transition functions* $\{P_t^x\}_{x \in X}$ *such that* $P_t^x \in \mathcal{P}_t$*, we have, for all* x*,*

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

As final remark, we underline that the empirical transition function $\widehat{P}_t$ belongs to $\mathcal{P}_t$ by construction. Thus, the aforementioned lemma immidiately holds for $\widehat{P}_t$.

APPENDIX B

Omitted Lemmas and Proofs of Chapter 4

B.1 Transitions Concentration for Algorithm 4.2

We state the following lemma, which is a generalization of the results from Jin et al. (2020a). Intuitively, the following result states that the distance between the estimated *non-safe* occupancy measure $\widehat{q}_t$ and the real one reduces as the number of episodes increases, paying a $1-\lambda_t$ factor. This is reasonable since, from the update of the *non-Markovian* policy π_t (see Algorithm 4.2), policy $\widehat{\pi}_t \leftarrow \widehat{q}_t$ is played with probability $1-\lambda_{t-1}$.

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

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

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

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

Now following standard analysis by Lemma A.9 from Jin et al. (2020a), we have that,

$$\sum_{t=1}^{T}(1-\lambda_{t-1})\sum_{x\in X,a\in A}\left|q^{P_t^x,\widehat{\pi}_t}(x,a)-q^{P,\widehat{\pi}_t}(x,a)\right| \leq$$

$$\sum_{0\le m<k<L}\sum_{t,w_m}(1-\lambda_{t-1})\epsilon_t^*(x_{m+1}|x_m,a_m)q^{P,\widehat{\pi}_t}(x_m,a_m)$$
$$+|X|\sum_{0\le m<h<L}\sum_{t,w_m,w'_h}(1-\lambda_{t-1})\cdot$$
$$\cdot\,\epsilon_{i_t}^*\left(x_{m+1}\mid x_m,a_m\right)q^{P,\widehat{\pi}_t}\left(x_m,a_m\right)\epsilon_t^*\left(x'_{h+1}\mid x'_h,a'_h\right)q^{P,\widehat{\pi}_t}\left(x'_h,a'_h\mid x_{m+1}\right),$$

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

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

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

where the last step holds following Lemma A.9 from Jin et al. (2020a).

Bound on the second term. Following Lemma A.9 from Jin et al. (2020a), the second term is bounded by (ignoring constants),

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

The last two terms are bounded logarithmically in T, employing the definition of $q^{P,\widehat{\pi}_t}$ and following Lemma A.9 from Jin et al. (2020a), while, similarly, the first term is bounded by:

$$\sum_{0\le m<h<L}\sqrt{|X_{m+1}|\sum_{t,x_m,a_m}\frac{(1-\lambda_{t-1})q^{P,\widehat{\pi}_t}\left(x_m,a_m\right)}{\max\left\{1,N_t\left(x_m,a_m\right)\right\}}}$$

$$\cdot \sqrt{|X_{h+1}| \sum_{t,x'_h,a'_h} \frac{(1-\lambda_{t-1}) q^{P,\widehat{\pi}_t}(x'_h,a'_h)}{\max\{1, N_t(x'_h,a'_h)\}}},$$

which is upper bounded by:

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

Employing the same argument as Lemma A.9 from Jin et al. (2020a) shows that the previous term is bounded logarithmically in T and concludes the proof. □

B.2 Auxiliary Lemmas from Existing Works

In this section, we provide auxiliary lemmas and results from existing works.

B.2.1 Transitions Estimation

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

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

Given the estimation of the occupancy measure space, it is possible to apply Lemma A.9.

We underline that the constrained space defined by Program (4.1) is a subset of $\Delta(\mathcal{P}_t)$. This implies that, in Algorithm 4.1, it holds $\widehat{q}_t \in \Delta(\mathcal{P}_t)$ and Lemma A.9 is valid.

B.2.2 Auxiliary Lemmas for the Optimistic Loss Estimator

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

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

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

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

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

Following the analysis of Lemma B.2, with $\alpha_t(x,a) = 2\gamma \mathbb{I}_t(x,a)$ and a Union Bound, the following corollary holds.

Corollary B.1 (Jin et al. (2020a)). *With probability at least $1 - \delta$:*

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

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

Lemma B.3 (Jin et al. (2020a)). *With probability at least $1 - 7\delta$,*

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

We notice that $\pi_t \leftarrow \widehat{q}_t$ holds only for Algorithm 4.1, since in Algorithm 4.2, $\pi_t \leftarrow \widehat{q}_t$ with probability $1 - \lambda_{t-1}$.

B.2.3 Auxiliary Lemmas for Online Mirror Descent

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

Lemma B.4 (Jin et al. (2020a)). *The OMD update with $\widehat{q}_1(x,a,x') = \frac{1}{|X_k||A||X_{k+1}|}$ for all $k < L$ and $(x,a,x') \in X_k \times A \times X_{k+1}$, and*

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

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

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

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

APPENDIX C

Omitted Lemmas and Proofs of Chapter 5

C.1 Events

In this section, we provide the definition of the events that we use in the chapter.

The following event states that the true occupancy measure space is always contained in the confidence set:

Event $E^{\Delta}(\delta)$**:** $\Delta(M) \subseteq \cap_j \Delta(\mathcal{P}_j)$.

In particular, under $E^{\Delta}(\delta)$, we have that $q^{\diamond}, q^{*} \in \cap_j \Delta(\mathcal{P}_j)$. $E^{\Delta}(\delta)$ holds with probability at least $1-\delta$ (see Lemma A.8).

The following event states that the cumulative error after T episodes due to the difference between q^{P,π_t} and $q^{P^{\widehat{q}_t},\pi_t}$ is small enough:

Event $E^{\widehat{q}}(\delta)$**:**

$$\sum_{t=1}^{T} ||q_t - \widehat{q}_t||_1 \leq \mathcal{E}^q_\delta,$$

where $\mathcal{E}^q_\delta := 4L|X|\sqrt{2T\ln\left(\frac{1}{\delta}\right)} + 6L|X|\sqrt{2T|A|\ln\left(\frac{T|X||A|}{\delta}\right)} \leq \widetilde{\mathcal{O}}(\sqrt{T})$.

In the next sections we will often condition on the intersection of the previous events:

Event $E^{\Delta,\widehat{q}}(\delta)$**:** $E^{\widehat{q}}(\delta) \cap E^{\Delta}(\delta)$.

$E^{\Delta,\widehat{q}}(\delta)$ holds with probability at least $1-2\delta$ (see Lemma 5.1).

The next event states that, in case the rewards are stochastic, the reward accumulated is not too far from the mean reward accumulated.

Event $E^{r}_{q^*}(\delta)$**:** $\left|\sum_{t=1}^{T}\left(r_t - \overline{r}\right)^{\top} q^*\right| \leq \mathcal{E}^r_\delta$, *where* $\mathcal{E}^r_\delta = \frac{L}{\sqrt{2}}\sqrt{T\ln\left(\frac{2}{\delta}\right)} \leq \widetilde{O}(\sqrt{T})$.

$E^r_{q^*}(\delta)$ holds with probability at least $1-\delta$ (see Lemma 5.3).
For the stochastic constraint setting, we define the quantity:

$$\mathcal{E}^G_{t_1,t_2,\delta} := 2L\sqrt{2(t_2-t_1+1)\ln\left(\frac{T^2}{\delta}\right)}$$

and then two events bounding the cumulative difference between the dual utility with the average constraints and that with the sampled constraints.

Event $E^G_{q^\diamond}(\delta)$**:** *for all* $[t_1,\dots,t_2]\subseteq[1,\dots,T]$, $\left|\sum_{t=t_1}^{t_2}\lambda_t^\top(G_t^\top-\overline{G}^\top)q^\diamond\right|\leq\lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2,\delta}$.

Event $E^G_{q^*}(\delta)$**:** *for all* $[t_1,\dots,t_2]\subseteq[1,\dots,T]$, $\left|\sum_{t=t_1}^{t_2}\lambda_t^\top(G_t^\top-\overline{G}^\top)q^*\right|\leq\lambda_{t_1,t_2}\mathcal{E}^G_{t_1,t_2,\delta}$.

$E^G_{q^\diamond}(\delta)$, $E^G_{q^*}(\delta)$ each hold with probability at least $1-\delta$ (See Lemma 5.4). We denote $\mathcal{E}^G_\delta := \mathcal{E}^G_{1,T,\delta}$.

C.2 Interval Regret

In this section, we provide the interval regret bounds attained by both primal and dual player in our specific framework.

C.2.1 Interval Regret of the Dual

In this section, we show the interval regret bound attained by dual algorithm. Recall that the dual variables are updated with projected online gradient descent as shown in Equation (5.1) or equivalently:

$$\lambda_{t+1,i} = \min\left\{\max\left\{0,\lambda_{t,i}+\eta[G_t^\top]_i\widehat{q}_t\right\},T^{1/4}\right\}, \tag{C.1}$$

where $\eta=\left[K\sqrt{T\ln\left(\frac{T^2}{\delta}\right)}\right]^{-1}$.
Let:

$$R^{\mathsf{D}}_{t_1,t_2}(\lambda) := \sum_{t=t_1}^{t_2}(\lambda-\lambda_t)^\top G_t^\top\widehat{q}_t,$$

denote the regret accumulated by OGD from episode t_1 to episode t_2 with respect to the constant multiplier λ. By standard analysis of OGD Orabona (2019) we have that:

$$R^{\mathsf{D}}_{t_1,t_2}(\lambda)\leq\frac{||\lambda_{t_1}-\lambda||_2^2}{2\eta}+\frac{\eta}{2}\sum_{t=t_1}^{t_2}||G_t^\top\widehat{q}_t||_2^2.$$

We can upper-bound the quantity $||G_t^\top\widehat{q}_t||_2^2$ as:

$$||G_t^\top\widehat{q}_t||_2^2=\sum_{i=1}^m\left(\sum_{x,a}g_{t,i}(x,a)\widehat{q}_t(x,a)\right)^2\leq\sum_{i=1}^m\left(\sum_{x,a}\widehat{q}_t(x,a)\right)^2\leq mL^2,$$

obtaining:

$$R^{\mathsf{D}}_{t_1,t_2}(\lambda)\leq D_1\frac{||\lambda_{t_1}-\lambda||_2^2}{\eta}+D_2\eta(t_2-t_1+1),$$

with $D_1 = \frac{1}{2}$, $D_2 = \frac{mL^2}{2}$.

We bound the distance between Lagrange multipliers for consecutive episodes.

Lemma C.1. *If the dual player employs projected online gradient descent as in Update (C.1), it holds:*

$$||\lambda_{t+1}||_1 - ||\lambda_t||_1 \leq m\eta L.$$

Proof. Since the dual minimizer is performing projected gradient descent with learning rate η, and the gradient of the Lagrangian at time t with respect to λ is equal to $\widehat{q}_t^\top G_t^\top$, element-wise it holds that:

$$\begin{aligned}
\lambda_{t+1,i} &= \min\left\{\max\left\{0, \lambda_{t,i} + \eta[G_t^\top]_i \widehat{q}_t\right\}, T^{\frac{1}{4}}\right\} \\
&\leq \max\left\{0, \lambda_{t,i} + \eta[G_t^\top]_i \widehat{q}_t\right\} \\
&\leq \max\left\{0, \lambda_{t,i} + \eta||[G_t^\top]_i||_\infty ||\widehat{q}_t||_1\right\} \\
&\leq \max\left\{0,\ \lambda_{t,i} + \eta L\right\} \\
&= \lambda_{t,i} + \eta L,
\end{aligned}$$

Thus,

$$||\lambda_{t+1}||_1 - ||\lambda_t||_1 = \sum_{i=1}^m \lambda_{t+1,i} - \sum_{i=1}^m \lambda_{t,i} \leq \sum_{i=1}^m \lambda_{t,i} + \sum_{i=1}^m \eta L - \sum_{i=1}^m \lambda_{t,i} = m\eta L.$$

This concludes the proof. □

C.2.2 Interval Regret of the Primal

Following the analysis of Theorem 5.1, it easy to obtain the following lemma.

Lemma C.2. *For any $q \in \cap_j \Delta(\mathcal{P}_j)$, the Projected OGD update:*

$$\widehat{q}_{t+1} = \Pi_{\Delta(\mathcal{P}_i)}\left(\widehat{q}_t - \eta_t \ell_t\right),$$

with $\eta_t = \frac{1}{\overline{\ell}_t C\sqrt{T}}$ and $\overline{\ell}_t = \max\{||\ell_t||_\infty\}_{t=1}^t$ ensures:

$$\sum_{t=t_1}^{t_2} \ell_t^\top(\widehat{q}_t - q) \leq U_1 \frac{\overline{\ell}_{t_2}}{2} C\sqrt{T} + U_2 \frac{\overline{\ell}_{t_1,t_2}}{2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}},$$

where $U_1 = 2L$, $U_2 = |X||A|$, $\overline{\ell}_{t_1,t_2} = \max\{||\ell_t||_\infty\}_{t=t_1}^{t_2}$.

Now, let:

$$\lambda_{t_1,t_2} := \max\{||\lambda_t||_1\}_{t=t_1}^{t_2}.$$

Then it holds $\overline{\ell}_{t_1,t_2} \leq 1 + \lambda_{t_1,t_2}$ and we can restate the interval regret of the primal in terms of the 1-norm of the Lagrange multipliers as:

$$\sum_{t=t_1}^{t_2} {r_t^{\mathcal{L}}}^\top(q - \widehat{q}_t) \leq U_1 \frac{(1 + \lambda_{1,t_2})}{2} C\sqrt{T} + U_2 \frac{(1 + \lambda_{t_1,t_2})}{2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}}. \tag{C.2}$$

C.3 Analysis with Stochastic Constraints

C.3.1 Lower Bound on the Dual Cumulative Utility

We start proving a useful lemma in which we lower bound the dual cumulative utility. This Lemma holds both for the stochastic constraints and the adversarial constraint setting.

Lemma C.3. *Under the event $E^{\widehat{q}}(\delta)$, the cumulative dual utility $\sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t$ is lower bounded as:*

$$\sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t \geq -\lambda_{1,T} \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}),$$

where $\lambda_{t_1,t_2} := \max\{\|\lambda_t\|_1\}_{t=t_1}^{t_2}$.

Proof. We exploit the fact that the dual is no-regret with respect to the $\underline{0}$ vector:

$$\begin{aligned}
\sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t &= \sum_{t=1}^{T} \lambda_t^\top G_t^\top (q_t - \widehat{q}_t) + \sum_{t=1}^{T} \lambda_t^\top G_t^\top \widehat{q}_t \\
&\geq \sum_{t=1}^{T} \lambda_t^\top G_t^\top (q_t - \widehat{q}_t) + \sum_{t=1}^{T} \underline{0}^\top G_t^\top \widehat{q}_t - R_T^{\mathsf{D}}(\underline{0}) \\
&\geq \sum_{t=1}^{T} - \underbrace{\|\lambda_t\|_1}_{\leq \lambda_{1,T}} \underbrace{\left\|G_t^\top\right\|_\infty}_{\leq 1} \|q_t - \widehat{q}_t\|_1 - R_T^{\mathsf{D}}(\underline{0}) \\
&\geq -\lambda_{1,T} \sum_{t=1}^{T} \|q_t - \widehat{q}_t\|_1 - R_T^{\mathsf{D}}(\underline{0}) \\
&\geq -\lambda_{1,T} \mathcal{E}_\delta^q - R_T^{\mathsf{D}}(\underline{0}),
\end{aligned}$$

where the last inequality holds under $E^{\widehat{q}}(\delta)$. □

C.3.2 Analysis when Condition 5.1 Holds

We start by introducing the notation $\widehat{v}_{t,i} := [G_t^\top]_i \widehat{q}_t$, that is the violation of the i-th constraint incurred by $\widehat{q}_t$. We further denote $\widehat{V}_{t,i} := \sum_{\tau=1}^{t} \widehat{v}_{\tau,i}$. Observe that, when Condition 5.1 holds, thanks to Theorem 5.2 we have $||\lambda_t||_1 \leq T^{\frac{1}{4}}$ for all t and thus $\lambda_{t,i} \leq T^{\frac{1}{4}}$. This means that $\lambda_{t,i}$ never gets past the upper extreme and the update of the dual is effectively equivalent to that of OGD working on the set $\mathrm{R}_{\geq 0}^m$:

$$\lambda_{t,i} = \max\{\lambda_{t,i} + \eta \widehat{v}_{t,i},\ 0\}.$$

Lemma C.4. *If Condition 5.1 holds, then for each episode $t \in [T]$ and each constraint i it holds:*

$$\lambda_{t,i} \geq \eta \widehat{V}_{t-1,i}.$$

Proof. We prove the result by induction. Suppose that the statement holds for episode t. Then

$$\lambda_{t+1,i} = \max\{\lambda_{t,i} + \eta \widehat{v}_{t,i},\ 0\}$$

$$\geq \lambda_{t,i} + \eta \widehat{v}_{t,i}$$
$$\geq \eta \widehat{V}_{t-1,i} + \eta \widehat{v}_{t,i}$$
$$= \eta \widehat{V}_{t,i}.$$

Observe that for $t = 1$ the statement holds as the sum on the RHS evaluates to 0. □

Lemma C.5. *If Condition 5.1 holds, under the events $E^{\Delta}(\delta)$, $E^{\widehat{q}}(\delta)$ and $E^{G}_{q^\diamond}(\delta)$ for the stochastic constraint setting and under the events $E^{\Delta}(\delta)$ and $E^{\widehat{q}}(\delta)$ for the adversarial constraints one, it holds:*

$$V_T \leq \widehat{V}_{T,i^*} + \mathcal{E}^q_\delta.$$

Proof. Let i^* denote the most violated constraint, e.g. $i^* = \arg\max_i \sum_{t=1}^T [G_t^\top q_t]_i$. Then we have:

$$\begin{aligned}
V_T &= \sum_{t=1}^T [G_t^\top q_t]_{i^*} \\
&= \sum_{t=1}^T [G_t^\top \widehat{q}_t]_{i^*} + \sum_{t=1}^T [G_t^\top (q_t - \widehat{q}_t)]_{i^*} \\
&= \widehat{V}_{T,i^*} + \sum_{t=1}^T [G_t^\top]_{i^*} (q_t - \widehat{q}_t) \\
&\leq \widehat{V}_{T,i^*} + \sum_{t=1}^T ||[G_t^\top]_{i^*}||_\infty ||q_t - \widehat{q}_t||_1 \\
&\leq \widehat{V}_{T,i^*} + \mathcal{E}^q_\delta,
\end{aligned}$$

where the last step holds under $E^{\widehat{q}}(\delta)$ since $||[G_t^\top]_{i^*}||_\infty \leq 1$. □

C.3.3 Analysis when Condition 5.1 Does Not Hold

Lemma C.6. *If Condition 5.1 does not hold, then*

$$\widehat{V}_{T,i} \leq (2 + 2L) \frac{1}{\eta} T^{\frac{1}{4}}, \qquad \forall T, i$$

holds under the event $E^{\Delta}(\delta)$ in the adversarial constraint setting and under the events $E^{\Delta}(\delta)$, $E^{G}_{q^\diamond}(\delta)$, in the stochastic constraint setting.

Proof. Assume events $E^{\Delta}(\delta)$, $E^{G}_{q^\diamond}(\delta)$ hold and suppose by absurd that $\widehat{V}_{T,i} = (2 + 2L + \epsilon) \frac{1}{\eta} T^{\frac{1}{4}}$, with $\epsilon > 0$, for some T and i.

