Truly Adapting to Adversarial Constraints in Constrained MABs

Francesco Emanuele Stradi FRANCESCOEMANUELE.STRADI@POLIMI.IT
Politecnico di Milano

Kalana Kalupahana KALANAKALPITHA.KALUPAHANA@MAIL.POLIMI.IT
Politecnico di Milano

Matteo Castiglioni MATTEO.CASTIGLIONI@POLIMI.IT
Politecnico di Milano

Alberto Marchesi ALBERTO.MARCHESI@POLIMI.IT
Politecnico di Milano

Nicola Gatti NICOLA.GATTI@POLIMI.IT
Politecnico di Milano

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the constrained variant of the *multi-armed bandit* (MAB) problem, in which the learner aims not only at minimizing the total loss incurred during the learning dynamic, but also at controlling the violation of multiple *unknown* constraints, under both *full* and *bandit feedback*. We consider a non-stationary environment that subsumes both stochastic and adversarial models and where, at each round, both losses and constraints are drawn from distributions that may change arbitrarily over time. In such a setting, it is provably not possible to guarantee both sublinear regret and sublinear violation. Accordingly, prior work has mainly focused either on settings with stochastic constraints or on relaxing the benchmark with fully adversarial constraints (*e.g.*, via competitive ratios with respect to the optimum). We provide the first algorithms that achieve optimal rates of regret and *positive* constraint violation when the constraints are stochastic while the losses may vary arbitrarily, and that simultaneously yield guarantees that degrade smoothly with the degree of adversariality of the constraints. Specifically, under *full feedback* we propose an algorithm attaining $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ regret and $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ positive violation, where C quantifies the amount of non-stationarity in the constraints. We then show how to extend these guarantees when only bandit feedback is available for the losses. Finally, when *bandit feedback* is available for the constraints, we design an algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ positive violation and $\widetilde{\mathcal{O}}(\sqrt{T}+C\sqrt{T})$ regret.

Keywords: Online Learning, Multi-Armed Bandits, Constraints

1. Introduction

Over the past few years, *constrained multi-armed bandit* (MAB) problems have attracted growing attention in learning theory (see, *e.g.*, (Liakopoulos et al., 2019; Pacchiano et al., 2021; Castiglioni et al., 2022b)). In the unconstrained MAB setting, the learner is evaluated solely in terms of *regret*, which captures the gap between the learner's performance and that of an *optimal-in-hindsight* (fixed) action. Constrained variants introduce additional challenges: while learning, the learner must not only minimize regret, but also ensure that the prescribed constraints are not violated excessively.

In these settings, a well-known impossibility result by Mannor et al. (2009) shows that it is provably impossible to achieve both regret and constraint violation that are sublinear in the number of rounds T when the constraints are chosen adversarially, *i.e.*, when they can change arbitrarily

across rounds.[1] For this reason, the literature has largely focused on different regimes. Indeed, many works study settings in which both losses and constraints are sampled i.i.d. from a fixed distribution at each round (Efroni et al., 2020; Gangrade et al., 2024), or where constraints are stochastic while losses may be adversarial (Qiu et al., 2020; Stradi et al., 2025a). In both cases, optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and constraint violation are attainable. More recent contributions aim to provide best-of-both-worlds guarantees, allowing constraints to be either stochastic or adversarial (Castiglioni et al., 2022b; Stradi et al., 2024; Bernasconi et al., 2024). These works typically guarantee $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation in the stochastic regime, while in the adversarial regime they ensure sublinear violation together with sublinear α-regret, *i.e.*, regret measured against a fraction of the optimal reward. However, these algorithms come with several drawbacks. For instance, when constraints are only mildly adversarial, they may fail to provide any sublinear regret guarantee. Furthermore, in the adversarial regime, they typically control only a notion of violation that allows for cancellation, meaning that strictly feasible decisions can compensate for unfeasible ones.

A recent attempt to overcome these challenges is (Stradi et al., 2025c), where the authors study constrained MABs in a regime more general than the fully adversarial one, in which both losses and constraints are drawn from distributions that may vary across rounds.[2] Crucially, however, they assume that the non-stationarity of both losses and constraint distributions is bounded. Under this assumption, they propose a meta-algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *positive* violation—thereby *not* allowing cancellation of violation across rounds—where C measures the maximum amount of non-stationarity in both losses and constraints. Their approach has two main drawbacks. First, the resulting regret and violation bounds become vacuous (*i.e.*, linear in T) in the adversarial-loss regime; in particular, they do not recover the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantees in the mixed setting where losses are adversarial and constraints are stochastic. Second, the meta-algorithm relies on a rather complex corralling technique, *i.e.*, a no-regret master procedure that selects among multiple subroutines instantiated with different values of C. This design makes the algorithm complex, and the analysis technically involved and somewhat intricate. Moreover, it yields a quadratic dependence on the number of constraints. We finally note that the coralling technique of Stradi et al. (2025c) cannot be applied when losses are fully adversarial, since bounding the violations would require a sublinear upper bound on the negative regret incurred by the subroutines, which is clearly unattainable in adversarial loss settings. Indeed, the analysis of Stradi et al. (2025c) relies on the fact that the negative regret of the subroutines is bounded by C, since the losses are non-stationary in the same way as the constraints.