We can lower bound the quantity $\sum_{t=1}^T r_t^{\mathcal{L}\top} \widehat{q}_t$:

$$\sum_{t=1}^T r_t^{\mathcal{L}\top} \widehat{q}_t = \underbrace{\sum_{t=1}^T r_t^\top q^\diamond}_{\geq 0} - \sum_{t=1}^T \lambda_t^\top G_t^\top q^\diamond - \sum_{t=1}^T r_t^{\mathcal{L}\top} (q^\diamond - \widehat{q}_t)$$

$$\geq \underbrace{-\sum_{t=1}^{T} \lambda_t^\top \overline{G}^\top q^\diamond}_{\geq 0} - \lambda_{1,T} \mathcal{E}_\delta^G - \sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t)$$

$$\geq -mT^{\frac{1}{4}} \mathcal{E}_\delta^G - \sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t), \tag{C.3}$$

where Inequality (C.3) holds since $||\lambda_t||_1 \leq mT^{\frac{1}{4}}$ by construction of the dual space. Observe that, if we are in the Adversarial setting, then from the (stronger) definition of ρ and $q^\diamond$ it holds $-\sum_{t=1}^{T} \lambda_t^\top G_t^\top q^\diamond \geq 0$ and we obtain the tighter bound,

$$\sum_{t=1}^{T} r_t^{\mathcal{L}\top} \widehat{q}_t \geq -\sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t).$$

The dual is no regret with respect to the vector $\tilde{\lambda}$, whose elements are 0 for $j \neq i$ and $T^{\frac{1}{4}}$ in position $j = i$:

$$\begin{aligned}
\sum_{t=1}^{T} r_t^{\mathcal{L}\top} \widehat{q}_t &= \sum_{t=1}^{T} r_t^\top \widehat{q}_t - \sum_{t=1}^{T} \lambda_t^\top G_t^\top \widehat{q}_t \\
&\leq \sum_{t=1}^{T} r_t^\top \widehat{q}_t - \sum_{t=1}^{T} \tilde{\lambda}^\top G_t^\top \widehat{q}_t + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
&= \sum_{t=1}^{T} r_t^\top \widehat{q}_t - T^{\frac{1}{4}} \sum_{t=1}^{T} [G_t^\top \widehat{q}_t]_i + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
&\leq LT - T^{\frac{1}{4}} \widehat{V}_{T,i} + R_T^{\mathsf{D}}(\tilde{\lambda}).
\end{aligned}$$

Combining the bounds we have:

$$\begin{aligned}
-mT^{\frac{1}{4}} \mathcal{E}_\delta^G - \sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t) &\leq LT - T^{\frac{1}{4}} \widehat{V}_{T,i} + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
T^{\frac{1}{4}} \widehat{V}_{T,i} &\leq LT + mT^{\frac{1}{4}} \mathcal{E}_\delta^G + \sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t) + R_T^{\mathsf{D}}(\tilde{\lambda}) \\
\frac{\sqrt{T}}{\eta}(2 + 2L + \epsilon) &\leq LT + mT^{\frac{1}{4}} \mathcal{E}_\delta^G + \sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t) + R_T^{\mathsf{D}}(\tilde{\lambda}).
\end{aligned} \tag{C.4}$$

Observe that:

$$\begin{aligned}
R_T^{\mathsf{D}}(\tilde{\lambda}) &\leq \frac{1}{2} \frac{||\tilde{\lambda}||_2^2}{\eta} + \frac{mL^2}{2} \eta T \\
&= \frac{\sqrt{T}}{2\eta} + \frac{mL^2}{2} \frac{1}{100m|X||A|\sqrt{T \ln\left(\frac{T^2}{\delta}\right)}} T
\end{aligned}$$

$$\leq L\frac{\sqrt{T}}{\eta},$$

since $|X| \geq L$.

For the primal it holds by Lemma C.2:

$$\begin{aligned}
\sum_{t=1}^{T} r_t^{\mathcal{L}\top}(q^\diamond - \widehat{q}_t) = \sum_{t=1}^{T} \ell_t{}^\top(\widehat{q}_t - q^\diamond) \\
&\leq \lambda_{1,T} U_1 C\sqrt{T} + \lambda_{1,T} U_2 \frac{\sqrt{T}}{C} \\
&\leq mT^{\frac{1}{4}}\sqrt{T}\left(U_1 C + \frac{U_2}{C}\right) \\
&= m\left(U_1\frac{U_2}{5} + 5\right)\sqrt{T}\,T^{\frac{1}{4}} \\
&= m\left(2L\frac{|X||A|}{5} + 5\right)\sqrt{T}\,T^{\frac{1}{4}} \\
&\leq 6mL|X||A|\sqrt{T}\,T^{\frac{1}{4}} \\
&\leq \frac{L}{\eta}T^{\frac{1}{4}} \\
&\leq L\frac{\sqrt{T}}{\eta}.
\end{aligned}$$

For the Azuma-Hoeffding term it holds:

$$mT^{\frac{1}{4}}\mathcal{E}_\delta^G = mT^{\frac{1}{4}}2L\sqrt{2T\ln\left(\frac{T^2}{\delta}\right)} \leq \frac{1}{\eta}T^{\frac{1}{4}} = \frac{\sqrt{T}}{\eta}.$$

Observe that $LT \leq \frac{\sqrt{T}}{\eta}$ holds trivially.
Dividing both the terms in Equation (C.4) by $\frac{\sqrt{T}}{\eta}$, we obtain

$$2 + 2L + \epsilon \leq 2 + 2L,$$

which is absurd. □

APPENDIX D

Omitted Lemmas and Proofs of Chapter 6

D.1 Dictionary

In the following, we provide the definition of different quantities which will be employed in the rest of the chapter. This is done for the ease of presentation.

- **Quantity $\mathcal{E}^P_{t_1,t_2}$:**

  $$\mathcal{E}^P_{t_1,t_2} = U_1 \Xi_{t_1,t_2} C\sqrt{T} + U_2 \Xi_{t_1,t_2} \frac{(t_2 - t_1 + 1)}{C\sqrt{T}} + U_3 \Xi_{t_1,t_2} \frac{1}{C\sqrt{T}} + U_4 \Xi_{t_1,t_2} \sqrt{T},$$

  where:

  - $U_1 = 6L^2 \ln\left(\frac{L|A|T^2}{\delta}\right)$
  - $U_2 = 9L|X||A|$
  - $U_3 = \frac{L}{2} \ln\left(\frac{LT^2}{\delta}\right)$
  - $U_4 = 30L^2|X|^2 \sqrt{2|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)}$.

  With probability at least $1 - 4\delta$ it holds $R^P_{t_1,t_2} \leq \mathcal{E}^P_{t_1,t_2}$, $\forall t_1, t_2 \in [T] : 1 \leq t_1 \leq t_2 \leq T$ by Theorem 6.1.

- **Quantity $\mathcal{E}^D(\underline{0})$:**

  $$\mathcal{E}^D(\underline{0}) = D_1 \frac{\|\lambda_{t_1}\|_2^2}{\eta} + D_2 \eta (t_2 - t_1 + 1),$$

  where:

- $D_1 = \frac{1}{2}$
- $D_2 = \frac{mL^2}{2}$.

It holds $R^D_{t_1,t_2}(\underline{0}) \leq \mathcal{E}^D(\underline{0}),\ \forall t_1, t_2 \in [T] : 1 \leq t_1 \leq t_2 \leq T$ by Theorem D.2.

- **Quantity $\mathcal{E}^G_{t_1,t_2}$:**

$$\mathcal{E}^G_{t_1,t_2} = B_1\sqrt{(t_2 - t_1 + 1)},$$

where:

- $B_1 = 2L\sqrt{\ln\left(\frac{T^2}{\delta}\right)}$.

Given a $q \in \Delta(M)$, with probability at least $1-\delta$ it holds in case of stochastic constraints $\sum_{t=t_1}^{t_2}(G_t^\top q - \overline{G}^\top q) \leq \mathcal{E}^G_{t_1,t_2}$, by Azuma-Hoeffding inequality.

- **Quantity $\mathcal{E}^{\mathbb{I}}_{t_1,t_2}$:**

$$\mathcal{E}^{\mathbb{I}}_{t_1,t_2} = F_1\sqrt{(t_2 - t_1 + 1)},$$

where:

- $F_1 = L\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}$.

With probability at least $1-\delta$ it holds $\sum_{t=t_1}^{t_2}\sum_{x,a}(\mathbb{I}_t(x,a) - q_t(x,a)) \leq \mathcal{E}^{\mathbb{I}}_{t_1,t_2}$, and with probability at least $1-\delta$ it holds $\sum_{t=t_1}^{t_2}\sum_{x,a}(q_t(x,a) - \mathbb{I}_t(x,a)) \leq \mathcal{E}^{\mathbb{I}}_{t_1,t_2}$, by Azuma-Hoeffding inequality.

- **Quantity C:**

$$C = 252|X||A|L$$

- **Quantity D:**

$$\begin{aligned} D &= 84672mL^2|X|^2|A| \\ &= 336mL|X|C. \end{aligned}$$

D.2 Additional Notation

We introduce a Q-function of a generic function f as:

$$\begin{cases} Q(x,a;f) = f(x,a) + \mathbb{E}_{x'\sim P(\cdot|x,a)}\left[V^\pi(x';f)\right] \\ V^\pi(x;f) = \mathbb{E}_{a\sim\pi(\cdot|x)}\left[Q^\pi(x,a;f)\right] \\ V^\pi(x_L;f) = 0. \end{cases}$$

In addition we will use the notation $Q_t(x,a)$ to indicate the Q-function computed with respect to the function ℓ_t, *i.e.* $Q(x,a;\ell_t)$.

D.3 Omitted Proofs for The Primal Algorithm

In this section, we study the guarantees attained by the primal procedure, which we completely provide in Algorithm D.1.

Algorithm D.1 Fixed Share Policy Optimization with Dilated Bonus (`FS-PODB`)

Require: $X, A, \sigma = \frac{1}{T}, C$

$\mathcal{P}_1 \leftarrow$ set of all possible transitions

$\pi_1(a|x) = \frac{1}{|A|} \quad \forall (x,a) \in X \times A$

$\Xi_0 \leftarrow 1$

$\gamma \leftarrow \frac{1}{C\sqrt{T}}$

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

Play π_t, observe $\{(x_k, a_k)\}_{k=0}^{L-1}$, losses $\{\ell_t(x_k, a_k)\}_{k=0}^{L-1}$ and Ξ_t

$\eta_t \leftarrow \frac{1}{2L\Xi_t C\sqrt{T}}$

For all $k = 0, \dots, L-1$ and $(x,a) \in X_k \times A$:

$$L_{t,k} = \sum_{j=k}^{L-1} \ell_t(x_k, a_k)$$

$$\widehat{Q}_t(x,a) = \frac{L_{t,k}}{\overline{q}_t(x,a) + \gamma} \mathbb{I}_t(x,a),$$

where $\overline{q}_t(x,a) = \max_{\widehat{P} \in \mathcal{P}_t} q^{\widehat{P}, \pi_t}(x,a)$ and $\mathbb{I}_t(x,a) = \mathbb{I}\{x_{t,k} = x, a_{t,k} = a\}$.

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

$$b_t(x) = \mathbb{E}_{a \sim \pi_t(\cdot|x)} \left[\frac{3\gamma L\Xi_t + L\Xi_t \left(\overline{q}_t(x,a) - \underline{q}_t(x,a) \right)}{\overline{q}_t(x,a) + \gamma} \right]$$

$$B_t(x,a) = b_t(x) + \left(1 + \frac{1}{L}\right) \max_{\widehat{P} \in \mathcal{P}_t} \mathbb{E}_{x' \sim \widehat{P}(\cdot|x,a)} \mathbb{E}_{a \sim \pi_t(\cdot|x')} \left[B_t(x', a') \right]$$

where $\underline{q}_t = \min_{\widehat{P} \in \mathcal{P}_t} q^{\widehat{P}, \pi_t}(x,a)$, and $B_t(x_H, a) = 0$ for all a.

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

$$w_{t+1}(a|x) = (1-\sigma) w_t(a|x) e^{-\eta_t \left(\widehat{Q}_t(x,a) - B_t(x,a) \right)} + \frac{\sigma}{|A|} \sum_{a' \in A} w_t(a'|x) e^{-\eta_t \left(\widehat{Q}_t(x,a') - B_t(x,a') \right)}.$$

$$\pi_{t+1}(a|x) = \frac{w_{t+1}(a|x)}{\sum_{a' \in A} w_{t+1}(a'|x)}.$$

$\mathcal{P}_{t+1} \leftarrow$`TRANSITION.UPDATE`$(\{(x_k, a_k)\}_{k=0}^{L-1})$

end for

Theorem D.1. *For any $\delta \in (0,1)$, Algorithm D.1 attains, with probability at least $1 - 4\delta$ and for all $[t_1, \dots, t_2] \subset [T]$:*

$$\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \left(\pi_t(a|x) - \pi^*(a|x) \right) \left(Q_t^{\pi_t}(x,a) - B_t(x,a) \right)$$

$$= \Xi_{t_1,t_2} o(T) + \sum_{t=t_1}^{t_2} V^{\pi^*}(b_t) + \frac{1}{L} \sum_{t=t_1}^{t_2} \sum_{x,a} q^*(x)\pi_t(a|x)B_t(x,a),$$

for all $t_1, t_2 \in [T]$ *s.t.* $1 \le t_1 \le t_2 \le T$ *and where* $\Xi_{t_1,t_2} \ge \max_{t\in[t_1,\dots,t_2]} \max_{x,a} \ell_t(x,a)$.

Proof. In the rest of the proof, we will refer as $\bar{L}_t$ to $\max_{\tau\in[t]} \max_{k\in[L]} L_{\tau,k}$ and $\bar{L}_{t_1,t_2}$ to $\max_{\tau\in[t_1,\dots,t_2]} \max_{k\in[L]} L_{\tau,k}$; therefore, by definition it holds $\bar{L}_t \le L\Xi_t$ for all $t \in [T]$.
As a first step, we decompose

$$\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \left(\pi_t(a|x) - \pi^*(a|x)\right) \left(Q_t^{\pi_t}(x,a) - B_t(x,a)\right)$$

in three different quantities, as follows:

$$\begin{aligned}
&\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \left(\pi_t(a|x) - \pi^*(a|x)\right) \left(Q_t^{\pi_t}(x,a) - B_t(x,a)\right) \\
&\quad = \underbrace{\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \left(\pi_t(a|x) - \pi^*(a|x)\right) \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)}_{①} \\
&\quad + \underbrace{\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \pi_t(a|x) \left(Q_t^{\pi_t}(x,a) - \widehat{Q}_t(x,a)\right)}_{②} \\
&\quad + \underbrace{\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \pi^*(a|x) \left(\widehat{Q}_t(x,a) - Q_t^{\pi_t}(x,a)\right)}_{③},
\end{aligned}$$

which we proceed to bound separately.

Bound on ①. The quantity of interest can be bounded after noticing that Algorithm D.1 employs a slightly modified version of OMD. In fact, recalling the definition of π_t, we can write:

$$\begin{aligned}
&\pi_{t+1}(a|x) \\
&= \frac{w_{t+1}(a|x)}{\sum_{a'} w_{t+1}(a'|x)} \\
&= \frac{(1-\sigma) w_t(a|x) e^{-\eta_t\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)} + \frac{\sigma}{|A|} \sum_{a'\in A} w_t(a'|x) e^{-\eta_t\left(\widehat{Q}_t(x,a') - B_t(x,a')\right)}}{\sum_{a'\in A} w_t(a'|x) e^{-\eta_t\left(\widehat{Q}_t(x,a') - B_t(x,a')\right)}} \\
&= (1-\sigma) \frac{\pi_t(a|x) e^{-\eta\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)}}{\sum_{a'} \pi_t(a'|x) e^{-\eta\left(\widehat{Q}_t(x,a') - B_t(x,a')\right)}} + \sigma \frac{1}{|A|}.
\end{aligned}$$

From now on we will refer to $\frac{\pi_t(a|x)e^{-\eta(\widehat{Q}_t(x,a)-B_t(x,a))}}{\sum_{a'}\pi_t(a'|x)e^{-\eta(\widehat{Q}_t(x,a')-B_t(x,a'))}}$ as $\widetilde{\pi}_{t+1}(x,a)$. Thus,

$$\pi_{t+1}(a|x) = (1-\sigma)\widetilde{\pi}_{t+1}(x,a) + \frac{\sigma}{|A|}.$$

Calling $\psi(\cdot)$ the negative entropy function defined as $\psi(\pi(\cdot|x)) := \sum_a \pi(a|x)\ln(\pi(a|x))$, by standard analysis (*e.g.* Orabona (2019)), it holds:

$$\widetilde{\pi}_{t+1}(\cdot|x) = \operatorname*{arg\,min}_{\pi(\cdot|x)\in\Delta(A)} \sum_a \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)\pi(a|x) + \frac{1}{\eta}D_\psi(\pi(\cdot|x);\pi_t(\cdot|x)),$$

where D_ψ is Bregman divergence w.r.t. the negative entropy function $\psi(\cdot)$. Thus, for all $\pi(\cdot|x)$ it holds $\eta_t\sum_a\left(\widehat{Q}_t(x,a)-B_t(x,a)\right)(\pi(a|x)-\widetilde{\pi}_{t+1}(x,a)) + \langle\nabla\psi(\widetilde{\pi}_{t+1}(\cdot|x)) - \nabla\psi(\pi_t(\cdot|x)), \pi(\cdot|x)-\widetilde{\pi}_{t+1}(\cdot|x)\rangle \geq 0$. So, for all $\pi(\cdot|x)$ the following holds:

$$\begin{aligned}
&\eta_t\langle\widehat{Q}_t(x,\cdot) - B_t(x,\cdot), \pi_t(\cdot|x) - \pi(\cdot|x)\rangle \\
&= \eta_t\langle\widehat{Q}_t(x,\cdot) - B_t(x,\cdot) + \nabla\psi(\widetilde{\pi}_{t+1}(\cdot|x)) - \nabla\psi(\pi_t(\cdot|x)), \widetilde{\pi}_{t+1}(\cdot|x) - \pi(\cdot|x)\rangle \\
&\quad + \eta_t\langle\widehat{Q}_t(x,\cdot) - B_t(x,\cdot), \pi_t(\cdot|x) - \widetilde{\pi}_{t+1}(\cdot|x)\rangle \\
&\quad + \langle\nabla\psi(\widetilde{\pi}_{t+1}(\cdot|x)) - \nabla\psi(\pi_t(\cdot|x)), \pi(\cdot|x) - \widetilde{\pi}_{t+1}(\cdot|x)\rangle \\
&\leq \langle\eta_t\left(\widehat{Q}_t(x,\cdot) - B_t(x,\cdot)\right), \pi_t(\cdot|x) - \widetilde{\pi}_{t+1}(\cdot|x)\rangle \\
&\quad + \langle\nabla\psi(\widetilde{\pi}_{t+1}(\cdot|x)) - \nabla\psi(\pi_t(\cdot|x)), \pi(\cdot|x) - \widetilde{\pi}_{t+1}(\cdot|x)\rangle \\
&\leq D_\psi(\pi(\cdot|x);\pi_t(\cdot|x)) - D_\psi(\pi(\cdot|x);\widetilde{\pi}_{t+1}(\cdot|x)) - D_\psi(\widetilde{\pi}_{t+1}(\cdot|x);\pi_t(\cdot|x)) \\
&\quad + \eta_t\langle\widehat{Q}_t(x,\cdot) - B_t(x,\cdot), \pi_t(\cdot|x) - \widetilde{\pi}_{t+1}(\cdot|x)\rangle \qquad \text{(D.1)} \\
&= D_\psi(\pi(\cdot|x);\pi_t(\cdot|x)) - D_\psi(\pi(\cdot|x);\widetilde{\pi}_{t+1}(\cdot|x)) \\
&\quad + \frac{\eta_t^2}{2}\sum_{a\in A}\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2\pi_t(a|x), \qquad \text{(D.2)}
\end{aligned}$$

where Inequality (D.1) and Inequality (D.2) are based on the proofs of Lemma 6.6. and Lemma 6.9. in Orabona (2019).
Additionally we can show that for all $t \in [T]$: $D_\psi(\pi(\cdot|x);\pi_t(\cdot|x)) - D_\psi(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x)) \leq \sigma\ln(|A|)$. Indeed,

$$\begin{aligned}
&D_\psi\left(\pi(\cdot|x);\pi_t(\cdot|x)\right) - D_\psi\left(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x)\right) \\
&\qquad = D_\psi\left(\pi(\cdot|x);(1-\sigma)\widetilde{\pi}_t(\cdot|x) + \sigma\pi^{\frac{1}{|A|}}\right) - D_\psi\left(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x)\right) \\
&\qquad \leq \sigma D_\psi\left(\pi(\cdot|x);\pi^{\frac{1}{|A|}}\right) - \sigma D_\psi\left(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x)\right) \\
&\qquad \leq \sigma\ln(|A|),
\end{aligned}$$

where the last inequality holds since $D_\psi(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x)) \geq 0$ and

$$D_\psi(\pi(\cdot|x);\pi^{\frac{1}{|A|}}) = \sum_{a\in A}\pi(a|x)\ln\left(\frac{\pi(a|x)}{\pi^{\frac{1}{|A|}}(a|x)}\right)$$

$$\leq \sum_{a\in A} \pi(a|x) \ln\left(\frac{1}{\pi^{\frac{1}{|A|}}(a|x)}\right)$$
$$= \sum_{a\in A} \pi(a|x) \ln\left(|A|\right)$$
$$= \ln(|A|).$$

Notice that with we refer as $\pi^{\frac{1}{|A|}}$ to the vector strategy in $[0,1]^{|A|}$ with all elements equal to $\frac{1}{|A|}$.
Moreover we bound $D_\psi(\pi(\cdot|x);\pi_{t_1}(\cdot|x))$, since $\pi_{t_1}(a|x) = (1-\sigma)\widetilde{\pi}_{t_1}(a|x)+\sigma(\frac{1}{|A|}) \geq \frac{\sigma}{|A|}$, as follows:

$$D_\psi(\pi(\cdot|x);\pi_{t_1}(\cdot|x)) = \sum_{a\in A} \pi(a|x) \ln\left(\frac{\pi(a|x)}{\pi_{t_1}(a|x)}\right)$$
$$\leq \sum_{a\in A} \pi(a|x) \ln\left(\frac{1}{\pi_{t_1}(a|x)}\right)$$
$$\leq \sum_{a\in A} \pi(a|x) \ln\left(\frac{|A|}{\sigma}\right)$$
$$= \ln\left(\frac{|A|}{\sigma}\right).$$

Putting everything together we have that:

$$① = \sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \left(\pi_t(a|x) - \pi^*(a|x)\right)\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)$$
$$\leq \sum_x q^*(x) \Bigg(\frac{D_\psi(\pi(\cdot|x);\pi_{t_1}(\cdot|x))}{\eta_{t_1}}$$
$$+ \sum_{t=t_1+1}^{t_2} \left(\frac{D_\psi(\pi(\cdot|x);\pi_t(\cdot|x))}{\eta_t} - \frac{D_\psi(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x))}{\eta_{t+1}} \right) \Bigg)$$
$$+ \sum_x q^*(x) \sum_{t=t_1}^{t_2} \frac{\eta_t}{2} \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \pi_t(a|x) \tag{D.3a}$$
$$\leq \sum_x q^*(x) \Bigg(\frac{D_\psi(\pi(\cdot|x);\pi_{t_1}(\cdot|x))}{\eta_{t_1}} \tag{D.3b}$$
$$+ \sum_{t=t_1+1}^{t_2} \left(\frac{D_\psi(\pi(\cdot|x);\pi_t(\cdot|x)) - D_\psi(\pi(\cdot|x);\widetilde{\pi}_t(\cdot|x))}{\eta_t} \right) \Bigg)$$
$$+ \sum_x q^*(x) \sum_{t=t_1}^{t_2} \frac{\eta_t}{2} \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \pi_t(a|x) \tag{D.3c}$$

$$\leq \frac{\ln\left(\frac{|A|}{\sigma}\right)}{\eta_{t_1}} + \sigma \sum_{t=t_1+1}^{t_2} \frac{\ln(|A|)}{\eta_{t_2}} + \sum_x q^*(x) \sum_{t=t_1}^{t_2} \frac{\eta_t}{2} \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \pi_t(a|x) \tag{D.3d}$$

$$\leq \frac{\ln\left(\frac{|A|}{\sigma}\right)}{\eta_{t_1}} + \frac{\sigma T \ln(|A|)}{\eta_{t_2}} + \sum_x q^*(x) \sum_{t=t_1}^{t_2} \frac{\eta_t}{2} \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \pi_t(a|x)$$

$$= \frac{\ln(|A|T)}{\eta_{t_1}} + \frac{\ln(|A|)}{\eta_{t_2}} + \sum_x q^*(x) \sum_{t=t_1}^{t_2} \frac{\eta_t}{2} \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \pi_t(a|x),$$

where $\sigma = \frac{1}{T}$, Inequality (D.3a) holds by Inequality (D.2) , Inequality (D.3c) holds since $\frac{1}{\eta_{t+1}} \geq \frac{1}{\eta_t}$ for all $t \in [T]$, and Inequality (D.3d) holds since $\eta_{t_2} \leq \eta_t$ for all t in $[t_1+1, \ldots, t_2]$. Focusing now on the last part of the right term, with probability at least $1-2\delta$ the following holds:

$$\sum_x \sum_a q^*(x) \sum_{t=t_1}^{t_2} \frac{\eta_t}{2} \left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \pi_t(a|x)$$

$$\leq \sum_{t=t_1}^{t_2} \eta_t \sum_x \sum_a q^*(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2 + \sum_{t=t_1}^{t_2} \eta_t \sum_x \sum_a q^*(x)\pi_t(a|x)B_t(x,a)^2 \tag{D.4a}$$

$$= \sum_{t=t_1}^{t_2} \eta_t \sum_x \sum_a q^*(x)\pi_t(a|x) \frac{L_{t,k}^2}{(\overline{q}_t(x,a)+\gamma)^2} \mathbb{I}_t(x,a) + \sum_{t=t_1}^{t_2} \eta_t \sum_x \sum_a q^*(x)\pi_t(a|x)B_t(x,a)^2 \tag{D.4b}$$

$$\leq \bar{L}_{t_1,t_2} \sum_{t=t_1}^{t_2} \eta_t \bar{L}_t \sum_x \sum_a \frac{q^*(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma} \frac{\mathbb{I}_t(x,a)}{\overline{q}_t(x,a)+\gamma} + \sum_{t=t_1}^{t_2} \eta_t \sum_x \sum_a q^*(x)\pi_t(a|x)B_t(x,a)^2 \tag{D.4c}$$

$$\leq \frac{\gamma}{2L} \bar{L}_{t_1,t_2} \sum_{t=t_1}^{t_2} \sum_x \sum_a \frac{q^*(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma} \frac{q_t(x,a)}{\overline{q}_t(x,a)+\gamma} + \frac{\gamma \bar{L}_{t_1,t_2}}{2} \ln\left(\frac{LT^2}{\delta}\right) + \sum_{t=t_1}^{t_2} \eta_t \sum_x \sum_a q^*(x)\pi_t(a|x)B_t(x,a)^2 \tag{D.4d}$$

$$\leq \sum_{t=t_1}^{t_2} \sum_x \sum_a q^*(x)\pi_t(a|x) \frac{\gamma \Xi_{t_1,t_2}}{2(\overline{q}_t(x,a)+\gamma)} + \frac{\gamma \bar{L}_{t_1,t_2}}{2} \ln\left(\frac{LT^2}{\delta}\right)$$

$$+ \frac{1}{2L} \sum_{t=t_1}^{t_2} \sum_x \sum_a q^*(x)\pi_t(a|x) B_t(x,a) \tag{D.4e}$$

$$= \sum_{t=t_1}^{t_2} \sum_x \sum_a q^*(x)\pi_t(a|x) \left(\frac{\gamma \Xi_{t_1,t_2}}{2(\overline{q}_t(x,a)+\gamma)} + \frac{B_t(x,a)}{2L}\right) + \frac{\gamma \bar{L}_{t_1,t_2}}{2} \ln\left(\frac{LT^2}{\delta}\right),$$

where Inequality (D.4a) holds since $(a-b)^2 \leq 2a^2 + 2b^2$, for all $a, b \in \mathbb{R}$, Equality (D.4b) holds by definition of $\widehat{Q}(x,a)$, Inequality (D.4c) is motivated by the fact that $L_{t,k} \leq \bar{L}_{t_1,t_2}$ by its definition, Inequality (D.4d) holds with probability at least $1-\delta$ by applying Lemma D.9 and taking $\alpha_t(x,a) = \frac{q^*(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma}$ since $\frac{q^*(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma} \leq \frac{1}{\gamma}$ and considering that by definition $\eta_t \Xi_t = \frac{\gamma}{2L}$, and finally Inequality (D.4e) holds since $\overline{q}_t(x,a) \geq q_t(x,a)$, $\forall (x,a) \in X \times A, \forall t \in [t_1 \dots t_2]$ with probability at least $1-\delta$ by definition of the confidence sets and by Lemma D.6. Setting $\gamma = 2\eta_t L \Xi_t$, we can conclude that, with probability at least $1-2\delta$, ① is bounded as:

$$\frac{4L\Xi_{t_1,t_2} \ln(|A|T)}{\gamma} + \gamma \frac{L\Xi_{t_1,t_2}}{2} \ln\left(\frac{LT^2}{\delta}\right)$$

$$+ \sum_{t=t_1}^{t_2} \sum_x \sum_a q^*(x)\pi_t(a|x) \left(\frac{\gamma \Xi_{t_1,t_2}}{2(\overline{q}_t(x,a)+\gamma)} + \frac{B_t(x,a)}{2L}\right).$$

Bound on ②. To bound ② we employ the same approach as in (Luo et al., 2021). First we define Y_t as $Y_t := \sum_x \sum_a q^*(x)\pi_t(a|x)\widehat{Q}_t(x,a)$, for all $t \in [T]$. Now since $\sum_{t=t_1}^{t_2} Y_t$ is a martingale sequence , we apply Freedman's inequality. First notice that under the event $P \in \mathcal{P}_t$ for all $t \in [T]$:

$$\mathbb{E}[Y_t^2] \leq \mathbb{E}_t \left[\left(\sum_x \sum_a q^*(x)\pi_t(a|x)\widehat{Q}_t(x,a)\right)^2\right]$$

$$\leq \mathbb{E}_t \left[\left(\sum_x \sum_a q^*(x)\pi_t(a|x)\right)\left(\sum_x \sum_a q^*(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2\right)\right]$$

$$= L\mathbb{E}_t \left[\sum_x \sum_a q^*(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2\right]$$

$$= L \sum_x \sum_a q^*(x)\pi_t(a|x) \frac{L_{t,k}^2}{(\overline{q}_t(x,a)+\gamma)^2} q_t(x,a)$$

$$\leq \sum_x \sum_a q^*(x)\pi_t(a|x) \frac{L\bar{L}_t^2}{\overline{q}_t(x,a)+\gamma}.$$

Thus, thanks to Lemma D.8, since $|Y_t| \leq L \sup_{x',a'} \widehat{Q}_t(x,a) \leq \frac{L\bar{L}_t}{\gamma}$, with probability

at least $1-\delta$ it holds simultaneously for all $t_1, t_2 : 1 \leq t_1 \leq t_2 \leq T$:

$$\sum_{t=t_1}^{t_2} (\mathbb{E}_t[Y_t] - Y_t)$$
$$\leq \frac{\gamma}{L\bar{L}_{t_1,t_2}} \sum_{t=t_1}^{t_2} \sum_x \sum_a q^*(x)\pi_t(a|x) \frac{L\bar{L}_t^2}{\overline{q}_t(x,a)+\gamma} + \frac{L\bar{L}_{t_1,t_2}}{\gamma} \ln\left(\frac{T^2}{\delta}\right).$$

We notice also the following result with probability at least $1-\delta$ for all $t \in [T]$:

$$\sum_x \sum_a q^*(x)\pi_t(a|x)Q_t(x,a) - \mathbb{E}[Y_t]$$
$$= \sum_x \sum_a q^*(x)\pi_t(a|x)Q_t(x,a) - \mathbb{E}\left[\sum_x \sum_a q^*(x)\pi_t(a|x)\widehat{Q}_t(x,a)\right]$$
$$= \sum_x \sum_a q^*(x)\pi_t(a|x)Q_t(x,a)\left(1 - \frac{q_t(x,a)}{\overline{q}_t(x,a)+\gamma}\right)$$
$$\leq \sum_x \sum_a q^*(x)\pi_t(a|x)L\Xi_t\left(\frac{\overline{q}_t(x,a) - q_t(x,a) + \gamma}{\overline{q}_t(x,a)+\gamma}\right)$$
$$\leq \sum_x \sum_a q^*(x)\pi_t(a|x)L\Xi_t\left(\frac{(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + \gamma}{\overline{q}_t(x,a)+\gamma}\right).$$

Finally we can bound ② with probability at least $1-2\delta$ as follows.

$$② = \sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \pi_t(a|x)\left(Q_t^{\pi_t}(x,a) - \widehat{Q}_t(x,a)\right)$$
$$= \sum_{t=t_1}^{t_2} (\mathbb{E}_t[Y_t] - Y_t) + \sum_{t=t_1}^{t_2}\left(\sum_x \sum_a q^*(x)\pi_t(a|x)Q_t(x,a) - \mathbb{E}[Y_t]\right)$$
$$\leq \sum_{t=t_1}^{t_2} \sum_x \sum_a q^*(x)\pi_t(a|x)L\Xi_t\left(\frac{(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + 2\gamma}{\overline{q}_t(x,a)+\gamma}\right) + \frac{L\bar{L}_{t_1,t_2}}{\gamma}\ln\left(\frac{T^2}{\delta}\right).$$

Bound on ③. With probability at least $1-2\delta$ it holds:

$$③ = \sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \pi^*(a|x)\left(\widehat{Q}_t(x,a) - Q_t^{\pi_t}(x,a)\right) \leq \frac{L^2\Xi_{t_1,t_2}}{2\gamma}\ln\left(\frac{LT^2}{\delta}\right),$$

by Corollary D.1.

Conclusion of the proof. Finally we notice that, with probability at least $1-4\delta$, we have the following result.

$$\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a (\pi_t(a|x) - \pi^*(a|x))\left(Q_t^{\pi_t}(x,a) - B_t(x,a)\right) = ① + ② + ③$$

$$\leq \gamma \frac{L\Xi_{t_1,t_2}}{2} \ln\left(\frac{HT^2}{\delta}\right) + \frac{6L^2\Xi_{t_1,t_2}}{\gamma} \ln\left(\frac{L|A|T^2}{\delta}\right) + \sum_{t=t_1}^{t_2} \sum_{x,a} q^*(x)\pi_t(a|x) \left(\frac{\Xi_t(3\gamma L + L(\overline{q}_t(x,a) - \underline{q}_t(x,a)))}{\overline{q}_t(x,a) + \gamma} + \frac{B_t(x,a)}{L}\right).$$

This concludes the proof. □

D.4 Omitted Proofs for the Dual Algorithm

In this section, we study the guarantees attained by the dual procedure.

Theorem D.2. *When employed by Algorithm 6.1, online projected gradient descent (OGD) attains:*

$$R^D_{t_1,t_2}(\lambda) = \sum_{t=t_1}^{t_2} (\lambda - \lambda_t)^\top \sum_{k=0}^{L-1} G_t(x_k, a_k) \leq \frac{\|\lambda_{t_1} - \lambda\|_2^2}{2\eta} + \frac{\eta}{2}(t_2 - t_1 + 1)mL^2.$$

Proof. We proceed to prove the theorem following (Orabona, 2019). Indeed, it holds:

$$\begin{aligned} R^D_{t_1,t_2}(\lambda) &= \sum_{t=t_1}^{t_2} (\lambda - \lambda_t)^\top \sum_{k=0}^{L-1} G_t(x_k, a_k) \\ &\leq \frac{\|\lambda_{t_1} - \lambda\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=t_1}^{t_2} \|\sum_{k=0}^{L-1} G_t(x_k, a_k)\|_2^2 \\ &\leq \frac{\|\lambda_{t_1} - \lambda\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=t_1}^{t_2} \sum_{i=1}^{m} \left(\sum_{k=0}^{L-1} g_{t,i}(x_k, a_k)\right)^2 \\ &\leq \frac{\|\lambda_{t_1} - \lambda\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=t_1}^{t_2} mL^2 \\ &\leq \frac{\|\lambda_{t_1} - \lambda\|_2^2}{2\eta} + \frac{\eta}{2}(t_2 - t_1 + 1)mL^2. \end{aligned}$$

This concludes the proof. □

We conclude with following result.

Lemma D.1. *When employed by Algorithm 6.1, online projected gradient descent (OGD) guarantees for all $t \in [T]$:*

$$\|\lambda_{t+1}\|_1 - \|\lambda_t\|_1 \leq mL\eta.$$

Proof. It holds:

$$\lambda_{t+1,i} = \min\left\{\max\left\{0, \lambda_{t,i} + \eta \sum_{k=0}^{L-1} g_{t,i}(x_k, a_k)\right\}, T^{\frac{1}{4}}\right\}$$

$$
\begin{aligned}
&\leq \max\left\{0, \lambda_{t,i} + \eta \sum_{h=0}^{L-1} g_{t,i}(x_k, a_k)\right\} \\
&\leq \max\left\{0, \lambda_{t,i} + \eta \sum_{k=0}^{L-1} 1\right\} \\
&= \lambda_{t,i} + \eta L,
\end{aligned}
$$

which concludes the proof when we take the sum over all $i \in [m]$. □

D.5 Preliminary Results

In this section, we provide some useful preliminary results, which are need to prove the final regret and violation bounds.

Lemma D.2. *The loss given to the primal algorithm at episode $t \in [T]$, which is defined as $\ell_t(x,a) = \Gamma_t + \sum_{i\in[m]} \lambda_{t,i} g_{t,i}(x,a) - r_t(x,a)$, is such that, considering the Lagrangian loss function $\ell_t^{\mathcal{L}}(x,a) = \sum_{i\in[m]} \lambda_{t,i} g_{t,i}(x,a) - r_t(x,a)$, it holds:*

$$
\ell_t^\top (q_t - q^*) = \ell_t^{\mathcal{L},\top} (q_t - q^*),
$$

and additionally, ℓ_t assume values in the bounded interval $[0, \Xi_t]$.