Due to space constraints, we refer to Appendix A for a complete discussion on related works.

In this work, we aim to advance the theoretical understanding of constrained MABs. In particular, we address the following research question:

Is it possible to achieve sublinear regret and sublinear positive constraint violation that degrade only with the degree of non-stationarity in the constraints?

In this paper, we provide an affirmative answer to the question above, as we describe next.

1. In this paper, we say that a quantity is sublinear in T if it is $o(T)$.
2. Indeed, Stradi et al. (2025c) study *constrained Markov decision processes* (CMDPs). Nonetheless, the way unknown constraints are handled is essentially equivalent in constrained MABs and CMDPs.

1.1. Original Contribution

We study the *constrained* MAB problem in which both the losses and the unknown constraints are drawn at each round from distributions that may change arbitrarily over time. We quantify the non-stationarity of the constraints via a corruption level C, which measures the distance between the mean values of the constraint distributions and a fictitious uncorrupted constraint vector. Crucially, our notion of corruption level accounts only for the non-stationarity of the constraints, and *not* for that of the losses. Clearly, when the constraints are stochastic but stationary, $C = 0$, while in the worst case $C = \Theta(T)$ (see Section 2 for further details on the setting). Throughout the paper, we consider different feedback models. Specifically, in the full feedback case, at the end of each round the learner observes the losses and the constraint violation of every possible action. In contrast, under bandit feedback, the learner observes only the loss and the violation of the chosen action.

As a warm-up, in Section 3 we provide a detailed discussion on the technical challenges posed by our setting. Then, we show how simple concentration results for the corrupted constraints yield $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret R_T and *positive* constraint violation V_T when the corruption level C is known.

Full feedback In Section 4, we focus on the full feedback setting. In this case, we propose an algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret and *positive* violation without any prior knowledge of the corruption. Our approach relies on two main components. First, at each round, the algorithm constructs an approximate feasible decision set $\mathcal{X}_t$, by using the *optimism* principle. Since C is unknown, this optimistic approximation does *not* ensure that an optimal strategy is included in the set at every round. To tackle this challenge, we leverage a second component: a no-regret optimization procedure that guarantees small *switching regret on moving decision spaces*. Specifically, we employ online mirror descent with a fixed-share update (Cesa-Bianchi et al., 2012), and we show that, when the number of switches of a dynamic benchmark—allowed to lie in moving decision spaces—is small, the procedure attains sublinear dynamic regret. This property enables us to effectively analyze the performance of our algorithm against a fixed benchmark that may not always belong to the per-round decision space, leading to our final result. Finally, in Section 4.3, we show how to extend this result to the case where only bandit feedback is available on the losses.

Bandit feedback In Section 5, we study the bandit feedback setting. We propose an algorithm that attains $\widetilde{\mathcal{O}}(T^{\max\{1/2,\beta\}} + C)$ *positive* violation and $\widetilde{\mathcal{O}}(T^{\beta} + CT^{1-\beta})$ regret, where β is given as input. Specifically, by setting $\beta = 1/2$ we obtain $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ *positive* violation and $\widetilde{\mathcal{O}}(\sqrt{T}+C\sqrt{T})$ regret. As we discuss in Section 5, under bandit feedback, we cannot rely on switching-regret guarantees to address the main technical challenges. Intuitively, in adversarial-loss settings, one cannot ensure convergence to an optimal strategy; consequently, the support of an optimal strategy is *not* guaranteed to be explored sufficiently. We overcome this issue via a two-phase algorithm. In the first phase, the learner enforces uniform exploration for $\Theta(T^{\beta})$ rounds. In the second phase, it runs online mirror descent (Orabona, 2019) over $\mathcal{X}_t$. The final guarantee follows from the fact that the forced exploration allows us to show that, with high probability, an approximately optimal strategy belongs to $\mathcal{X}_t$ throughout the entire second phase.