Proof. By simple computation, it holds:

$$
\begin{aligned}
\ell_t^\top (q_t - q^*) - \ell_t^{\mathcal{L},\top} (q_t - q^*)
&= \Bigg(\sum_{x,a} \Gamma_t (q_t(x,a) - q^*(x,a)) \\
&\quad + \sum_{i\in[m]} \sum_{x,a} \lambda_{t,i} g_{t,i}(x,a)(q_t(x,a) - q^*(x,a)) \\
&\quad - \sum_{x,a} r_t(x,a)(q_t(x,a) - q^*(x,a)) \Bigg) \\
&\quad - \Bigg(\sum_{i\in[m]} \sum_{x,a} \lambda_{t,i} g_{t,i}(x,a)(q_t(x,a) - q^*(x,a)) \\
&\quad - \sum_{x,a} r_t(x,a)(q_t(x,a) - q^*(x,a)) \Bigg) \\
&= \sum_{x,a} \Gamma_t (q_t(x,a) - q^*(x,a)) \\
&= \Gamma_t (L - L) \\
&= 0,
\end{aligned}
$$

where the last steps hold since Γ_t is a constant and by the definition of *valid* occupancy measures.

In addition it holds:

$$\ell_t(x,a) = \Gamma_t + \sum_{i\in[m]} \lambda_{t,i} g_{t,i}(x,a) - r_t(x,a) \geq 1 + \|\lambda_t\|_1 - \sum_{i\in[m]} \lambda_{t,i} - 1 = 0,$$

and similarly,

$$\ell_t(x,a) = \Gamma_t + \sum_{i\in[m]} \lambda_{t,i} g_{t,i}(x,a) - r_t(x,a) \leq \Gamma_t + \sum_{i\in[m]} \lambda_{t,i} = 1 + 2\|\lambda_t\| \leq \Xi_t.$$

This concludes the proof. □

We proceed with some useful lemmas which holds under Condition 6.1.

Lemma D.3. *If Condition 6.1 holds, for all $t \in [T]$ and for each constraints $i \in [m]$, it holds:*

$$\lambda_{t,i} \geq \eta \widehat{V}_{t-1,i},$$

where $\widehat{V}_{t,i} := \sum_{\tau=1}^{t} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a)$.

Proof. First observe that with $t = 1$ we have that $\widehat{V}_{t-1,i}$ is the sum of zero elements and as such, it is equal to zero. This means that for $t = 1$ the inequality $\lambda_{t,i} \geq \eta \widehat{V}_{t-1,i}$ is equivalent to

$$\lambda_{t,i} \geq 0,$$

which is true by construction. We finish the proof by induction. Suppose $\lambda_{t,i} \geq \eta \widehat{V}_{t-1,i}$ is true for a $t \in [T]$, we show that it also holds for $t+1$, indeed:

$$\begin{aligned}
\lambda_{t+1,i} &= \max\left\{\lambda_{t,i} + \eta \sum_{k=0}^{L-1} g_{t,i}(x_k,a_k), 0\right\} \\
&= \max\left\{\lambda_{t,i} + \eta \sum_{x,a} g_{t,i}(x,a)\mathbb{I}_t(x,a), 0\right\} \\
&\geq \lambda_{t,i} + \eta \sum_{x,a} g_{t,i}(x,a)\mathbb{I}_t(x,a) \\
&\geq \eta \widehat{V}_{t-1,i} + \eta \sum_{x,a} g_{t,i}(x,a)\mathbb{I}_t(x,a) \\
&= \eta \left(\sum_{\tau=1}^{t-1} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) + g_{t,i}(x,a)\mathbb{I}_t(x,a)\right) \\
&= \eta \sum_{\tau=1}^{t} g_{\tau,i}(x,a)\mathbb{I}_t(x,a) \\
&= \eta \widehat{V}_{t,i}.
\end{aligned}$$

This concludes the proof. □

Lemma D.4. *If Condition 6.1 holds, referring as i^* to the element in $[m]$ such that $i^* =$*

$\arg\max_{i\in[m]} \sum_{t=1}^{T} \left[G_t^\top q^{P,\pi_t}\right]_i$, *then with probability at least* $1-\delta$, *it holds:*

$$V_T \leq \widehat{V}_{T,i^*} + \mathcal{E}^{\mathbb{I}},$$

where $\widehat{V}_{t,i} = \sum_{\tau=1}^{t} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a)$.

Proof. We observe with probability at least $1-\delta$:

$$\begin{aligned}
V_T &= \sum_{t=1}^{T} \left[G_t^\top q^{P,\pi_t}\right]_{i^*} \\
&= \sum_{t=1}^{T} \sum_{x,a} g_{t,i^*}(x,a)\left(q_t(x,a) - \mathbb{I}_t(x,a)\right) + \sum_{t=1}^{T} \sum_{x,a} g_{t,i^*}(x,a)\mathbb{I}_t(x,a) \\
&\leq \sum_{t=1}^{T} \sum_{x,a} g_{t,i^*}(x,a)\left(q_t(x,a) - \mathbb{I}_t(x,a)\right) + \widehat{V}_{T,i^*} \\
&\leq \mathcal{E}^{\mathbb{I}} + V_{T,i^*}.
\end{aligned}$$

This concludes the proof. □

We conclude with a result in the case Condition 6.1 does not hold.

Lemma D.5. *When Condition 6.1 does not hold, with probability at least* $1-10\delta$ *in case of stochastic costs and* $1-9\delta$ *in case of adversarial costs it holds for all* $i \in [m]$*:*

$$\widehat{V}_{T,i} \leq \frac{4T^{\frac{1}{4}}}{\eta}.$$

Proof. Recall the definition of $\widehat{V}_{t,i}$ as $\widehat{V}_{t,i} = \sum_{\tau=1}^{t} \sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a)$. We first focus on the stochastic setting. Thus, with probability at least $1-\delta$, it holds:

$$\begin{aligned}
\sum_{t=1}^{T} r_t^\top q_t - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t &= \sum_{t=1}^{T} r_t^\top q^\diamond - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^\diamond + \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top}(q^\diamond - q_t) \\
&\geq -\sum_{t=1}^{T} \lambda_t^\top \overline{G}^\top q^\diamond - \lambda_{1,T}\mathcal{E}_T^G + \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top}(q^\diamond - q_t) \\
&\geq -mT^{\frac{1}{4}}\mathcal{E}_T^G + \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top}(q^\diamond - q_t).
\end{aligned}$$

On the other hand, in case of adversarial constraints, it holds:

$$\begin{aligned}
\sum_{t=1}^{T} r_t^\top q_t - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t &= \sum_{t=1}^{T} r_t^\top q^\diamond - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q^\diamond + \sum_{t=1}^{T} \ell_t^{\mathcal{L}\top}(q^\diamond - q_t) \\
&\geq \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top}(q^\diamond - q_t).
\end{aligned}$$

Define a vector $\widetilde{\lambda} \in [0, T^{\frac{1}{4}}]^m$ as $\widetilde{\lambda}_j = 0$ if $j \neq i$ and $\widetilde{\lambda}_j = T^{\frac{1}{4}}$ if $j = i$. Simultaneously

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

$$\begin{aligned}
&\sum_{t=1}^{T} r_t^\top q_t - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t \\
&\leq \sum_{t=1}^{T} r_t^\top q_t - \sum_{t=1}^{T} \widetilde{\lambda}^\top \sum_{x,a} G_t(x,a)\mathbb{I}_t(x,a) \\
&\qquad + \sum_{t=1}^{T} (\widetilde{\lambda} - \lambda_t)^\top \sum_{x,a} G_t(x,a)\mathbb{I}_t(x,a) + \lambda_{1,T}\mathcal{E}^{\mathbb{I}} \\
&\leq \sum_{t=1}^{T} r_t^\top q_t - \sum_{t=1}^{T} \widetilde{\lambda}^\top \sum_{x,a} G_t(x,a)\mathbb{I}_t(x,a) + \mathcal{E}_T^D(\widetilde{\lambda}) + mT^{\frac{1}{4}}\mathcal{E}^{\mathbb{I}} \\
&\leq LT - T^{\frac{1}{4}}\widehat{V}_{T,i} + \mathcal{E}_T^D(\widetilde{\lambda}) + mT^{\frac{1}{4}}\mathcal{E}^{\mathbb{I}},
\end{aligned}$$

where in the first equality we used the definition of $\mathcal{E}_T^{\mathbb{I}}$, in the first inequality we used the definition of the dual space $[0, T^{\frac{1}{4}}]^m$ to bound $\lambda_{1,T}$ as $mT^{\frac{1}{4}}$, and in the last inequality we used the definition of $\widetilde{\lambda}$. We can then compare the lower and the upper bound for $\sum_{t=1}^{T} r_t^\top q_t - \sum_{t=1}^{T} \lambda_t^\top G_t^\top q_t$ obtaining the following inequality, which holds with probability at least $1-\delta$ with adversarial constraints and with probability at least $1-2\delta$ with stochastic constraints:

$$-mT^{\frac{1}{4}}\mathcal{E}_T^G + \sum_{t=1}^{T} \ell_t^{\mathcal{L},\top}(q^\diamond - q_t) \leq LT - T^{\frac{1}{4}}\widehat{V}_{T,i} + \mathcal{E}_T^D(\widetilde{\lambda}) + mT^{\frac{1}{4}}\mathcal{E}^{\mathbb{I}},$$

from which we can write the following inequality that holds with probability at least $1-9\delta$ with adversarial constraints and $1-10\delta$ with stochastic constraints:

$$T^{\frac{1}{4}}\widehat{V}_{T,i} \leq mT^{\frac{1}{4}}(\mathcal{E}_T^G + \mathcal{E}_T^{\mathbb{I}}) + \mathcal{E}_T^P + \mathcal{E}_T^D(\widetilde{\lambda}) + LT. \tag{D.5}$$

We proceed now to bound each element of the right side of the inequality.
To bound $\mathcal{E}^P$ we use the fact that $\Xi_{1,T} \leq (1+\lambda_{1,T}) \leq (1+mT^{\frac{1}{4}})$ and the definition of η as following:

$$\begin{aligned}
\mathcal{E}_T^P &= \Xi_{1,T}\left(U_1 C\sqrt{T} + U_2\frac{\sqrt{T}}{C} + U_3\frac{1}{C\sqrt{T}}\right) + U_4\sqrt{T} \\
&\leq 2(1+mT^{\frac{1}{4}})\sqrt{T}\left(1512L^3|X||A|\ln\left(\frac{L|A|T^2}{\delta}\right) + \frac{9L|X||A|}{252L|X||A|} + \frac{\frac{L}{2}\ln\left(\frac{LT^2}{\delta}\right)}{252L|X||A|}\right) \\
&\quad + \sqrt{T}30L^2|X|^2\sqrt{2|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)} \\
&\leq T^{\frac{1}{4}}\sqrt{T}6056L^3 m|X||A|\ln\left(\frac{L|A|T^2}{\delta}\right) + \sqrt{T}30L^2|X|^2\sqrt{2|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}
\end{aligned}$$

$$\leq T^{\frac{1}{4}}\sqrt{T}6116L^2m|X|^2|A|\ln\left(\frac{|X|^2|A|T^2}{\delta}\right)$$
$$\leq \frac{T^{\frac{1}{4}}}{\eta}.$$

To bound $\mathcal{E}_T^D(\widetilde{\lambda})$ we use Theorem D.2, the fact that by its definition $\|\widetilde{\lambda}\|_2^2 = \left(T^{\frac{1}{4}}\right)^2 = \sqrt{T}$, the initialization of the dual $\lambda_1 = \underline{0}$ and the definition of η in the following way:

$$\begin{aligned}
\mathcal{E}_T^D(\widetilde{\lambda}) &\leq \frac{\|\lambda_1 - \widetilde{\lambda}\|_2^2}{2\eta} + \frac{\eta}{2}TmL^2 \\
&= \frac{\|\widetilde{\lambda}\|_2^2}{2\eta} + \frac{\eta}{2}TmL^2 \\
&= \frac{\sqrt{T}}{2\eta} + \frac{\eta}{2}TmL^2 \\
&\leq \frac{\sqrt{T}}{2\eta} + \frac{mL^2T}{2}\frac{1}{84672mL^2|X|^2|A|\ln\left(\frac{|A||X|^2T^2}{\delta}\right)\sqrt{T}} \\
&\leq \frac{\sqrt{T}}{2\eta} + \frac{\sqrt{T}}{2\eta} = \frac{\sqrt{T}}{\eta}
\end{aligned}$$

We proceed to simply bound also $mT^{\frac{1}{4}}\left(\mathcal{E}_T^G + \mathcal{E}_T^{\mathbb{I}}\right)$ through their definition:

$$\begin{aligned}
mT^{\frac{1}{4}}\left(\mathcal{E}_T^G + \mathcal{E}_T^{\mathbb{I}}\right) &= mT^{\frac{1}{4}}\sqrt{T}\left(2L\sqrt{\ln\left(\frac{T^2}{\delta}\right)} + L\sqrt{2\ln\left(\frac{T^2}{\delta}\right)}\right) \\
&\leq \frac{T^{\frac{1}{4}}}{\eta}
\end{aligned}$$

Finally we bound LT as $LT \leq \frac{\sqrt{T}}{\eta}$.
Thus, Inequality (D.5) becomes

$$\widehat{V}_{T,i} \leq \frac{1}{T^{\frac{1}{4}}}\left(mT^{\frac{1}{4}}(\mathcal{E}_T^G + \mathcal{E}_T^{\mathbb{I}}) + \mathcal{E}_T^P + \mathcal{E}_T^D(\widetilde{\lambda}) + LT\right) \leq \frac{4\sqrt{T}}{T^{\frac{1}{4}}\eta} = \frac{4T^{\frac{1}{4}}}{\eta},$$

which concludes the proof. □

D.6 Auxiliary Lemmas

In this section, we provide some technical lemmas which are adapted from existing works.

Lemma D.6 (Adapted from (Luo et al., 2021) Lemma C.4)**.**

$$\eta_t\widehat{Q}_t(x,a) \leq \frac{1}{2} \quad \wedge \quad \eta_t B_t(x,a) \leq \frac{1}{2L}.$$

Proof. Recall $\gamma = 2\eta_t L \Xi_t$. Thus, it holds:

$$\eta_t \widehat{Q}_t(x,a) \leq \frac{\eta_t L \Xi_t}{\gamma} = \frac{\eta_t L \Xi_t}{2\eta_t L \Xi_t} = \frac{1}{2}$$

and

$$\eta_t b_t(x,a) = \frac{3\eta_t L \Xi_t \gamma + \eta_t \Xi_t L(\overline{q}_t(x,a) - \underline{q}_t(x,a))}{\overline{q}_t(x,a) + \gamma} \leq 3\eta_t \Xi_t L + \eta_t L \Xi_t = 2\gamma.$$

Finally,

$$\begin{aligned}
\eta_t B_t(x,a) &\leq L\left(1 + \frac{1}{L}\right)^L \eta_t \sup_{x',a'} b_t(x',a') \\
&\leq 3L2\gamma \\
&= 6L\gamma \\
&= \frac{6L}{C\sqrt{T}} \\
&= \frac{6L}{252|X||A|L\sqrt{T}} \\
&\leq \frac{1}{42L} \\
&\leq \frac{1}{2L}.
\end{aligned}$$

This concludes the poof. □

Lemma D.7 (Adapted from (Luo et al., 2021), Lemma B.1)**.** *If the following inequality holds:*

$$\begin{aligned}
&\sum_x q^*(x) \sum_{t=t_1}^{t_2} \sum_a \left(\pi_t(a|x) - \pi^*(a|x)\right)\left(Q_t^{\pi_t}(x,a) - B_t(x,a)\right) \\
&\qquad \leq o(T) + \sum_{t=t_1}^{t_2} V^{\pi^*}(b_t) + \frac{1}{L}\sum_{t=t_1}^{t_2}\sum_{x,a} q^*(x)\pi_t(a|x)B_t(x,a),
\end{aligned} \tag{D.6}$$

where B_t is defined as:

$$B_t(x,a) = b_t(x,a) + \left(1 + \frac{1}{L}\right)\mathbb{E}_{x'\sim P(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right], \tag{D.7}$$

for all $t \in [T], x \in X, a \in A$, then it holds that:

$$R_{t_1,t_2} \leq o(T) + 3\sum_{t=t_1}^{t_2} \widehat{V}^{\pi_t}(x_0; b_t),$$

where $\widehat{V}^{\pi_t}(x_0;b_t) = \sum_{x,a} q^{\widehat{P}_t,\pi_t}(x,a)\left(\frac{L\Xi_t(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + 3L\Xi_t\gamma}{\overline{q}_t(x,a)+\gamma}\right)$.

Proof. The proof is analogous to the one proposed by (Luo et al., 2021), Lemma B.1, since the proof is episode based and then the sum over t is taken. □

Lemma D.8 (Adapted from (Luo et al., 2021), Lemma A.1)**.** *Let $\mathcal{F}_0, \dots, \mathcal{F}_T$ be a filtration and $X_1, \dots, X_T$ be real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_t] = 0$, $|X_t| \leq b$ for all $t \in [T]$ and $\sum_{t=t_1}^{t_2} \mathbb{E}[X_t^2|\mathcal{F}_t] \leq V_{t_1,t_2}$ for some fixed $V_{t_1,t_2} > 0$ and $b > 0$ for every $t_1, t_2 \in [T]$ such that $1 \leq t_1 \leq t_2 \leq T$. Then with probability at least $1 - \delta$ it holds simultaneously for all $[t_1, \dots, t_2] \subset [T]$:*

$$\sum_{t=t_1}^{t_2} X_t \leq \frac{V_{t_1,t_2}}{b} + b \ln\left(\frac{T^2}{\delta}\right).$$

Proof. For all $\delta' \in (0, 1)$ by Lemma A.1 (Luo et al., 2021) it holds:

$$\mathbb{P}\left(\sum_{t=t_1}^{t_2} X_t \geq \frac{V_{t_1,t_2}}{b} + b \ln\left(\frac{1}{\delta'}\right)\right) \leq \delta'.$$

It is sufficient to consider the intersection of all events for all possible intervals $[t_1, \dots, t_2]$, that are less than T^2.

$$\mathbb{P}\left(\bigcap_{t_1,t_2}\left\{\sum_{t=t_1}^{t_2} X_t \geq \frac{V_{t_1,t_2}}{b} + b \ln\left(\frac{1}{\delta'}\right)\right\}\right) \leq T^2\delta'.$$

To conclude the proof we take δ as $T^2\delta'$. □

Consider a loss function $f_t(x, a) \in [0, Z]$, for all $t \in [T], (x, a) \in X \times A$, with $Z > 0$. Define another function $\widetilde{f}_t \in [0, Z]^{|X \times A|}$. If we define the estimator $\widehat{f}_t(x, a) = \frac{\widetilde{f}_t(x,a)\mathbb{I}_t(x,a)}{\overline{q}_t(x,a)+\gamma}$ where $\mathbb{E}[\widetilde{f}_t(x, a)] = f_t(x, a)$, we can state the following result.