Comparison with the state-of-the-art To conclude the section, we provide a brief comparison between the regret and violation bounds derived in this work and the state-of-the-art:

- Our results match those in the *stochastic constraints* literature, including settings with stochastic losses (Efroni et al., 2020), adversarial losses with full feedback (Qiu et al., 2020), and adversarial

Protocol 1 Learner-Environment Interaction

for $t \in [T]$ **do**
  The adversary selects $\mathcal{L}_t$, $\mathcal{G}_{t,i}$ for all $i \in [m]$
  $\boldsymbol{\ell}_t \sim \mathcal{L}_t$ and $\boldsymbol{g}_{t,i} \sim \mathcal{G}_{t,i}$ for all $i \in [m]$ are drawn
  Select strategy $\boldsymbol{x}_t$ and choose $a_t \sim \boldsymbol{x}_t$
  Suffer loss $\ell_t(a_t)$ and violation $g_{t,i}(a_t)$ for all $i \in [m]$
  Observe $\boldsymbol{\ell}_t$ and $\boldsymbol{g}_{t,i}$ for all $i \in [m]$ `// Full Feedback`
  Observe $\ell_t(a_t)$ and $g_{t,i}(a_t)$ for all $i \in [m]$ `// Bandit Feedback`
end

losses with bandit feedback (Stradi et al., 2025a).[3] Specifically, we match their $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation bounds, while additionally providing guarantees when the constraints are adversarial.

- Comparing to the *best-of-both-worlds* results of Bernasconi et al. (2024), we match their guarantees when the constraints are stochastic. When the constraints are (even only mildly) adversarial, they do *not* provide any sublinear regret guarantee with respect to an optimal strategy; specifically, they obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret against a *fraction* of the optimum. The *same* guarantee can be easily recovered by our algorithm, as discussed in Section 3. In contrast, we show that our algorithm achieves sublinear regret whenever the constraints are mildly adversarial. Moreover, under *adversarial* constraints, Bernasconi et al. (2024) do *not* provide guarantees on the *positive* violation, thus allowing for cancellations. Finally, their techniques do *not* extend to adversarial settings with noise, namely when both losses and constraints are drawn from distributions that vary over time.
- Comparing to Stradi et al. (2025c), who provide regret and *positive* violation bounds that degrade with the amount of corruption affecting *both* losses and constraints, we obtain guarantees that degrade only with the degree of non-stationarity of the constraints, and are thus optimal in the setting with adversarial losses and stochastic constraints.

2. Preliminaries

We study *online learning* problems (Cesa-Bianchi and Lugosi, 2006) in which a learner interacts with an unknown environment over T rounds, so as to minimize long-term losses subject to $m \in \mathbb{N}_+$ *unknown* constraints. At each round $t \in [T]$,[4] the learner selects a strategy $\boldsymbol{x}_t \in \Delta_K$ over $K \in \mathbb{N}_+$ arms, where Δ_K is the $(K-1)$-dimensional simplex. Then, they select arm $a_t \sim \boldsymbol{x}_t$, suffering a loss $\ell_t(a_t) \in [0,1]$ and a constraint violation $g_{t,i}(a_t) \in [-1,1]$ for each $i \in [m]$, where non-strictly-positive values stand for satisfaction of the constraint. The loss vector $\boldsymbol{\ell}_t \in [0,1]^K$ is sampled at each round from a distribution $\mathcal{L}_t$ while, for all $i \in [m]$, the constraint vector $\boldsymbol{g}_{t,i} \in [-1,1]^K$ is sampled at each round from a distribution $\mathcal{G}_{t,i}$. $\mathcal{L}_t$ and $\mathcal{G}_{t,i}$ are allowed to change arbitrarily across rounds, *i.e.*, they may be chosen adversarially. We refer to $\bar{\boldsymbol{\ell}}_t \in [0,1]^K$ as the expected value of $\mathcal{L}_t$, and to $\bar{\boldsymbol{g}}_{t,i} \in [-1,1]^K$ as the expected value of $\mathcal{G}_{t,i}$. When *full feedback* (on the losses and/or constraints) is available, the learner observes the full (loss and/or constraint) vectors at the end of each round. When only *bandit feedback* is available, the learner observes the loss and the constraint violation only of the selected arm. In Protocol 1, we provide the learner-environment interaction.

3. We again acknowledge that these works address the more general CMDP setting.
4. In this paper, we denote by $[a \dots b]$ the set of all the natural numbers from $a \in \mathbb{N}$ to $b \in \mathbb{N}$ (both included), while $[b] := [1 \dots b]$ denotes the set of the first $b \in \mathbb{N}$ natural numbers.