Lemma D.9 (Adapted from (Jin et al., 2020a))**.** *For every sequence of functions $\alpha_1, \dots \alpha_T$ such that $\alpha_t \in [0, \frac{2\gamma}{Z}]^{|X \times A|}$ is $\mathcal{F}_t$ measurable for all $t \in [T]$, we have with probability at least $1 - \delta$ that simultaneously for all $t_1, t_2 \in [T]$ such that $1 \leq t_1 \leq t_2 \leq T$ it holds:*

$$\sum_{t=t_1}^{t_2} \sum_{x,a} \alpha_t(x, a)\left(\widehat{f}_t(x, a) - \frac{q_t(x, a)}{\overline{q}_t(x, a)} f_t(x, a)\right) \leq L \ln\left(\frac{LT^2}{\delta}\right).$$

Proof. It holds:

$$\begin{aligned}
\widehat{\ell}_t(x, a) &= \frac{\widetilde{f}_t(x, a)\mathbb{I}_t(x, a)}{\overline{q}_t(x, a) + \gamma} \\
&\leq \frac{\widetilde{f}_t(x, a)\mathbb{I}_t(x, a)}{\overline{q}_t(x, a) + \frac{\widetilde{f}_t(x,a)}{Z}\gamma} \\
&= \frac{\mathbb{I}_t(x, a)Z}{2\gamma} \frac{2\gamma\frac{\widetilde{f}_t(x,a)}{Z}}{\overline{q}_t(x, a) + \gamma\frac{\widetilde{f}_t(x,a)}{Z}}
\end{aligned}$$

$$= \frac{\mathbb{I}_t(x,a)Z}{2\gamma} \frac{2\gamma \frac{\widetilde{f}_t(x,a)}{Z\overline{q}_t(x,a)}}{1+\gamma \frac{\widetilde{f}_t(x,a)}{Z\overline{q}_t(x,a)}}$$
$$\leq \frac{Z}{2\gamma} \ln\left(1 + 2\gamma \frac{\mathbb{I}_t(x,a)\widetilde{f}_t(x,a)}{Z\overline{q}_t(x,a)}\right).$$

For each layer $k \in [0,\dots,L-1]$ we define $\widehat{S}_{t,k} := \sum_{x\in X_k, a\in A} \alpha_t(x,a)\widehat{f}_t(x,a)$ and $S_{t,k} := \sum_{x\in X_k, a\in A} \alpha_t(x,a)\frac{q_t(x,a)}{\overline{q}_t(x,a)} f_t(x,a)$. Thus, it holds:

$$\begin{aligned}
\mathbb{E}_t[\exp(\widehat{S}_{t,k})] &= \mathbb{E}\left[\exp\left(\sum_{x\in X_k, a\in A} \alpha_t(x,a)\widehat{f}_t(x,a)\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\sum_{x\in X_k, a\in A} \alpha_t(x,a)\frac{Z}{2\gamma}\ln\left(1+2\gamma\frac{\mathbb{I}_t(x,a)\widetilde{f}_t(x,a)}{Z\overline{q}_t(x,a)}\right)\right)\right] \\
&\leq \mathbb{E}\left[\prod_{x\in X_k, a\in A}\left(1+\alpha_t(x,a)\frac{\mathbb{I}_t(x,a)\widetilde{f}_t(x,a)}{\overline{q}_t(x,a)}\right)\right] \\
&\leq 1 + \sum_{x\in X_h, a\in A} \alpha_t(x,a)\frac{q_t(x,a)f_t(x,a)}{\overline{q}_t(x,a)} \\
&= 1 + S_{t,k} \\
&\leq \exp(S_{t,k}).
\end{aligned}$$

For each interval $[t_1,\dots,t_2] \subset [T]$ it holds:

$$\mathbb{P}\left[\sum_{t=t_1}^{t_2}(\widehat{S}_{t,k} - S_{t,k}) \geq \ln\left(\frac{L}{\delta'}\right)\right] \leq \frac{\delta'}{L}.$$

Taking the intersection event for all intervals $[t_1,\dots,t_2] \subset [T]$:

$$\mathbb{P}\left[\bigcap_{t_1,t_2}\left\{\sum_{t=t_1}^{t_2}(\widehat{S}_{t,k} - S_{t,k}) \geq \ln\left(\frac{L}{\delta'}\right)\right\}\right] \leq T^2\frac{\delta'}{L}.$$

$$\delta = T^2\delta',$$

and

$$\mathbb{P}\left[\bigcap_{t_1,t_2}\left\{\sum_{t=t_1}^{t_2}(\widehat{S}_{t,k} - S_{t,k}) \geq \ln\left(\frac{LT^2}{\delta}\right)\right\}\right] \leq \frac{\delta}{L}.$$

Finally we take the sum over $k \in [0,\dots,L-1]$:

$$\mathbb{P}\left[\sum_{t=t_1}^{t_2}\sum_{x,a}\alpha_t(x,a)\left(\widehat{f}_t(x,a) - \frac{q_t(x,a)}{u_t(x,a)}f_t(x,a)\right) \leq L\ln\left(\frac{LT^2}{\delta}\right)\right] \leq \delta.$$

This concludes the proof. □

Corollary D.1. *Given $\delta \in (0,1)$, it holds with probability at least $1-2\delta$ simultaneously for all $t_1, t_2 \in [T]$ such that $1 \leq t_1 \leq t_2 \leq T$:*

$$\sum_{t=t_1}^{t_2} \sum_{x,a} \left(\widehat{f}_t(x,a) - f_t(x,a) \right) \leq \frac{ZL}{2\gamma} \ln \left(\frac{LT^2}{\delta} \right).$$

We conclude with the following concentration result on the transitions.

Lemma D.10. *Let $\{\pi_t\}_{t=1}^T$ policies, then for any collection of transition $P_t^x \in \mathcal{P}_t$ with probability at least $1-2\delta$,*

$$\sum_{t=1}^{T} \|q^{P,\pi_t} - q^{P_t^x,\pi_t}\|_1 \leq 2L|X|^2 \sqrt{2T \ln \left(\frac{L|X|}{\delta} \right)} + 3L|X|^2 \sqrt{2T|A| \ln \left(\frac{T|X|^2|A|}{\delta} \right)}.$$

Proof. It holds:

$$\begin{aligned}
&\sum_{t=1}^{T} \|q^{P,\pi_t} - q^{P_t^x,\pi_t}\|_1 \\
&\quad = \sum_{t=1}^{T} \sum_{x,a} |q^{P,\pi_t}(x,a) - q^{P_t^x,\pi_t}(x,a)| \\
&\quad \leq \sum_{t=1}^{T} \sum_{x,a} \sum_{x'} |q^{P,\pi_t}(x',a) - q^{P_t^x,\pi_t}(x',a)| \\
&\quad = \sum_{x} \sum_{t=1}^{T} \sum_{x',a} |q^{P,\pi_t}(x',a) - q^{P_t^x,\pi_t}(x',a)| \\
&\quad \leq \sum_{x} \left(2L|X| \sqrt{2T \ln \left(\frac{L|X|}{\delta} \right)} + 3L|X| \sqrt{2T|A| \ln \left(\frac{T|X|^2|A|}{\delta} \right)} \right) \\
&\quad \leq |X| \left(2L|X| \sqrt{2T \ln \left(\frac{|X|L}{\delta} \right)} + 3L|X| \sqrt{2T|A| \ln \left(\frac{T|X|^2|A|}{\delta} \right)} \right),
\end{aligned}$$

by Lemma 5.1, taking the union bound over X $\left(\delta' = \frac{\delta}{|X|}\right)$. This concludes the proof. □

APPENDIX E

Omitted Lemmas and Proofs of Chapter 7

E.1 Results on the Optimization Update

In this section, we provide the results associated with the optimization update performed by Algorithm 7.1. We start with the following lemma.

Lemma E.1. *For any $\delta \in (0,1)$ and for any $q \in \bigcap_{t\in[T]} \widehat{\Delta}_t(\mathcal{P}_t)$, Algorithm 7.1 attains:*

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

with probability at least $1-\delta$.

Proof. The result follows from Lemma 12 of (Jin et al., 2020a), considering a general $q \in \bigcap_{t\in[T]} \widehat{\Delta}_t(\mathcal{P}_t)$. □

We conclude by showing the following performance bound.

Theorem E.1. *For any $\delta \in (0,1)$ and for any $q \in \bigcap_{t\in[T]} \widehat{\Delta}_t(\mathcal{P}_t)$, Algorithm 7.1, with $\eta = \gamma = \sqrt{\frac{L\ln\left(\frac{L|X||A|}{\delta}\right)}{T|X||A|}}$, attains:*

$$\sum_{t=1}^{T} r_t^\top (q - q_t) \leq 14L|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)},$$

with probability at least $1-15\delta$.

Proof. It holds:

$$\sum_{t=1}^{T} \ell_t^\top (q_t - q) = \sum_{t=1}^{T} (\ell_t - \widehat{\ell}_t)^\top \widehat{q}_t + \sum_{t=1}^{T} (\widehat{\ell}_t - \ell_t)^\top q \tag{E.1}$$

$$+ \sum_{t=1}^{T} \widehat{\ell}_t^\top (\widehat{q}_t - q) + \sum_{t=1}^{T} \ell_t^\top (q_t - \widehat{q}_t)$$

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

$$+ 3L|X|^2 \sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)} + \sum_{t=1}^{T} (\widehat{\ell}_t - \ell_t)^\top q$$

$$+ \sum_{t=1}^{T} \widehat{\ell}_t^\top (\widehat{q}_t - q) + \sum_{t=1}^{T} \ell_t^\top (q_t - \widehat{q}_t) \tag{E.2}$$

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

$$+ 3L|X|^2 \sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)} + \frac{L \ln(|X||A|/\delta)}{\gamma}$$

$$+ \sum_{t=1}^{T} \widehat{\ell}_t^\top (\widehat{q}_t - q) + \sum_{t=1}^{T} \ell_t^\top (q_t - \widehat{q}_t) \tag{E.3}$$

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

$$+ 3L|X|^2 \sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)} + \frac{L \ln(|X||A|/\delta)}{\gamma}$$

$$+ L\frac{\ln(|X|^2|A|)}{\eta} + \eta |X||A|T + \frac{\eta L \ln \frac{L}{\delta}}{\gamma} + \sum_{t=1}^{T} \ell_t^\top (q_t - \widehat{q}_t) \tag{E.4}$$

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

$$+ 3L|X|^2 \sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)} + \frac{L \ln(|X||A|/\delta)}{\gamma}$$

$$+ L\frac{\ln(|X|^2|A|)}{\eta} + \eta |X||A|T + \frac{\eta L \ln \frac{L}{\delta}}{\gamma}$$

$$+ 2L|X| \sqrt{2T \ln \frac{2L}{\delta}} + 3L|X| \sqrt{2T|A| \ln \frac{2T|X||A|}{\delta}} \tag{E.5}$$

$$
\begin{aligned}
&\leq \sqrt{|X||A|TL \ln\left(\frac{L|X||A|}{\delta}\right)} + 2L|X|^2\sqrt{2T \ln\left(\frac{L|X|}{\delta}\right)} \\
&+ 3L|X|^2\sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)} + 3\sqrt{|X||A|TL \ln\left(\frac{L|X||A|}{\delta}\right)} \\
&+ 2L|X|\sqrt{2T \ln\frac{2L}{\delta}} + 3L|X|\sqrt{2T|A| \ln\frac{2T|X||A|}{\delta}} \\
&\leq (4+2+3+2+3)L|X|^2\sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)} \\
&= 14L|X|^2\sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)},
\end{aligned}
$$

where Inequality (E.2) holds by Lemma E.8 with probability $1-7\delta$, Inequality (E.3) holds by Lemma 14 of (Jin et al., 2020a) with probability $1-5\delta$, Inequality (E.4) holds by Lemma E.1 with probability $1-\delta$ and Inequality (E.5) holds by Lemma B.3 of (Rosenberg and Mansour, 2019b) with probability $1-2\delta$. By Union Bound, the final result holds with probability $1-15\delta$. Since by definition $\ell_t(x,a) = 1 - r_t(x,a)\mathbb{I}_t(x,a)$ for all $x \in X, a \in A$, it holds:

$$
\sum_{t=1}^{T} \ell_t^\top (q_t - q) = \sum_{t=1}^{T} r_t^\top (q - q_t) \leq 14L|X|^2\sqrt{2T|A| \ln\left(\frac{T|X|^2|A|}{\delta}\right)},
$$

which concludes the proof. □

E.2 Results on the Decision Space

In this section, we provide the results on the decision space definition of Algorithm 7.1.

We start by showing that, in the stochastic setting, the confidence bound plays a central role in the definition of the decision space.

Theorem E.2. *In the stochastic setting, let $\delta \in (0,1)$ and $b_t(x,a)$ such that with probability at least $1-\delta$, it holds $|\widehat{g}_{t,i}(x,a) - \bar{g}_i(x,a)| \leq b_t(x,a)$, for all $(x,a) \in X \times A, i \in [m], t \in [T]$. Furthermore, let $\Delta^\star = \left\{q \in \Delta(M) : \bar{g}_i^\top q \leq 0, \forall i \in [m]\right\}$. Then, with probability at least $1-2\delta$ it holds:*

$$
\Delta^\star \subseteq \widehat{\Delta}_t(\mathcal{P}_t),\ \forall t \in [T].
$$

Proof. Assume the condition of the theorem holds. Let $q \in \Delta^\star$ and consider the following inequalities:

$$
\begin{aligned}
\widehat{g}_{t,i}^\top q &= (\widehat{g}_{t,i} - \bar{g}_i)^\top q + \bar{g}_i^\top q \\
&\leq (\widehat{g}_{t,i} - \bar{g}_i)^\top q \\
&= \sum_{x \in X, a \in A} (\widehat{g}_{t,i}(x,a) - \bar{g}_i(x,a))q(x,a) \\
&\leq b_t^\top q,
\end{aligned}
$$

where the first inequality holds by definition of $\Delta^\star$ and the second inequality follows from the definition of b_t. Thus, $(\widehat{g}_{t,i} - b_t)^\top q \leq 0$, which by definition proves that $q \in \widehat{\Delta}_t(\mathcal{P}_t)$. The final results follow from noticing that by Lemma 4.1 of (Rosenberg and Mansour, 2019b) $P \in \mathcal{P}_t$ with probability at least $1-\delta$. A final Union Bound concludes the proof. □

E.3 Results on the Weights

In this section, we provide some fundamental results on the weights employed by Algorithm 7.1. Specifically, we relate the violation attained by Algorithm 7.1 to the weighted estimators.

Theorem E.3. *Given an interval $[t_1, t_2] \subseteq [T]$, $i \in [m]$ and $\delta \in (0,1)$, Algorithm 7.1 attains the following bound with probability at least $1-3\delta$:*

$$V_{[t_1,t_2],i} \leq \sum_{x\in X, a\in A} \sum_{\tau \in \mathcal{T}_{t_2,x,a} \cap [t_1,t_2]} \frac{1}{\beta_{\tau,i}(x,a)} \left(\widehat{g}_{\tau,i}(x,a) - \widehat{g}_{\tau-1,i}(x,a)\right) + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^\top \widehat{q}_\tau + 7L|X|\sqrt{2(t_2-t_1)|A| \ln \frac{2T|X||A|}{\delta}},$$

where $V_{[t_1,t_2],i} \coloneqq \sum_{\tau=t_1}^{t_2} g_{\tau,i}^\top q_\tau$.

Proof. It holds:

$$\begin{aligned}
V_{[t_1,t_2],i} &= \sum_{\tau=t_1}^{t_2} g_{\tau,i}^\top q_\tau \\
&\leq \sum_{\tau=t_1}^{t_2} g_{\tau,i}^\top q_\tau + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^\top \widehat{q}_\tau - \sum_{\tau=t_1}^{t_2} \widehat{g}_{\tau-1,i}^\top \widehat{q}_\tau && \text{(E.6)} \\
&= \sum_{\tau=t_1}^{t_2} g_{\tau,i}^\top q_\tau + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^\top \widehat{q}_\tau - \sum_{\tau=t_1}^{t_2} \widehat{g}_{\tau-1,i}^\top \widehat{q}_\tau + \sum_{\tau=t_1}^{t_2} \widehat{g}_{\tau-1,i}^\top q_\tau - \sum_{\tau=t_1}^{t_2} \widehat{g}_{\tau-1,i}^\top q_\tau \\
&= \sum_{\tau=t_1}^{t_2} (g_{\tau,i} - \widehat{g}_{\tau-1,i})^\top q_\tau + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^\top \widehat{q}_\tau + \sum_{\tau=t_1}^{t_2} \widehat{g}_{\tau-1,i}^\top (q_\tau - \widehat{q}_\tau) \\
&\leq \sum_{\tau=t_1}^{t_2} \sum_{x\in X, a\in A} (g_{\tau,i}(x,a) - \widehat{g}_{\tau-1,i}(x,a)) \mathbb{I}_\tau(x,a) + 2L\sqrt{2(t_2-t_1)\ln\frac{1}{\delta}} \\
&\quad + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^\top \widehat{q}_\tau + \|q_\tau - \widehat{q}_\tau\|_1 && \text{(E.7)} \\
&\leq \sum_{\tau=t_1}^{t_2} \sum_{x\in X, a\in A} (g_{\tau,i}(x,a) - \widehat{g}_{\tau-1,i}(x,a)) \mathbb{I}_\tau(x,a) \\
&\quad + 2L\sqrt{2(t_2-t_1)\ln\frac{1}{\delta}} + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^\top \widehat{q}_\tau
\end{aligned}$$

$$
\begin{aligned}
&+ 2L|X|\sqrt{2(t_2 - t_1)\ln\frac{2L}{\delta}} \\
&+ 3L|X|\sqrt{2(t_2 - t_1)|A|\ln\frac{2T|X||A|}{\delta}} \qquad \text{(E.8)} \\
&= \sum_{x\in X, a\in A}\sum_{\tau\in\mathcal{T}_{t_2,x,a}\cap[t_1,t_2]}\frac{(\widehat{g}_{\tau,i}(x,a) - \widehat{g}_{\tau-1,i}(x,a))}{\beta_{\tau,i}(x,a)} \\
&+ \sum_{\tau=t_1}^{t_2} b_{\tau-1}^{\top}\widehat{q}_{\tau} + 2L|X|\sqrt{2(t_2 - t_1)\ln\frac{2L}{\delta}} \\
&+ 3L|X|\sqrt{2(t_2 - t_1)|A|\ln\frac{2T|X||A|}{\delta}} + 2L\sqrt{2(t_2 - t_1)\ln\frac{1}{\delta}} \qquad \text{(E.9)} \\
&\leq \sum_{x\in X, a\in A}\sum_{\tau\in\mathcal{T}_{t_2,x,a}\cap[t_1,t_2]}\frac{(\widehat{g}_{\tau,i}(x,a) - \widehat{g}_{\tau-1,i}(x,a))}{\beta_{\tau,i}(x,a)} \\
&+ \sum_{\tau=t_1}^{t_2} b_{\tau-1}^{\top}\widehat{q}_{\tau} + 7L|X|\sqrt{2(t_2 - t_1)|A|\ln\frac{2T|X||A|}{\delta}},
\end{aligned}
$$

where Equation (E.6) is due to the fact that $\widehat{q}_{\tau+1} \in \widehat{\Delta}_{\tau}(\mathcal{P}_{\tau})$, Inequality (E.7) holds by Lemma E.6, from which we have, with probability $1-\delta$:

$$
\begin{aligned}
&\sum_{\tau=t_1}^{t_2}(g_{\tau,i} - \widehat{g}_{\tau-1,i})^{\top} q_{\tau} \\
&\qquad \leq \sum_{\tau=t_1}^{t_2}\sum_{x\in X, a\in A}(g_{\tau,i}(x,a) - \widehat{g}_{\tau-1,i}(x,a))\mathbb{I}_{\tau}(x,a) + L\sqrt{8(t_2 - t_1)\ln\frac{1}{\delta}}.
\end{aligned}
$$

Inequality (E.8) follows from Lemma B.3 of (Rosenberg and Mansour, 2019b), with probability at least $1-2\delta$—notice that all the results mentioned above can be trivially extended to hold in the interval $[t_1, t_2]$—. Equation (E.9) holds by the definition of the update:

$$
\widehat{g}_{\tau,i}(x,a) = (1 - \beta_{\tau,i}(x,a))\,\widehat{g}_{\tau-1,i}(x,a) + \beta_{\tau,i}(x,a)g_{\tau,i}(x,a),
$$

for all (x,a) such that $\mathbb{I}_{\tau}(x,a) = 1$. A final Union Bound concludes the proof. □

We proceed with the following corollary.