Given the impossibility of learning under adversarial constraints (Mannor et al., 2009)—and noting that our setting is a generalization of standard adversarial ones, where no noise is added to adversarial losses and constraints—we aim at results parametrized by the amount of adversariality of the constraints. Formally, we introduce the notion of *(adversarial) corruption* of the constraints defined as $C := \max_{i\in[m]} C_i$ where $C_i := \min_{\boldsymbol{g}_i\in[-1,1]^K} \sum_{t=1}^T \|\bar{\boldsymbol{g}}_{t,i} - \boldsymbol{g}_i\|_1$. For every $i \in [m]$, we let $\boldsymbol{g}_i^\circ \in [-1,1]^K$ be the constraint vector that attains the minimum in the definition of C_i. Intuitively, C_i can be interpreted as a corruption measure on the constraint i. Specifically, suppose there exists a fixed constraint distribution with mean $\boldsymbol{g}_i^\circ$. Then, an adversary is allowed, for C rounds, to arbitrarily alter the mean of this distribution. This modeling choice is primarily motivated by impossibility results for learning under adversarial constraints (Balseiro and Gur, 2019; Bernasconi et al., 2025), where the lower bounds are established through a step change in the constraints, corresponding to the regime $C = \Theta(T)$.

2.1. Performance Metrics

To define the performance metrics used to evaluate our learning algorithms, we need to introduce an *offline* optimization problem. Program (1) defines an (offline) *optimal feasible solution in hindsight*:

$$\text{OPT} := \begin{cases} \min_{\boldsymbol{x}\in\Delta_K} & \sum_{t=1}^T \bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x} \quad \text{s.t.} \\ & \sum_{t=1}^T \bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x} \le 0 \quad \forall i \in [m]. \end{cases} \tag{1}$$

Notice that, differently from standard MAB problems (Orabona, 2019), in their constrained counterpart an optimal solution in hindsight may need to randomize over arms, since the constraints can potentially cut out vertices of the simplex, thus making choosing any arm potentially suboptimal.

Throughout the paper, we make use of the following Slater's like assumption, which is standard in adversarial online learning with unknown constraints (see, *e.g.*, (Immorlica et al., 2022; Castiglioni et al., 2022b; Stradi et al., 2024; Bernasconi et al., 2025)).

Assumption 1 *There exists a strategy $\boldsymbol{x}^\circ \in \Delta_K$ such that $\max_{t\in[T]} \bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}^\circ < 0$ for all $i \in [m]$.*

We also introduce a problem-specific *feasibility parameter* $\rho \in [0,1]$ related to Assumption 1, defined as $\rho := \max_{\boldsymbol{x}\in\Delta_K} \min_{t\in[T]} \min_{i\in[m]} -\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}$. We denote by $\boldsymbol{x}^\diamond \in \Delta_K$ a strategy that attains the value ρ, which we refer to in the following as a *strictly feasible strategy*. Intuitively, ρ represents by how much $\boldsymbol{x}^\diamond$ strictly satisfies the constraints. Assumption 1 is equivalent to assuming that $\rho > 0$.

Now, we introduce the notion of *(cumulative) regret* and *(cumulative) positive constraint violation*, the performance metrics used to evaluate algorithms. The regret over the T rounds is defined as $R_T := \sum_{t=1}^T \ell_t(a_t) - \text{OPT}$. In the following, we denote by $\boldsymbol{x}^*$ a strategy solving Program (1). Thus, $\text{OPT} = \sum_{t=1}^T \bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^*$ and the regret can be written as $R_T := \sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^*$.

The cumulative positive constraint violation over T rounds is $V_T := \max_{i\in[m]} \sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+$, where we let $[\cdot]^+ := \max\{0,\cdot\}$. To be consistent with the definition of OPT, we use the randomized strategy played by the learner instead of the actual arm (and similarly, we take the expectation over the constraint distribution) in the constraint violation definition. Notice that replacing the strategy with the realized arm (and the expectation with the sampled constraint) in the violation definition would lead to linear violation even if the optimal solution $\boldsymbol{x}^*$ is played for T rounds, due to the $[\cdot]^+$ operator. We remark that a sublinear bound on V_T directly implies a bound on $\max_{i\in[m]} \sum_{t\in[T]} g_{t,i}(a_t)$ (where $[\cdot]^+$ is not used) up to a $\widetilde{\mathcal{O}}(\sqrt{T})$ factor, thanks to a straightforward application of the Azuma–Hoeffding inequality.

The goal of the learner is to attain sublinear regret and sublinear positive constraint violation with bounds that degrade gracefully with respect to the constraint corruption term C.

3. Warm-Up: The Trade-Off Between Regret and Violation

We start highlighting the technical challenges of our setting, namely, the trade-off that any algorithm has to face when dealing with an adversarial online learning problem under unknown constraints.

It is well known that adversarial regret minimizers are capable of being no-regret with respect to any strategy that is included in their decision space $\mathcal{X}_t$ at every round $t \in [T]$ (see, *e.g.*, (Jin et al., 2020; Stradi et al., 2025a; Bernasconi et al., 2024)). Thus, the fundamental challenge of our setting is to simultaneously ensure that: (i) an optimal feasible solution is included in the decision space at every round, so that no-regret guarantees can be easily attained; and (ii) the decision spaces are *not* "too large", which would lead to large constraint violation.

When the losses are adversarial, and each constraint $i \in [m]$ is sampled from a fixed distribution $\mathcal{G}_i$ at each round, the most natural idea is to build an optimistic feasible set by employing Hoeffding inequality. Specifically, referring to $\widehat{\boldsymbol{g}}_{t,i}$ as the empirical mean of the observed constraint violation, it can be easily shown that the expected value of $\mathcal{G}_i$ lies in $\widehat{\boldsymbol{g}}_{t,i} \pm \boldsymbol{\xi}_t$ with high probability, where $\boldsymbol{\xi}_t$ is of order $\mathcal{O}(1/\sqrt{t})$ under full feedback, while it is of order $\mathcal{O}(1/\sqrt{N_t(a)})$, where $N_t(a)$ is the number of pulls of arm a, when only bandit feedback is available. Thus, by taking at each round $(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x} \leq 0$ as an optimistic estimated constraint and optimizing over the set of strategies that satisfy it, one can show that an optimal feasible solution $\boldsymbol{x}^*$ is included in $\mathcal{X}_t$ at every round t. Similarly, by definition of $\boldsymbol{\xi}_t$, $\mathcal{X}_t$ concentrates to the true feasible decision space at a rate of $\mathcal{O}(1/\sqrt{t})$, and this allows to show that $\mathcal{X}_t$ is *not* "too large" and, thus, V_T is sublinear.

When both the losses and the constraints are adversarial,[5] and it is thus provably *not* possible to simultaneously attain sublinear regret and sublinear violation, state-of-the-art results (Castiglioni et al., 2022b; Bernasconi et al., 2024) focus on attaining sublinear (*non-positive*) violation and sublinear regret with respect to a $\rho/(1+\rho)$ fraction of the optimal *reward*. To have a high-level intuition on how to get these results, it is sufficient to notice that Slater's condition implies that a $\rho/1+\rho$ convex combination between the strictly feasible strategy $\boldsymbol{x}^\diamond$ and the optimal strategy $\boldsymbol{x}^*$ is feasible at every round and, thus, it is included in any "reasonable" per-round decision space $\mathcal{X}_t$.[6] Hence, building a decision space that simply moves toward feasible strategies, based solely on the observed violations, is sufficient to attain sublinear regret w.r.t. a fraction $\rho/1+\rho$ of the optimal *reward*.

In this work, the goal is to attain sublinear regret. Thus, we cannot simply play strategy that are (approximately) feasible according to observed violations, since we cannot cut out arms that are unfeasible only in some rounds. Indeed, notice that $\boldsymbol{x}^*$ may violate the constraints in many rounds. Furthermore, the use of confidence intervals is *not* sufficient to guarantee that $\boldsymbol{x}^*$ is included in $\mathcal{X}_t$, since the constraints are corrupted by C, making standard confidence intervals ineffective. In the following section, as a warm-up, we show how to deal with this problem when C is known.

3.1. A Simple Approach When C is Known

Throughout this section, we assume that the value of C is *known* to the learner. We show that, in such a case, it is easy to build suitable decision spaces $\mathcal{X}_t$. Specifically, we define an unbiased estimator

5. The same reasoning holds even when only the losses are stochastic.
6. By "reasonable" decision space, here we mean one that does not cut out strategies that are feasible at every round.

of the constraint violation $\widehat{g}_{t,i}(a) := \frac{1}{N_t(a)} \sum_{\tau=1}^{t} \mathbb{I}_\tau(a) g_\tau(a)$ for all $a \in [K], i \in [m], t \in [T]$, where $N_t(a)$ is the number of pulls of arm a up to round t and $\mathbb{I}_\tau(a)$ is the indicator function of a being selected at round $\tau \in [T]$. We refer to $\widehat{\boldsymbol{g}}_{t,i}$ as the vector whose entries are the estimates $\widehat{g}_{t,i}(a)$ for all $a \in [K]$. We remark that, when full feedback is available, we can define $\widehat{\boldsymbol{g}}_{t,i}$ by using t in place of $N_t(a)$ and setting $\mathbb{I}_\tau(a) = 1$ for all $a \in [K]$ and $\tau \in [T]$. Now, we are able to bound the distance between the estimator and the constraint violation in hindsight, as follows.