Corollary E.1. *Given an interval $[t_1, t_2] \subseteq [T]$, $i \in [m]$ and $\delta > 0$, assume that for any $(x,a) \in X \times A$ it holds $\beta_{\tau,i}(x,a) \geq \beta_{\tau',i}(x,a)$ for each $\tau' \leq \tau \in \mathcal{T}_{t_2,x,a} \cap [t_1, t_2]$. Then, with probability at least $1-3\delta$ it holds:*

$$
V_{[t_1,t_2],i} \leq \sum_{x\in X, a\in A}\frac{2}{\beta_{\ell(x,a,[t_1,t_2]),i}(x,a)} + \sum_{\tau=t_1}^{t_2} b_{\tau-1}^{\top}\widehat{q}_{\tau} + 7L|X|\sqrt{2(t_2 - t_1)|A|\ln\frac{2T|X||A|}{\delta}},
$$

where $\ell(x,a,[t_1,t_2])$ are the last rounds in the interval $[t_1,t_2]$ in which the pair (x,a) is visited.

Proof. Assuming Theorem E.3 holds with probability $1-3\delta$, it is sufficient to show $\sum_{x\in X,a\in A}\sum_{\tau\in\mathcal{T}_{t_2,x,a}\cap[t_1,t_2]}\frac{1}{\beta_{\tau,i}(x,a)}\left(\widehat{g}_{\tau,i}(x,a)-\widehat{g}_{\tau-1,i}(x,a)\right) \leq \sum_{x\in X,a\in A}\frac{2}{\beta_{t_2,i}(x,a)}$. Fixing a $(x,a)\in X\times A$ and defining $h=|\mathcal{T}_{t_2,x,a}\cap[t_1,t_2]|$ as the number of times the pair (x,a) is visited in the interval $[t_1,t_2]$,let $\tau(j)$ be the rounds in which the pair (x,a) is visited the j^{th} time in $[t_1,t_2]$.
Then we have:

$$
\begin{aligned}
&\sum_{\tau\in\mathcal{T}_{t_2,x,a}\cap[t_1,t_2]}\frac{1}{\beta_{\tau,i}(x,a)}\left(\widehat{g}_{\tau,i}(x,a)-\widehat{g}_{\tau-1,i}(x,a)\right)\\
&\quad=\sum_{j\in[h]}\frac{1}{\beta_{\tau(j),i}(x,a)}\left(\widehat{g}_{\tau(j),i}(x,a)-\widehat{g}_{\tau(j-1),i}(x,a)\right)\\
&\quad=\sum_{j\in[h-1]}\left(\frac{1}{\beta_{\tau(j),i}(x,a)}\left(\widehat{g}_{\tau(j),i}(x,a)-\widehat{g}_{\tau(j-1),i}(x,a)\right)\right)\\
&\qquad+\frac{1}{\beta_{\tau(h),i}(x,a)}\left(\widehat{g}_{\tau(h),i}(x,a)-\widehat{g}_{\tau(h-1),i}(x,a)\right)\\
&\quad\leq\sum_{j\in[h-1]}\left(\frac{1}{\beta_{\tau(j+1),i}(x,a)}\widehat{g}_{\tau(j),i}(x,a)-\frac{1}{\beta_{\tau(j),i}(x,a)}\widehat{g}_{\tau(j-1),i}(x,a)\right)\\
&\qquad+\frac{1}{\beta_{\tau(h),i}(x,a)}\left(\widehat{g}_{\tau(h),i}(x,a)-\widehat{g}_{\tau(h-1),i}(x,a)\right) &\text{(E.10)}\\
&\quad=\frac{1}{\beta_{\tau(h),i}(x,a)}\widehat{g}_{\tau(h),i}(x,a)-\frac{1}{\beta_{\tau(1),i}(x,a)}\widehat{g}_{\tau(1)-1,i}(x,a) &\text{(E.11)}\\
&\quad\leq\frac{2}{\beta_{\tau(h),i}(x,a)} &\text{(E.12)}\\
&\quad=\frac{2}{\beta_{\ell(x,a,[t_1,t_2]),i}(x,a)},
\end{aligned}
$$

where Inequality (E.10) and Inequality (E.12) follow from the hypothesis that the learning rates are decreasing in the interval, and Equation (E.11) follows from evaluating the telescoping sum. □

E.4 Concentration Results

In this section, we provide a fundamental result on the concentration of the confidence bounds parameter employed by Algorithm 7.1. This is done in the following lemma.

Lemma E.2. *Given $c>0$, $\alpha\in(0,1)$, $t\in[T]$ and $\delta\in(0,1)$, let the bonus $b_t(x,a)=\max\left\{1,\frac{c}{N_t(x,a)^\alpha}\right\}$ for all $(x,a)\in X\times A$. Then, with probability $1-3\delta$ it holds:*

$$
\sum_{\tau=1}^{t}b_{\tau-1}^{\top}\widehat{q}_\tau\leq\frac{c}{1-\alpha}|X|^\alpha|A|^\alpha L^{1-\alpha}t^{1-\alpha}+7L|X|\sqrt{2t|A|\ln\frac{2T|X||A|}{\delta}}
$$

Proof. It holds:

$$\begin{aligned}
\sum_{\tau=1}^{t}\sum_{x\in X,a\in A} b_\tau(x,a)\mathbb{I}_\tau(x,a) &= c\sum_{\tau=1}^{t}\sum_{x\in X,a\in A}\frac{1}{N_\tau(x,a)^\alpha}\mathbb{I}_\tau(x,a)\\
&= c\sum_{x\in X,a\in A}\sum_{h=1}^{N_t(x,a)}\frac{1}{h^\alpha}\\
&\leq \frac{c}{1-\alpha}\sum_{x\in X,a\in A} N_t(x,a)^{1-\alpha}\\
&\leq \frac{c}{1-\alpha}|X|^\alpha|A|^\alpha L^{1-\alpha}t^{1-\alpha}, \qquad \text{(E.13)}
\end{aligned}$$

where Inequality (E.13) holds by Jensen's inequality. Moreover, notice that:

$$\begin{aligned}
\sum_{\tau=1}^{t} b_{\tau-1}^\top \widehat{q}_\tau &= \sum_{\tau=1}^{t} b_{\tau-1}^\top \widehat{q}_\tau + \sum_{\tau=1}^{t} b_{\tau-1}^\top q_\tau - \sum_{\tau=1}^{t} b_{\tau-1}^\top q_\tau\\
&\leq \|q_\tau - \widehat{q}_\tau\|_1 + \sum_{\tau=1}^{t} b_{\tau-1}^\top q_\tau\\
&\leq 2L|X|\sqrt{2T\ln\frac{2L}{\delta}} + 3L|X|\sqrt{2T|A|\ln\frac{2T|X||A|}{\delta}}\\
&\quad + \sum_{\tau=1}^{t} b_{\tau-1}^\top q_\tau \qquad \text{(E.14)}\\
&\leq 2L|X|\sqrt{2T\ln\frac{2L}{\delta}} + 3L|X|\sqrt{2T|A|\ln\frac{2T|X||A|}{\delta}}\\
&\quad + 2L\sqrt{2T\ln\frac{1}{\delta}} + \sum_{\tau=1}^{t}\sum_{x\in X,a\in A} b_{\tau-1}(x,a)\mathbb{I}_\tau(x,a) \qquad \text{(E.15)}\\
&\leq 2L|X|\sqrt{2T\ln\frac{2L}{\delta}} + 3L|X|\sqrt{2T|A|\ln\frac{2T|X||A|}{\delta}}\\
&\quad + 2L\sqrt{2T\ln\frac{1}{\delta}} + L + \sum_{\tau=1}^{t}\sum_{x\in X,a\in A} b_\tau(x,a)\mathbb{I}_\tau(x,a)\\
&\leq 2L|X|\sqrt{2T\ln\frac{2L}{\delta}} + 3L|X|\sqrt{2T|A|\ln\frac{2T|X||A|}{\delta}} + L\\
&\quad + 2L\sqrt{2T\ln\frac{1}{\delta}} + \frac{c}{1-\alpha}|X|^\alpha|A|^\alpha L^{1-\alpha}t^{1-\alpha} \qquad \text{(E.16)}\\
&\leq 7L|X|\sqrt{2T|A|\ln\frac{2T|X||A|}{\delta}} + \frac{c}{1-\alpha}|X|^\alpha|A|^\alpha L^{1-\alpha}t^{1-\alpha},
\end{aligned}$$

where Inequality (E.14) follows from Lemma B.3 of (Rosenberg and Mansour, 2019b), with probability at least $1-2\delta$, Inequality (E.15) holds by Lemma E.6 with probability at least $1-\delta$, Inequality (E.16) follows from Inequality (E.13). A Union Bound concludes the proof. □

E.5 Violation Bound

In this section, we provide the violation bound of Algorithm 7.1.

Theorem E.4. *Let $\delta \in (0,1)$. Both in the stochastic and in the adversarial setting, with probability at least $1-4\delta$, Algorithm 7.1 attains:*

$$V_t \leq 61L|X|\sqrt{2t|A|\ln\left(\frac{2mT^2|X||A|}{\delta}\right)},$$

for all $t \in [T]$.

Proof. Given an $i \in [m]$, we assume that Corollary E.1 holds with probability $1-3\delta$ for any interval. If $V_{t,i} \leq 61L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)}$ then the statement is trivially satisfied. Otherwise, let us suppose that there exists a $\bar{t} \in T$ for which $V_{\bar{t},i} \geq 61L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)}$. This implies that there exists a $\underline{t} < \bar{t}$ such that $V_{t,i} \geq 44L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)}$ for all $t \in [\underline{t},\bar{t}]$ and $V_{\underline{t}-1,i} \leq 44L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)}$. By Lemma E.6 it holds:

$$V_{t,i} = \sum_{\tau\in[t]} g_{\tau,i}^\top q_\tau \leq \sum_{\tau\in[t]}\sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) + 2L\sqrt{2t\ln\frac{1}{\delta}}.$$

with probability at least $1-\delta$. Therefore, since $V_{t,i} \geq 44L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)}$ for all $t \in [\underline{t},\bar{t}]$, it holds:

$$\begin{aligned}
\sum_{\tau\in[t]}\sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) &\geq 44L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)} - 2L\sqrt{2t\ln\frac{1}{\delta}} \\
&\geq 44L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)} - 2L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)} \\
&= 42L\sqrt{|X||A|t\ln\left(\frac{T^2}{\delta}\right)}.
\end{aligned}$$

Thus, we can write:

$$\begin{aligned}
&\sum_{\tau\in[t]}\sum_{x,a} g_{\tau,i}(x,a)\mathbb{I}_\tau(x,a) - 21L|X|\sqrt{2t|A|\ln\frac{2mT^2|X||A|}{\delta}} \\
&\qquad \geq 42L|X|\sqrt{2t|A|\ln\frac{2mT^2|X||A|}{\delta}} \\
&\qquad\quad - 21L|X|\sqrt{2t|A|\ln\frac{2mT^2|X||A|}{\delta}} \\
&\qquad \geq 21L|X|\sqrt{2t|A|\ln\frac{2mT^2|X||A|}{\delta}}.
\end{aligned} \tag{E.17}$$

and thus $\Gamma_{t,i} = 21L|X|\sqrt{2t|A| \ln \frac{2mT^2|X||A|}{\delta}}$ for all $t \in [\underline{t}, \bar{t}]$.
Therefore on $t \in [\underline{t}, \bar{t}]$ the learning rate can be lower-bounded as:

$$\begin{aligned}
\beta_{t,i}(x,a) &= \frac{(1+\Gamma_t)}{N_t(x,a)} \\
&= \frac{1 + 21L|X|\sqrt{2t|A| \ln \frac{2mT^2|X||A|}{\delta}}}{N_t(x,a)} \\
&\geq 21L|X|\sqrt{\frac{2|A| \ln \frac{2mT^2|X||A|}{\delta}}{N_t(x,a)}},
\end{aligned}$$

exploiting the fact that $N_t(x,a) \leq t$ for all $t \in [T]$.
Therefore, by Corollary E.1, since the constraints learning rates are decreasing in the interval, the following holds:

$$\begin{aligned}
V_{[\underline{t},\bar{t}],i} &\leq \frac{2}{21L\sqrt{|X||A|t \ln\left(\frac{T^2}{\delta}\right)}} \sum_{x \in X, a \in A} \sqrt{N_{\bar{t}}(x,a)} + \sum_{\tau=\underline{t}}^{\bar{t}} b_{\tau-1}^\top \widehat{q}_\tau \\
&\quad + 7L|X|\sqrt{2t|A| \ln \frac{2T^2|X||A|}{\delta}} \\
&\leq \frac{2\sqrt{|X||A|L\bar{t}}}{21L\sqrt{|X||A|t \ln\left(\frac{T^2}{\delta}\right)}} + \sum_{\tau=\underline{t}}^{\bar{t}} b_{\tau-1}^\top \widehat{q}_\tau \\
&\quad + 7L|X|\sqrt{2t|A| \ln \frac{2T^2|X||A|}{\delta}} \qquad \text{(E.18)} \\
&\leq \frac{2\sqrt{|X||A|L\bar{t}}}{21L\sqrt{|X||A|t \ln\left(\frac{T^2}{\delta}\right)}} + 2\sqrt{2|X||A|Lt \ln\left(\frac{2T^2|X||A|}{\delta}\right)} \\
&\quad + 14L|X|\sqrt{2t|A| \ln \frac{2T^2|X||A|}{\delta}} \qquad \text{(E.19)} \\
&\leq \left(\frac{1}{10} + 2 + 14\right) L|X|\sqrt{2t|A| \ln\left(\frac{2T^2|X||A|}{\delta}\right)},
\end{aligned}$$

where Inequality (E.18) holds by Jensen's Inequality and Inequality (E.19) holds by Lemma E.2, under the same event of Corollary E.1. Thus, we have:

$$\begin{aligned}
V_{\bar{t},i} &\leq V_{\underline{t},i} + V_{[\underline{t},\bar{t}],i} \\
&\leq \left(44 + \frac{1}{10} + 2 + 14\right) L|X|\sqrt{2t|A| \ln\left(\frac{2T^2|X||A|}{\delta}\right)} \\
&< 61L|X|\sqrt{2t|A| \ln\left(\frac{2T^2|X||A|}{\delta}\right)}.
\end{aligned}$$

This shows a contradiction, so there is no such $\bar{t}$. Taking a Union Bound on all $i \in [m]$ concludes the proof. □

E.6 Towards the Regret Bound in the Stochastic Setting

In this section, we provide some preliminary results for the stochastic setting. Specifically, throughout the section, we show that the violations are kept small during the learning dynamic, thus making $\widehat{g}_{t,i}$ the empirical mean estimator of the constraint functions. This step is fundamental to show that the decision space of Algorithm 7.1 is suited to guarantee sublinear regret.

Lemma E.3. *Let $\delta \in (0,1)$. With probability at least $1-2\delta$ it holds:*

$$V_{t,i} \leq \sum_{\tau=1}^{t} g_{\tau,i}^{\top} \widehat{q}_{\tau} + 5L|X|\sqrt{2T|A|\ln\frac{2mT|X||A|}{\delta}}, \quad \forall t \in [T], i \in [m].$$

Proof. It holds:

$$\begin{aligned}
V_{t,i} &= \sum_{\tau=1}^{t} g_{\tau,i}^{\top} q^{P,\pi_\tau} \\
&= \sum_{\tau=1}^{t} g_{\tau,i}^{\top} q^{P,\pi_\tau} + \sum_{\tau=1}^{t} g_{\tau,i}^{\top} \widehat{q}_{\tau} - \sum_{\tau=1}^{t-1} g_{\tau,i}^{\top} \widehat{q}_{\tau} \\
&\leq \sum_{\tau=1}^{t} g_{\tau,i}^{\top} \widehat{q}_{\tau} + \|q_\tau - \widehat{q}_\tau\|_1 \\
&\leq \sum_{\tau=1}^{t} g_{\tau,i}^{\top} \widehat{q}_{\tau} + 2L|X|\sqrt{2t\ln\frac{2L}{\delta}} + 3L|X|\sqrt{2t|A|\ln\frac{2T|X||A|}{\delta}} \qquad \text{(E.20)} \\
&\leq \sum_{\tau=1}^{t} g_{\tau,i}^{\top} \widehat{q}_{\tau} + (2+3)L|X|\sqrt{2t|A|\ln\frac{2T|X||A|}{\delta}} \\
&= \sum_{\tau=1}^{t} g_{\tau,i}^{\top} \widehat{q}_{\tau} + 5L|X|\sqrt{2t|A|\ln\frac{2T|X||A|}{\delta}},
\end{aligned}$$

where Inequality (E.20) follows from Lemma B.3 of (Rosenberg and Mansour, 2019b), with probability at least $1-2\delta$. □

We proceed with the following result.

Lemma E.4. *Let $\delta \in (0,1)$. With probability at least $1-3\delta$ it holds:*

$$\sum_{\tau=1}^{t} \bar{g}_i^{\top} \widehat{q}_{\tau} \leq \sum_{\tau=1}^{t} \sum_{x \in X, a \in A} \bar{g}_i(x,a) \mathbb{I}_\tau(x,a) + 7L|X|\sqrt{2t|A|\ln\frac{2mT|X||A|}{\delta}},$$

for all $t \in [T], i \in [m]$.

Proof. It holds:

$$\sum_{\tau=1}^{t} \bar{g}_i^{\top} \widehat{q}_{\tau} = \sum_{\tau=1}^{t} \bar{g}_i^{\top} \widehat{q}_{\tau} + \sum_{\tau=1}^{t} \bar{g}_i^{\top} q^{P,\pi_\tau} - \sum_{\tau=1}^{t} \bar{g}_i^{\top} q^{P,\pi_\tau}$$

$$
\begin{aligned}
&\leq \sum_{\tau=1}^{t} \bar{g}_i^\top q^{P,\pi_\tau} + \|q_\tau - \widehat{q}_\tau\|_1 \\
&\leq \sum_{\tau=1}^{t} \bar{g}_i^\top q^{P,\pi_\tau} + 2L|X|\sqrt{2t \ln \frac{2L}{\delta}} \\
&\quad + 3L|X|\sqrt{2t|A| \ln \frac{2T|X||A|}{\delta}} \qquad \text{(E.21)} \\
&\leq \sum_{\tau=1}^{t} \sum_{x\in X, a\in A} \bar{g}_i(x,a)\mathbb{I}_\tau(x,a) + 2L\sqrt{2t \ln \frac{1}{\delta}} \\
&\quad + 2L|X|\sqrt{2t \ln \frac{2L}{\delta}} + 3L|X|\sqrt{2t|A| \ln \frac{2T|X||A|}{\delta}} \qquad \text{(E.22)} \\
&\leq \sum_{\tau=1}^{t} \sum_{x\in X, a\in A} \bar{g}_i(x,a)\mathbb{I}_\tau(x,a) \\
&\quad + (2+2+3)L|X|\sqrt{2t|A| \ln \frac{2T|X||A|}{\delta}} \\
&= \sum_{\tau=1}^{t} \sum_{x\in X, a\in A} \bar{g}_i(x,a)\mathbb{I}_\tau(x,a) + 7L|X|\sqrt{2t|A| \ln \frac{2T|X||A|}{\delta}},
\end{aligned}
$$

where Inequality (E.21) follows from Lemma B.3 of (Rosenberg and Mansour, 2019b) with probability at least $1-2\delta$, Inequality (E.22) follows from Lemma E.6, with probability at least $1-2\delta$. A Union Bound concludes the proof. □

E.7 Technical Lemmas

In this section, we provide some auxiliary lemmas that are needed throughout the paper.