Lemma 1 *Let $\delta \in (0, 1)$. When full feedback is available, with probability at least $1 - \delta$ it holds:*

$$\max_{i \in [m]} \left| \widehat{g}_{t,i}(a) - \frac{1}{T} \sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \leq 4 \sqrt{\frac{1}{t} \ln \left(\frac{TKm}{\delta} \right)} + \frac{C}{t} + \frac{C}{T} \quad \forall t \in [T], a \in [K].$$

Similarly, when only bandit feedback is available, with probability at least $1 - \delta$ it holds:

$$\max_{i \in [m]} \left| \widehat{g}_{t,i}(a) - \frac{1}{T} \sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \leq 4 \sqrt{\frac{1}{N_t(a)} \ln \left(\frac{TKm}{\delta} \right)} + \frac{C}{N_t(a)} + \frac{C}{T} \quad \forall t \in [T], a \in [K].$$

By Lemma 1, when C is known, one can define $\zeta_t(a) := 4\sqrt{\ln(TKm/\delta)/N_t(a)} + C/N_t(a) + C/T$ for every action $a \in [K]$, so that the constraint violation in hindsight belongs to the interval $\widehat{\boldsymbol{g}}_{t,i} \pm \boldsymbol{\zeta}_t$ with high probability. Thus, a simple instantiation of online mirror descent with implicit exploration (Neu, 2015; Jin et al., 2020) on the per-round decision space $\{\boldsymbol{x} \in \Delta_K : (\boldsymbol{g}_{t,i} - \boldsymbol{\zeta}_t)^\top \boldsymbol{x} \leq 0\}$ attains sublinear regret and *positive* violation of order $\widetilde{\mathcal{O}}(\sqrt{T} + C)$, by a reasoning similar to the one for the stochastic setting with fixed distributions.

In the rest of the paper, we focus on overcoming the challenge that C is *not* known, precluding the possibility of building meaningful confidence intervals that depend on C. In the following, we show a substantially less trivial way to exploit the previous inequalities even when C is unknown.

4. Learning With Full Feedback: Switching Regret on Moving Decision Spaces

In this section, we focus on the *full feedback* setting, in which the entire loss and constraint vectors are observed by the learner at the end of each round. A simple application of Lemma 1 leads to $\max_{i \in [m]} \left| \widehat{g}_{t,i}(a) - \frac{1}{T} \sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \leq 4 \sqrt{\frac{1}{t} \ln \left(\frac{TKm}{\delta} \right)} + \frac{C}{t} + \frac{C}{T}$ for all $t \in [T], a \in [K]$, which holds with probability at least $1 - \delta$. The first component of the right-hand side can be easily computed by a learning algorithm, since all the quantities are known beforehand. Instead, the term $C/t + C/T$ is unknown to the learner. Fortunately, the term concentrates independently of the arms pulled. As we show next, this feature is fundamental for the analysis provided in this section.

4.1. Algorithm

In Algorithm 1, we provide the pseudocode of `ConOMD-FS`.

The algorithm relies on two components. First, it builds a per-round approximate feasible decision space $\mathcal{X}_t$. This is done by computing the empirical mean of the observed violation $\widehat{\boldsymbol{g}}_{t,i}$ for each $i \in [m]$ (Line 6) and by constructing a confidence interval based solely on the only term in Lemma 1 known to the learner, namely $\boldsymbol{\xi}_t$ (Line 7). Then, $\mathcal{X}_t$ is built optimistically, by taking the lower bound on the estimated constraint violation (Line 8). The second component

Algorithm 1 Constrained OMD with Fixed Share (`ConOMD-FS`)

Input: $T \in \mathbb{N}$, $K \in \mathbb{N}$, $\delta \in (0, 1)$

1 Initialize $\boldsymbol{x}_1 \leftarrow \boldsymbol{x}_{\mathcal{U}}$ where $\boldsymbol{x}_{\mathcal{U}}(a) \coloneqq 1/K$ for $a \in [K]$

2 Initialize $\eta \leftarrow \sqrt{\ln(KT)/T}$

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

4 Choose $a_t \sim \boldsymbol{x}_t$

5 Observe loss $\boldsymbol{\ell}_t$ and violation $\boldsymbol{g}_{t,i}$ for all $i \in [m]$

6 Update estimator $\widehat{\boldsymbol{g}}_{t,i}$ for all $i \in [m]$

7 Define $\boldsymbol{\xi}_t$ as $\xi_t(a) \coloneqq 4\sqrt{\frac{1}{t}\ln\left(\frac{TKm}{\delta}\right)}$ for $a \in [K]$

8 Build $\mathcal{X}_t \coloneqq \{\boldsymbol{x} \in \Delta_K : (\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x} \leq 0\}$

9 Compute $\widetilde{\boldsymbol{x}}_{t+1} \leftarrow \arg\min_{\boldsymbol{x} \in \mathcal{X}_t} \boldsymbol{\ell}_t^\top \boldsymbol{x} + \frac{1}{\eta} D(\boldsymbol{x} || \boldsymbol{x}_t)$

10 Select $\boldsymbol{x}_{t+1} \coloneqq (1 - \frac{1}{T})\widetilde{\boldsymbol{x}}_{t+1} + \frac{1}{T}\boldsymbol{x}_{\mathcal{U}}$

11 **end**

is a *regret minimizer* for the losses over the sets $\mathcal{X}_t$. In the following, we show that we cannot employ an arbitrary regret minimizer due to the nature of the sets $\mathcal{X}_t$, which are only approximations of the true feasible decision space and, thus, are *not* guaranteed to contain $\boldsymbol{x}^*$ at every round. We choose *online mirror descent* (OMD) with entropic regularizer (Orabona, 2019) (Line 9), where $D(\boldsymbol{x}_1 || \boldsymbol{x}_2) \coloneqq \sum_{a \in [K]} x_1(a) \ln\left({}^{x_1(a)}/_{x_2(a)}\right) - \sum_{a \in [K]}(x_1(a) - x_2(a))$, and a fixed-share update (Cesa-Bianchi et al., 2012) (Line 10). Intuitively, the strategy selected by OMD is combined with the uniform strategy in order to make the learning dynamic more stable.

4.2. Analysis and Theoretical Guarantees

In this section, we provide the theoretical guarantees of Algorithm 1, in terms of regret and violation.

We start by providing some intuitions about the regret analysis. To do that, we first study the approximate feasible decision space $\mathcal{X}_t$ and its connection to the optimal solution $\boldsymbol{x}^*$. As previously discussed, since the corruption value C is *not* known *a priori*, it is *not* guaranteed that $\boldsymbol{x}^* \in \mathcal{X}_t$ at every round $t \in [T]$. Nonetheless, we can still show the following fundamental results. First, given Assumption 1, it is possible to show that $\mathcal{X}_t$ is never empty. This is a consequence of both the definition of $\boldsymbol{x}^\diamond$ and the optimism employed in Algorithm 1. Specifically, it is possible to show that, with probability at least $1 - \delta$, it holds $(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x}^\diamond \leq -\rho < 0$. We provide a similar result for the optimal solution, that is, $(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x}^* \leq \frac{C}{T} + \frac{C}{t}$, which follows from Lemma 1 with probability at least $1 - \delta$. By combining the previous results, we can always build a t-dependent convex combination between $\boldsymbol{x}^\diamond$ and $\boldsymbol{x}^*$, defined as $\boldsymbol{x}^*_{\alpha_t} \coloneqq (1 - \alpha_t)\boldsymbol{x}^\diamond + \alpha_t \boldsymbol{x}^*$, which is always included in $\mathcal{X}_t$. Indeed, by setting $\alpha_t = \frac{\rho}{\rho + {}^{2C}/_t}$, with probability at least $1 - \delta$:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} \leq -(1 - \alpha_t)\rho + \alpha_t \frac{2C}{t} \leq 0. \tag{2}$$

Thus, on the one hand, Equation (2) shows that $\mathcal{X}_t$ is *not* too far from $\boldsymbol{x}^*$, while, on the other hand, a general adversarial regret minimizer is guaranteed to be no-regret only against benchmarks that belong to $\mathcal{X}_t$ *at every round*, which is *not* the case for $\boldsymbol{x}^*$. We overcome this technical challenge by showing that OMD with an entropic regularizer and a fixed-share update attains sublinear *switching regret* of order $\widetilde{\mathcal{O}}(S\sqrt{T})$ on *moving decision spaces*, that is, sublinear dynamic regret with respect to S*-switch dynamic benchmarks on moving decision spaces*, which are formally defined as follows.

Definition 2 (S-switch dynamic benchmark) *Let $\{\mathcal{X}_t\}_{t=1}^T$ be the sequence of decision spaces available to the learner. Define $S \leq T$ consecutive phases over the T rounds, and let $\mathcal{I}_1, \dots, \mathcal{I}_S$ be the associated partition of $[T]$. We say that $\{\boldsymbol{u}_t\}_{t=1}^T$ is a S-switch dynamic benchmark on moving decision spaces if and only if the following two conditions hold: (i) $\boldsymbol{u}_t = \boldsymbol{u}_{t'}$ for all $j \in [S]$, $t, t' \in \mathcal{I}_j$; and (ii) $\boldsymbol{u}_t \in \mathcal{X}_j$ for all $t \in [T]$, where $\mathcal{X}_j = \bigcap_{t \in \mathcal{I}_j} \mathcal{X}_t$.*