We start by the following application of the Hoeffding inequality on the constraints.

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

$$
\left| \frac{1}{N_t(x,a)} \sum_{\tau \in \mathcal{T}_{t,x,a}} g_{\tau,i}(x,a) - \bar{g}_i(x,a) \right| \leq \sqrt{\frac{2 \ln \frac{2m|X||A|T}{\delta}}{N_t(x,a)}}.
$$

Proof. The proof is a simple application of Hoeffding's inequality and a Union Bound. □

Thus, we provide concentration results for the constraints.

Lemma E.6. *For any $\delta \in (0,1)$, let $f_t : X \times A \to [-1,1]$ be a sequence of functions that is $t-1$ predictable, and let π_t be a randomized policy. Then, with probability at least $1-\delta$, it holds:*

$$
\left| \sum_{t\in[T]} \sum_{x\in X, a\in A} f_t(x,a)\mathbb{I}_t(x,a) - \sum_{t\in[T]} f_t^\top q^{P,\pi_t} \right| \leq 2L\sqrt{2T \ln \frac{1}{\delta}},
$$

where $\mathbb{I}_t(x,a) = 1$ if and only if the pair (x,a) is visited in episode t.

Proof. By definition of the occupancy measure, it holds:

$$\mathbb{E}\left[f_t(x,a)\mathbb{I}_t(x,a)|P,\pi_t\right] = \sum_{x\in X}\sum_{a\in A} q_t(x,a)f_t(x,a) = f_t^\top q_t.$$

We defined the following sequence:

$$X_t = \sum_{\tau=1}^{t}\left[\sum_{x\in X,a\in A} f_\tau(x,a)\mathbb{I}_\tau(x,a) - f_\tau^\top q^{P,\pi_\tau}\right].$$

X_t is a Martingale difference sequence and $|X_t - X_{t-1}| \le 2L$. Applying the Azuma inequality, we obtain that with probability at least $1-\delta$:

$$\left|\sum_{t\in[T]}\sum_{x\in X,a\in A} f_t(x,a)\mathbb{I}_t(x,a) - \sum_{t\in[T]} f_t^\top q^{P,\pi_t}\right| \le 2L\sqrt{2T\ln\frac{1}{\delta}}.$$

This concludes the proof. □

Lemma E.7. *For any $\delta\in(0,1)$, for any sequence of occupancy measure $\bar{q}_t \in \widehat{\Delta}_t(\mathcal{P}_t)$ and any function $f_t(x,a)$ sampled from a distribution with mean $\bar{f}(x,a)$, i.e., $\mathbb{E}[f_t(x,a)] = \bar{f}(x,a)$ and $\mathbb{P}(|f_t(x,a)|\le 1) = 1$, it holds that with probability at least $1-\delta$:*

$$\left|\sum_{t\in[T]} \bar{f}^\top \bar{q}_t - \sum_{t\in[T]} f_t^\top \bar{q}_t\right| \le 2L\sqrt{2T\ln\frac{1}{\delta}}.$$

Proof. The proof follows the one of Lemma E.6, after noticing that the quantity of interest is a Martingale difference sequence. □

To conclude the section, we provide an auxiliary result on the concentration of the optimistic loss estimator.

Lemma E.8. *For any $\delta\in(0,1)$, Algorithm 7.1 attains, with probability at least $1-7\delta$:*

$$\sum_{t=1}^{T}(\ell_t - \widehat{\ell}_t)^\top \widehat{q}_t \le \gamma|X||A|T + 2L|X|^2\sqrt{2T\ln\left(\frac{L|X|}{\delta}\right)} + 3L|X|^2\sqrt{2T|A|\ln\left(\frac{T|X|^2|A|}{\delta}\right)}$$

Proof. The result follows from the proof of Lemma 6 from (Jin et al., 2020a) and employing Lemma D.10. □

APPENDIX $\mathcal{F}$

Omitted Lemmas and Proofs of Chapter 9

F.1 Events Dictionary

In the following, we introduce the main events that are related to the 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 A.8. The event is related to the estimation of the unknown transition function.

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

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

  Similarly,

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

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

  $\mathcal{E}_G$ holds with probability at least $1 - \delta$ by Corollary F.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 F.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 A.9. 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.

F.2 Confidence Intervals

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

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

Lemma F.1. *For all $i \in [m]$, fixing $(x,a) \in X \times A$, it holds:*

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

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

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

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

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

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

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

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

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 F.2. *Fixing $i \in [m]$, $(x,a) \in X \times A$, $t \in [T]$, for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$:*

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

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

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

where we employed the triangle inequality and the definition of $\widehat{g}_{t,i}(x,a)$.
We bound the two terms in Equation (F.1) separately. For the first term, by Hoeffding's inequality and noticing that constraints values are bounded in $[0,1]$, it holds that:

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

where,

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

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

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

$$\leq \sqrt{\frac{1}{2\max\{N_t(x,a),1\}}\ln\left(\frac{2}{\delta}\right)}.$$

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

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

which concludes the proof. □

We now prove a similar result for the rewards function.

Lemma F.3. *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 F.2. □

We now generalize the previous results as follows.

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

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

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

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

From Lemma F.2, given $\delta'\in(0,1)$, we have, fixed any $i\in[m]$, $t\in[T]$ and $(x,a)\in X\times A$:

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

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

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

Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,i,t}\left\{\left|\widehat{g}_{t,i}(x,a)-g_i^\circ(x,a)\right|\leq\zeta_t(x,a)\right\}\right]\\
&=1-\mathbb{P}\left[\bigcup_{x,a,i,t}\left\{\left|\widehat{g}_{t,i}(x,a)-g_i^\circ(x,a)\right|\leq\zeta_t(x,a)\right\}^c\right]\\
&\geq 1-\sum_{x,a,i,t}\mathbb{P}\left[\left\{\left|\widehat{g}_{t,i}(x,a)-g_i^\circ(x,a)\right|\leq\zeta_t(x,a)\right\}^c\right] \qquad \text{(F.2)}\\
&\geq 1-|X||A|mT\delta',
\end{aligned}$$

where Inequality (F.2) holds by Union Bound. Noticing that $g_{t,i}(x,a)\leq 1$, substituting δ' with $\delta:=\delta'/|X||A|mT$ in $\zeta_t(x,a)$ with an additional Union Bound over the possible values of $N_t(x,a)$, we have, with probability at least $1-\delta$:

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

which concludes the proof. □

We provide a similar result for the rewards function.

Lemma F.5. *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 F.3, given $\delta'\in(0,1)$, we have fixed any $t\in[T]$ and $(x,a)\in X\times A$:

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

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

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

Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{x,a,t}\left\{\left|\widehat{r}_t(x,a)-r^\circ(x,a)\right|\leq\psi_t(x,a)\right\}\right]\\
&\qquad=1-\mathbb{P}\left[\bigcup_{x,a,t}\left\{\left|\widehat{r}_t(x,a)-r^\circ(x,a)\right|\leq\psi_t(x,a)\right\}^c\right]\\
&\qquad\geq 1-\sum_{x,a,t}\mathbb{P}\left[\left\{\left|\widehat{r}_t(x,a)-r^\circ(x,a)\right|\leq\psi_t(x,a)\right\}^c\right] \qquad \text{(F.3)}\\
&\qquad\geq 1-|X||A|T\delta',
\end{aligned}$$

where Inequality (F.3) holds by Union Bound. Noticing that $r_t(x,a)\leq 1$, substituting δ' with $\delta:=\delta'/|X||A|T$ in $\psi_t(x,a)$ with an additional Union Bound over the possible values of $N_t(x,a)$, we have, with probability at least $1-\delta$:

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

which concludes the proof. □

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

Lemma F.6. *Given $\delta\in(0,1)$, for all episodes $t\in[T]$, state-action pairs $(x,a)\in X\times A$ and constraint $i\in[m]$,it holds, with probability at least $1-\delta$:*

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

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

Proof. We first notice that if $\xi_t(x,a)=1$, the results is derived trivially by definition on the cost function. We prove now the non trivial case:

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

Employing Lemma F.1 and Lemma F.4, 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:

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

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

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

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

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

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

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

Proof. Following the same analysis of Lemma F.6 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 F.2. *Taking the definition of ξ_t employed in Lemma F.6 and defining $\mathcal{E}_G$ as the intersection event:*

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

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

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

Notice that by Corollary F.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 F.7. *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 F.1 and Lemma F.5, with probability at least $1-\delta$, for all $(x,a) \in X \times A$ and for all $t \in [T]$, it holds:

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

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 F.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 F.1. □

Corollary F.4. *Taking the definition of ϕ_t employed in Lemma F.7 and defining $\mathcal{E}_r$ as the intersection event:*

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

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

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

F.3 Omitted Proofs when the Corruption is Known

In the following, we provide the additional result attained by Algorithm 9.1 when the corruption of the environment is known to the learner.

We show that the linear program solved by Algorithm 9.1 at each $t \in [T]$ admits a feasible solution, with high probability.

Lemma F.8. *For any* $\delta \in (0, 1)$*, for all episodes* $t \in [T]$*, with probability at least* $1 - 5\delta$*, the space defined by linear constraints* $\left\{q \in \Delta(\mathcal{P}_t) : \underline{G}_t^\top q \leq \theta\right\}$ *admits a feasible solution and it holds:*

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

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

F.4 Omitted Proofs when the Knowledge of C is not Precise

In this section, we focus on the performances of Algorithm 9.1 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 F.9. *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 9.1 using* $\widehat{C}_G$ *as corruption value, for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, for all episodes* $t \in [T]$*, state-action pair* $(x, a) \in X \times A$ *and constraint* $i \in [m]$*, the following result holds:*

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

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

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

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

and constraints $i \in [m]$:

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

which can be rewritten as:

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

where

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

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

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

Theorem F.1. *For any $\delta \in (0,1)$, Algorithm 9.1, when instantiated with corruption value $\widehat{C}$ which is an overestimate of the true value of C, i.e. $\widehat{C} > C_G$ and $\widehat{C} > C_r$, attains with probability at least $1 - 8\delta$:*

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

Proof. By Corollary F.1, it holds that the decision space of the linear program performed by Algorithm 9.1 contains with high probability the optimal solution that respects to the constraints. Furthermore, employing Corollary F.3 and following the proof of Theorem 9.2 concludes the proof. □

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

Theorem F.2. *For any $\delta \in (0,1)$, Algorithm 9.1, 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 $1 - 9\delta$:*

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

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

$$\mathcal{V}_T = \max_{i\in[m]} \sum_{t\in[T]} \left[\mathbb{E}[G_t]^\top q_t - \theta\right]_i^+ \tag{F.4a}$$

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

$$\leq \max_{i\in[m]} \sum_{t\in[T]} \Bigg[\left(\mathbb{E}[g_{t,i}] - g_i^\circ\right)^\top q_t + \underline{g}_{t-1,i}^\top (q_t - \widehat{q}_t) + \underline{g}_{t-1,i}^\top \widehat{q}_t + 2\xi_{t-1}^\top q_t$$

$$+ \sum_{x,a} \left(\frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} + \frac{\epsilon}{T}\right) q_t(x,a) - \theta_i\Bigg]^+ \tag{F.4b}$$

$$\leq C_G + 2\max_{i\in[m]} \sum_{t\in[T]} \xi_{t-1}^\top q_t + \sum_{t\in[T]} \|q_t - \widehat{q}_t\|_1$$

$$+ \sum_{t\in[T]} \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a) + \epsilon L, \tag{F.4c}$$

where Inequality (F.4b) follows from Lemma F.9 and Inequality (F.4c) is derived as in the proof of Theorem 9.1, and considering that $\|q_t\|_1 = L,\ \forall t \in [T]$. Now, employing the Azuma-Hoeffding inequality, we can bound, with probability at least $1-\delta$ the term $\sum_{t=1}^T \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a)$ as follows:

$$\sum_{t\in[T]} \sum_{x,a} \frac{\epsilon}{\max\{N_{t-1}(x,a),1\}} q_t(x,a)$$

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

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

where we applied Azume Hoeffding inequality and the fact that $\sum_{t\in[N_T(x,a)]} \frac{1}{t} \leq 1 + \ln(T)$. Finally, following the steps of the proof of Theorem 9.1 to bound the first 3 elements of Inequality (F.4c) 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 $1-9\delta$,

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

which concludes the proof. □

Finally, we provide the violation bound attained by Algorithm 9.1 when an overestimate of the corruption value is given as input.

Theorem F.3. *For any $\delta \in (0,1)$, Algorithm 9.1, 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 $1 - 8\delta$:

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

Proof. The proof follows by employing Corollary F.1 to the proof of Theorem 9.1. □

F.5 Omitted Proofs when the Corruption is *not* Known

In the following section, we provide the omitted proofs of the theoretical guarantees attained by Algorithm 9.2. The algorithm is designed to work when the corruption value is *not* known.

F.5.1 Additional Notation

In the following sections, we will refer as:

$$\widehat{V}_T := \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j\top}\widehat{q}_t^{\,j} - \theta_i\right]^+, \tag{F.5}$$

to the estimated violation attained by the instances of Algorithm 9.2. Furthermore, we will refer as:

$$\widehat{V}_{T,j^*} := \sum_{t\in[T]}\frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*}+\gamma}\sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j^*\top}\widehat{q}_t^{\,j^*} - \theta_i\right]^+, \tag{F.6}$$

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 $\mathbb{I}_t(x,a)\mathbb{I}(j_t = j) = 1$.

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

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

F.5.2 Theoretical Results and Analysis

We start providing some preliminary results on the optimistic estimator employed by Algorithm 9.2.

Lemma F.10. *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:

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

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 9.2.

Lemma F.11. *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 F.5. We define the loss $\bar{\ell}_t=\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))$, the optimistic loss estimator $\widehat{\ell}_t:=\frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))$ and the unbiased estimator $\widetilde{\ell}_t:=\frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}}\sum_{(x_k^t,a_k^t)}(1-r_t(x_k^t,a_k^t))$. Employing the same argument as Neu (2015) it holds:

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

since $\frac{z}{1+z/2}\leq\ln(1+z), z\in\mathbb{R}_{\geq 0}$. 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:

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

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

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 F.12. *For any $\delta\in(0,1)$ and properly setting the learning rate η such that $\eta\leq\frac{1}{2\Lambda m\left(\sqrt{\beta_1 T}+\beta_2+\beta_5+\sqrt{\beta_4 T}\right)}$, Algorithm 9.2 attains, with probability at least $1-2\delta$:*

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

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

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

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

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

with $\nu_{t,j}=\max_{\tau\in[t]}\frac{1}{w_{\tau,j}}$.
First we prove that $\eta w_{t,j}|\ell_{t,j}-b_{t,j}|\leq 1/2$ for all $t\in[T],j\in[M]$, to apply Lemma F.15. It holds that $\eta w_{t,j}|\ell_{t,j}|\leq\frac{\eta(L\rho+L^2m^2+Lm)}{\rho}\leq\frac{1}{2}$ for all $j\in[M]$, for all $t\in[T]$ as long as $\eta\leq\frac{\rho}{2(L\rho+L^2m^2+Lm)}\leq\frac{\rho}{2(L^2m^2+Lm)}$, which is true if $\eta\leq\frac{\rho}{2Lm(Lm+1)}$. It also holds that:

$$\eta w_{t,j}|b_{t,j}|$$
$$=\eta w_{t,j}\left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)$$

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

if $\eta \leq \frac{1}{2\Lambda m\left(\sqrt{\beta_1 T} + \beta_2 + \beta_5 + \sqrt{\beta_4 T}\right)}$, where we used the fact that $\nu_{t,j} \neq \nu_{t-1,j} \iff {}^{1}/_{w_{t,j}} = \nu_{t,j}$. Thus, if the previous conditions on η hold, and notice that the second condition implies the first, Algorithm 9.2 attains, by Lemma F.15 :

$$
\begin{aligned}
\sum_{t\in[T]} \Bigg[\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t = j)}{w_{t,j} + \gamma} \sum_{(x_k^t, a_k^t)} (1 - r_t(x_k^t, a_k^t)) \\
- \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*} + \gamma} \sum_{(x_k^t, a_k^t)} (1 - r_t(x_k^t, a_k^t)) \Bigg] + \frac{Lm+1}{\rho} \widehat{V}_T \\
\leq \frac{M \ln T}{\eta} + 2\eta \frac{TM(L\rho + L^2m^2 + Lm)^2}{\rho^2} \\
+ 2\eta \Bigg(2 \left(\frac{m(mL+1)}{\rho} \beta_5 + \beta_2 \right)^2 M \ln(T) \\
+ 2T \left(\sqrt{\beta_1} + \left(\frac{m(Lm+1)}{\rho} \right) \sqrt{\beta_4} \right)^2 M \ln(T) \Bigg) \\
+ \frac{Lm+1}{\rho} \widehat{V}_{T,j^*} + \sum_{t\in[T]} \sum_{j\in[M]} w_{t,j} b_{t,j} - \sum_{t\in[T]} b_{t,j^*},
\end{aligned} \tag{F.7}
$$

where we used the following inequalities:

- First inequality:

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

- Second inequality:

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

- Third inequality:

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

and that, it holds:

$$\begin{aligned}
&\sum_{t\in[T]}\sum_{j\in[M]} w_{t,j}^2 b_{t,j}^2 = \sum_{t\in[T]}\sum_{j\in[M]} (w_{t,j}b_{t,j})^2 \\
&\leq \left(\left(\frac{m(Lm+1)}{\rho}\beta_5+\beta_2\right)+\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)\sqrt{T}\right)^2 \\
&\qquad\qquad \cdot \sum_{j\in[M]}\sum_{t\in[T]}\left(\frac{1}{\nu_{t,j}}(\nu_{t,j}-\nu_{t-1,j})\right)^2 && \text{(F.8a)} \\
&\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right) \\
&\qquad\qquad \cdot \sum_{j\in[M]}\sum_{t\in[T]}\left(1-\frac{\nu_{t-1,j}}{\nu_{t,j}}\right)^2 \\
&\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right) \\
&\qquad\qquad \cdot \sum_{j\in[M]}\sum_{t\in[T]}\left(1-\frac{\nu_{t-1,j}}{\nu_{t,j}}\right) \\
&\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right) \\
&\qquad\qquad \cdot \sum_{j\in[M]}\sum_{t\in[T]}\ln\left(\frac{\nu_{t,j}}{\nu_{t-1,j}}\right) && \text{(F.8b)} \\
&\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right) \\
&\qquad\qquad \cdot \sum_{j\in[M]}\ln\left(\prod_{t\in[T]}\frac{\nu_{t,j}}{\nu_{t-1,j}}\right) \\
&\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right) \\
&\qquad\qquad \cdot \sum_{j\in[M]}\ln\left(\frac{\nu_{T,j}}{\nu_{0,j}}\right) \\
&\leq \left(2\left(\frac{m(mL+1)}{\rho}\beta_5+\beta_2\right)^2+2T\left(\frac{m(Lm+1)}{\rho}\sqrt{\beta_4}+\sqrt{\beta_1}\right)^2\right) \\
&\qquad\qquad \cdot M\ln(T), && \text{(F.8c)}
\end{aligned}$$

where Inequality (F.8a) is true since $\nu_{t,j}-\nu_{t-1,j}\neq 0$ only when $w_{t,j}=1/\nu_{t,j}$ by definition, Inequality (F.8b) holds since $1-a\leq -\ln a$, and Inequality (F.8c) holds since by definition $\nu_{T,j}\leq T$ and $\nu_{0,j}=M$. Notice also that, following a similar

reasoning, it holds:

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

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

$$\begin{aligned}
&\sum_{t\in[T]} \mathbb{E}[r_t]^\top q_t^{j^*} - \sum_{t\in[T]}\sum_{j\in[M]} w_{t,j}\mathbb{E}[r_t]^\top q_t^j + \frac{Lm+1}{\rho}\widehat{V}_T \\
&= \sum_{t\in[T]}\sum_{j\in[M]} w_{t,j}\left(L - \mathbb{E}[r_t]^\top q_t^j\right) - \sum_{t\in[T]}\left(L - \mathbb{E}[r_t]^\top q_t^{j^*}\right) + \frac{Lm+1}{\rho}\widehat{V}_T \qquad \text{(F.9)} \\
&\leq \mathcal{O}\Bigg(\frac{M\ln T}{\eta} + \eta\, m^4 L^4 T M + \eta\, M \ln(T) m^4 L^2 \beta_5^2 + \eta\, M \ln(T)\beta_2^2 \\
&\qquad + \eta T(\beta_1 + L^2 m^4 \beta_4) M \ln(T) + \gamma T L M + L\sqrt{T\ln(1/\delta)} + \frac{L}{\gamma}\ln(1/\delta)\Bigg) \\
&\qquad + \frac{Lm+1}{\rho}\widehat{V}_{T,j^*} - \left(\frac{m(mL+1)}{\rho}\beta_5 + \beta_2\right)\nu_{T,j^*} \\
&\qquad - \left(\sqrt{\beta_1} + \left(\frac{m(Lm+1)}{\rho}\right)\sqrt{\beta_4}\right)\sqrt{T}\nu_{T,j*}, \qquad \text{(F.10)}
\end{aligned}$$

where Equation (F.9) holds since $\sum_{j\in[M]} w_{t,j} = 1$, $\forall t \in [T]$, and Inequality (F.10) holds, with probability at least $1-2\delta$, by Lemma F.10, Lemma F.11 and Equation (F.7). This concludes the proof. □

In order to provide the desired bound R_T and $\mathcal{V}_T$ for Algorithm 9.2, it is necessary to study the relation between the aforementioned performance measures and the terms appearing from the FTRL analysis in Lemma F.12.

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

Lemma F.13. *For any $\gamma \in \mathbb{R}_{\geq 0}$, given $\delta \in (0,1)$, with probability at least $1-10\delta$, it*

holds:

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

Proof. We start defining the quantity $\widehat{\xi}_{t,j}(x,a)$ – for all episode $t \in [T]$, for all state-action pairs $(x,a) \in X \times A$, for all instance $j \in [M]$ – as in Theorem 9.1 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,

$$\widehat{\xi}_{t,j}(x,a) := \min\Bigg\{1, \sqrt{\frac{1}{2\max\{N_t^j(x,a),1\}}\ln\left(\frac{2mMT|X||A|}{\delta}\right)} + \frac{C}{\max\{N_t^j(x,a),1\}} + \frac{C}{T}\Bigg\}, \tag{F.11}$$

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

$$\widehat{g}_{t,i}^{j}(x,a) + \widehat{\xi}_{t,j}(x,a) \geq g_i^{\circ}(x,a). \tag{F.12}$$

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

$$\begin{aligned}
\widehat{V}_T &= \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\sum_{i\in[m]}\left[\widehat{g}_{t,i}^{j\top}\widehat{q}_t^j - \theta_i\right]^+ \\
&= \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\cdot \\
&\quad\cdot\sum_{i\in[m]}\left[(\widehat{g}_{t,i}^{j\top}q_t^j + \widehat{\xi}_{t,j}^{\top}q_t^j - \theta_i) - \widehat{\xi}_{t,j}^{\top}q_t^j - \widehat{g}_{t,i}^{j\top}(q_t^j - \widehat{q}_t^j)\right]^+ \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\cdot \\
&\quad\cdot\sum_{i\in[m]}\left(\left[(\widehat{g}_{t,i}^{j}+\widehat{\xi}_{t,j})^{\top}q_t^j - \theta_i\right]^+ - \widehat{\xi}_{t,j}^{\top}q_t^j - \widehat{g}_{t,i}^{j\top}|q_t^j - \widehat{q}_t^j|\right) \qquad \text{(F.13a)} \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\cdot \\
&\quad\cdot\sum_{i\in[m]}\left(\left[g_i^{\circ\top}q_t^j - \theta_i\right]^+ - \widehat{\xi}_{t,j}^{\top}q_t^j - \|q_t^j - \widehat{q}_t^j\|_1\right) \qquad \text{(F.13b)} \\
&\geq \sum_{t\in[T]}\sum_{j\in[M]}\frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma}\cdot
\end{aligned}$$

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

$$\cdot \sum_{i\in[m]} \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^j \right]^+ - \mathcal{O}\left(mL|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)} \right), \quad \text{(F.13c)}$$

where Inequality (F.13a) holds since $[a-b]^+ \geq [a]^+ - b$, $a \in \mathbb{R}, b \in \mathbb{R}_{\geq 0}$, Inequality (F.13b) follows from Inequality (F.12) 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 (F.13c) holds under event $\mathcal{E}_{\widehat{q}}$ by Lemma A.9 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 $\sum_{t\in[T]}\sum_{j\in[M]} \frac{w_{t,j}\mathbb{I}(j_t=j)}{w_{t,j}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^j - \theta_i\right]^+$.
We bound the term by the Azuma-Hoeffding inequality. Indeed, with probability at least $1-\delta$, it holds:

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

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

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

Moreover, we observe the following bounds:

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

and,

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

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

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

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:

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

where we employed the Azuma-Hoeffding inequality and where the last step holds following the proof of Theorem 9.1.

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

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

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

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

Thus, plugging the previous bounds in Equation (F.13c), we obtain, with probability at least $1-10\delta$:

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

This concludes the proof. □

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

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

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

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:

$$\begin{aligned}
&\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[\widehat{g}_{t,i}^{j^*\top}\widehat{q}_t^{j^*} - \theta_i\right]^+ \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \Bigg(\left[\widehat{g}_{t,i}^{j^*\top}(\widehat{q}_t^{j^*} - q_t^{j^*}) + \widehat{g}_{t,i}^{j^*\top} q_t^{j^*} + \right. \\
&\qquad\qquad \left. - \widehat{\xi}_{t,j^*}^\top q_t^{j^*} - \theta_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*}\Bigg) \qquad \text{(F.14a)} \\
&\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \Bigg(\left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \theta_i\right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*} + \\
&\qquad\qquad \left. \right.
\end{aligned}$$

$$+ \left[g_i^{\circ\top} q_t^{j^*} - \mathbb{E}[g_{t,i}]^\top q_t^{j^*} \right]^+ + \|\widehat{q}_t^{j^*} - q_t^{j^*}\|_1 \Bigg) \tag{F.14b}$$

$$\leq \sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*} + \gamma} \sum_{i\in[m]} \Bigg(\left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \theta_i \right]^+ + \widehat{\xi}_{t,j^*}^\top q_t^{j^*} + \\ + \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^{j^*} \right]^+ \Bigg) \\ + \mathcal{O}\left(L|X| \sqrt{|A| T \ln\left(\frac{T|X||A|}{\delta} \right)} \right), \tag{F.14c}$$

where Inequality (F.14a) holds since $[a+b]^+ \leq [a]^+ + [b]^+$, $\forall a, b \in \mathbb{R}$ and by the definition of $\widehat{\xi}_{t,j^*}$ (see Equation (F.11)) which implies that all its elements are positive, Inequality (F.14b) holds with probability at least $1-\delta$ by Corollary F.2 and by union bound over M, and since that $\|\widehat{g}_{t,i}\|_\infty \leq 1$ and Inequality (F.14c) holds with probability at least $1-6\delta$ by Lemma A.9.

Upper-bound to $\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[(g_i^\circ - \mathbb{E}[g_{t,i}])^\top q_t^{j^*} \right]^+$.
It is immediate to bound the quantity of interest employing the definition of corruption C and by Lemma F.16. Indeed, with probability at least $1-\delta$:

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

Upper-bound to $\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \theta_i \right]^+$.
We bound the quantity of interest as follows. With probability at least $1-11\delta$, it holds:

$$\sum_{t\in[T]} \frac{\mathbb{I}(j_t = j^*)}{w_{t,j^*} + \gamma} \sum_{i\in[m]} \left[\mathbb{E}[g_{t,i}]^\top q_t^{j^*} - \theta_i \right]^+ \\ \leq m\sqrt{\beta_4 T} \nu_{T,j^*} + m\beta_5 \nu_{T,j^*} + 2m\beta_6 C + Lm \frac{\ln\left(\frac{M}{\delta}\right)}{2\gamma}, \tag{F.15a}$$

thank to Corollary F.5 and Corollary F.6 .

Upper-bound to $\sum_{t\in[T]} \frac{\mathbb{I}(j_t=j^*)}{w_{t,j^*}+\gamma} \sum_{i\in[m]} \widehat{\xi}_{t,j^*}^\top q_t^{j^*}$.
First, notice that, with probability at least $1-\delta$, it holds:

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

where we employed Lemma F.16. Now we observe that, with probability at least $1-\delta$, it holds:

$$\sum_{t=1}^{T} \widehat{\xi}_{t-1,j^*}^\top q_t \mathbb{I}(j_t = j^*)$$

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

where employed the same steps as in the proof of Theorem 9.1, considering that the counter increases if and only if $\mathbb{I}_t(x,a)\mathbb{I}(j_t = j^*) = 1$.
Combining the previous bounds concludes the proof. □

F.6 Auxiliary Lemmas from Existing Works

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

F.6.1 Auxiliary Lemmas for the FTRL Master Algorithm

In the following, we provide the optimization bound attained by the FTRL instance employed by Algorithm 9.2.

Lemma F.15 (Jin et al. (2023)). *The FTRL algorithm over a convex subset Ω of the $(M-1)$-dimensional simplex Δ_M :*

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

ensures for all $u \in \Omega$:

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

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

F.6.2 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 F.16 (Neu (2015)). *Let (γ_t) be a fixed non-increasing sequence with $\gamma_t \geq 0$ and let $\alpha_{t,j}$ be nonnegative $\mathcal{F}_{t-1}$-measurable random variables satisfying $\alpha_{t,j} \leq 2\gamma_t$ for all t and j. Then, with probability at least $1 - \delta$,*

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

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

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

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

F.7 Auxiliary Lemmas for Stability

In this section, we state the results related to the stability of the arm-algorithms when C is not known. The procedure is inspired by Jin et al. (2023) and Agarwal et al. (2017), but adapted to the case of *Constrained* MDP in high probability. We first give some important definitions. In these definitions we will use C_t as the value of adversarial corruption at episode $t \in [T]$, where C_t is defined as $C_t := \max\{C_t^G, C_t^r\}$, which meets the requirement of upper bounding the adversarial corruption at each considered episode. In addition it holds that $\sum_{t\in[T]} C_t \leq C_r + C_G$ or equivalently $C \leq \sum_{t\in[T]} C_t \leq 2C$, which does not influence the order of the analysis.

Definition F.1. *A CMDP algorithm is **corruption-robust** if it takes ϕ (a guess on the corruption amount) as input, and achieves for any random stopping time $t' \leq T$, whenever $\sum_{t\in[t']} C_t < \phi$:*

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

and

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

Notice that Algorithm 9.1 is corruption-robust after applying a doubling trick to make it work for any stopping time, with probability at least $1 - 9\delta$ thank to Theorem F.1 and Theorem F.3. Furthermore, we introduce the notion of α-stability. An algorithm is considered to be α-stable, if its regret under condition imposed by Algorithm 9.2 is of order $\nu_T^\alpha \cdot \widetilde{\mathcal{O}}\left(R_T\right)$, where R_T is the upper bound on the regret attained by the algorithm if it receives feedback at each episode. In particular, we are interested in the 1-stability.

Definition F.2. *An algorithm is 1-**stable** if, under the condition imposed by Algorithm 9.2, it holds:*

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

and

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

We can use the procedure defined by Algorithm F.1 - and originally proposed by Jin et al. (2023) - to transform a generic corruption robust algorithm to a 1-stable algorithm. Differently from Jin et al. (2023), in our setting, we use the natural symmetry between regret and strong cumulative constraint violation to stabilize both the regret and the strong cumulative constraint violation. We have a different bound for C_t (value of adversarial

corruption at episode t): indeed, $C_t \leq \max\{\|\mathbb{E}[r_t] - r^\circ\|_1, \max_{i\in[m]}\|\mathbb{E}[g_{t,i}] - g_i^\circ\|_1\}$ is bounded by $|X||A|$. Finally, we are interested in obtaining results that hold in high probability rather than in expectation. To do so, we focus on 1-stability guarantee rather than $1/2$-stability as in Jin et al. (2023) since removing the expectation prevents us from achieving the result above with lower coefficients. We can state the following result.

Lemma F.17. *Given an algorithm which is corruption robust according to Definition F.1 with parameters* $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ *and* $\beta_1 \geq \mathcal{O}(L^2 \ln(T/\delta))$, $\beta_4 \geq \mathcal{O}(L^2 \ln(T/\delta))$, *with probability at least* $1 - p$ *with* $p \in (0, 1)$, *then, it is possible convert it to an* 1*-stable algorithm with probability at least* $1 - p - 2\delta$ *according to Definition F.2 with parameters* $(\beta_1', \beta_2', \beta_3', \beta_4', \beta_5', \beta_6')$ *as* $\beta_1' = \mathcal{O}(\beta_1), \beta_2' = \mathcal{O}(\beta_2 + \beta_3|X||A|\ln(\ln(T)/\delta)), \beta_3' = \mathcal{O}(\beta_3 \ln(T)), \beta_4' = \mathcal{O}(\beta_4), \beta_5' = \mathcal{O}(\beta_5 + \beta_6|X||A|\ln(\ln(T)/\delta)), \beta_6' = \mathcal{O}(\beta_6 \ln(T))$, *employing Algorithm F.1.*

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

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

by the corruption-robust property of instance k. We study now the quantity $\mathbb{P}\left(\bigcup_{k\in[M]}\{\sum_{t\in[T]} C_t d_{t,k} h_{t,k} > \phi_k\}\right)$. Notice that $\mathbb{E}[h_{t,k}|d_{t,k}] = 2^{-k-1} d_{t,k}$, and since $d_{t,k}$ is an indicator function then $\mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} = \mathbb{E}[h_{t,k}|d_{t,k}]$. In addition, since $\sum_{t\in[T]} C_t \leq 2C$, it holds:

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

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

$$\begin{aligned}
&\sum_{t\in[T]} C_t d_{t,k} h_{t,k} - \sum_{t\in[T]} C_t \mathbb{E}[h_{t,k}|d_{t,k}] d_{t,k} \\
&\qquad \leq 2\sqrt{\sum_{t\in[T]} C_t^2 d_{t,k} \mathbb{E}[h_{t,k}|d_{t,k}] \ln\left(\frac{\log_2(T)}{\delta}\right)} \\
&\qquad\quad + |X||A| \ln\left(\frac{\log_2(T)}{\delta}\right) \qquad \text{(F.16a)} \\
&\qquad \leq 2\sqrt{|X||A| \sum_{t\in[T]} C_t d_{t,k} \mathbb{E}[h_{t,k}|d_{t,k}] \ln\left(\frac{\log_2(T)}{\delta}\right)} \\
&\qquad\quad + |X||A| \ln\left(\frac{\log_2(T)}{\delta}\right) \qquad \text{(F.16b)}
\end{aligned}$$

$$\leq \sum_{t\in[T]} C_t \mathbb{E}[h_{t,k}|d_{t,k}]d_{t,k} + 2|X||A| \ln\left(\frac{\log_2(T)}{\delta}\right), \tag{F.16c}$$

where Inequality (F.16a) holds with probability at least $1 - \delta/\log_2(T)$ by Freedman inequality, Inequality (F.16b) holds since $C_t \leq |X||A|$, and Inequality (F.16c) holds by AM-GM inequality. Therefore, it holds simultaneously for all $k \in [M]$:

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

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

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

noticing that $\mathbb{E}\left[d_{t,k}\left(\mathbb{E}\left[h_{t,k}|d_{t,k}\right] - h_{t,k}\right)\right] = \mathbb{E}\left[h_{t,k}|d_{t,k}\right] - \mathbb{E}[h_{t,k}]d_{t,k} = \mathbb{E}\left[h_{t,k}|d_{t,k}\right] - \mathbb{E}\left[h_{t,k}|d_{t,k}\right] = 0$, since the expectation is taken w.r.t. the randomization of Algorithm F.1 and the distribution generated given the external probability of receiving feedback w_t.
To conclude with probability at least $1 - p - 2\delta$:

$$\begin{aligned}&\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)\mathbb{I}\left(w_t \geq \frac{1}{T}\right)\\ &\quad\leq \sum_{k\in[M]}\sum_{t\in[T]} \overline{r}^\top (q^* - q_t)d_{t,k}\end{aligned}$$

$$\leq \mathcal{O}\left(\sqrt{\beta_1 T}\max_{t\in[T]}\frac{1}{w_t} + (\beta_2 + \beta_3|X||A|\ln(\ln(T)/\delta))\max_{t\in[T]}\frac{1}{w_t} + \beta_3\ln(T)C\right)$$
$$\leq \mathcal{O}\left(\left(\sqrt{\beta_1' T} + \beta_2'\right)\nu_T + \beta_3' C\right),$$

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

Algorithm F.1 Adapted `STABILIZE` (Jin et al., 2023)

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

Initialize $M = \log_2(T)$ instance of Algorithm 9.1, each instance $k \in [M]$ initialized with corruption parameter:

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

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

We conclude providing the following corollaries.

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

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

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

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

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

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

Bibliography

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

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

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

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

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

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

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

Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119, 2023.

Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957. ISSN 00959057, 19435274. URL http://www.jstor.org/stable/24900506.

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

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Beyond primal-dual methods in bandits with stochastic and adversarial constraints. *Advances in Neural Information Processing Systems*, 37: 8541–8568, 2024.

Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization. In *International Conference on Machine Learning*, pages 3877–3898. PMLR, 2025.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2767–2783. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/castiglioni22a.html.

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

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

Nicolo Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). *Advances in Neural Information Processing Systems*, 25, 2012.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pages 3123–3148. PMLR, 2022.

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

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, pages 3304–3312. PMLR, 2021.

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

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

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

Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. Safe linear bandits over unknown polytopes. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1755–1795. PMLR, 2024.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free RL. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1054–1062. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/ghosh24a.html.

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

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

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

Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pages 402–411. PMLR, 2016.

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 2020a. URL https://proceedings.mlr.press/v119/jin20c.html.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020b.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

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*, 2023.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. URL https://books.google.it/books?id=bydXzAEACAAJ.

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

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

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

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

Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1):2503–2528, 2012.

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

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

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

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

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

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

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

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15277–15287. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ae95296e27d7f695f891cd26b4f37078-Paper.pdf.

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

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

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

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

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

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained mdps via policy optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025a.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning adversarial mdps with stochastic hard constraints. In *Forty-Second International Conference on Machine Learning*, 2025b.

Francesco Emanuele Stradi, Eleonora Fidelia Chiefari, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Beyond slater's condition in online cmdps with stochastic and adversarial constraints, 2025c. URL https://arxiv.org/abs/2509.20114.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Taming adversarial constraints in cmdps. *Advances in Neural Information Processing Systems*, 2025d.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

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, 2022a.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3274–3307. PMLR, 28–30 Mar 2022b. URL https://proceedings.mlr.press/v151/wei22a.html.

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

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

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

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

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30, 2017.

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