Thus, the regret associated with $\{\boldsymbol{u}_t\}_{t=1}^T$ is defined as $R_T(\{\boldsymbol{u}_t\}_{t=1}^T) \coloneqq \sum_{t=1}^T \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{u}_t\right]$, which extends the well-known notion of switching regret (Cesa-Bianchi et al., 2012) to benchmarks that may belong to different decision spaces at each round.[7]

A key final challenge is to deal with the dependence on the number of phases S. Indeed, we cannot select $S = T$ (and $\boldsymbol{u}_t = \boldsymbol{x}^*_{\alpha_t}$), since it would lead to a superlinear regret bound. Nonetheless, by noticing that $\alpha_t \leq \alpha_{t+1}$ and thus, $\boldsymbol{x}^*_{\alpha_t} \in \mathcal{X}_\tau$ for $\tau \geq t$, we can employ a doubling trick approach. Specifically, we define $S = \log_2(T)$ phases, so that $\underline{t}_1 = 1$ and $\underline{t}_{j+1} = 2\underline{t}_j$ for all $j \in [S-1]$, where $\underline{t}_j \in [T]$ denotes the first round of phase $j \in [S]$. Moreover, we set $\boldsymbol{u}_t = \boldsymbol{x}^*_{\alpha_{\underline{t}_j}}$ for all $j \in [S], t \in \mathcal{I}_j$. This results in a logarithmic error in the final regret bound. In the following, for ease of notation and with a slight abuse of notation, we let $\alpha_j \coloneqq \alpha_{\underline{t}_j}$ for all $j \in [S]$.

We show the final regret bound in the following theorem.

Theorem 3 *Let $\delta \in (0,1)$. With probability at least $1 - 3\delta$, Algorithm 1 attains:*

$$R_T \leq 4\log_2(T)\sqrt{T\ln(KT)} + \frac{2C}{\rho}\log_2(T) + 4\sqrt{T\ln\left(\frac{TK}{\delta}\right)}.$$

Theorem 3 shows that Algorithm 1 attains regret of order $\widetilde{\mathcal{O}}(\sqrt{T} + C)$. Intuitively, the result is proved by employing the bound on the switching regret on moving decision spaces and the doubling trick approach. By letting $\phi : [T] \to [S]$ be a mapping such that $\phi(t) \coloneqq \sum_{j \in [S]} j \cdot \mathbb{I}\{t \in \mathcal{I}_j\}$, the final regret guarantee follows by $\sum_{t \in [T]} \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{\phi(t)}} - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \leq \sum_{t \in [T]} \left(1 - \alpha_{\phi(t)}\right) \leq \frac{2C}{\rho}\log_2(T)$.

We are now ready to provide the violation bound. This is done in the following theorem, where we show that Algorithm 1 effectively attains a positive violation of order $\widetilde{\mathcal{O}}(\sqrt{T} + C)$.

Theorem 4 *Let $\delta \in (0,1)$. With probability at least $1 - \delta$, Algorithm 1 attains:*

$$V_T \leq 2 + 2C + C\ln(T) + 16\sqrt{T\ln\left(\frac{TKm}{\delta}\right)}.$$

Intuitively, Theorem 4 relies on two components. First, similarly to Theorem 3, the fact that $\mathcal{X}_t$ is non-empty at every round, thanks to both optimism and Assumption 1. Thus, Algorithm 1 effectively works inside $\mathcal{X}_t$, up to the fixed-share update. This leads to a negligible (constant) violation term. By noticing that the positive violation in hindsight is far from the uncorrupted vector $\boldsymbol{g}_i^\circ$ of a C_i factor, it is sufficient to bound how the terms in Lemma 1 concentrate. Specifically, $\boldsymbol{\xi}_t$ concentrates as $1/\sqrt{t}$, leading to the $\mathcal{O}(\sqrt{T\ln(TKm/\delta)}$ term in the violation bound, C/t concentrates as $1/t$, leading to $C\ln(T)$ violation, while we pay an additional C term for the concentration of C/T.

7. A similar switching-regret result has been developed by Moreno et al. (2025) in the context of unconstrained MDPs.

4.3. Extension to Bandit Feedback on the Losses

In this section, we show how to extend the previous result to the case where only bandit feedback is available on the losses, *i.e.*, the learner only observes $\ell_t(a_t)$ at each round $t \in [T]$.

From an algorithmic perspective, we simply modify Algorithm 1 in the following way. The loss is estimated by employing an implicit exploration approach (Neu, 2015). Specifically, at each round $t \in [T]$, we build $\widehat{\boldsymbol{\ell}}_t$ such that $\widehat{\ell}_t(a) := \frac{\ell_t(a)}{x_t(a)+\gamma}\mathbb{I}_t(a)$ for all $a \in [K]$, where $\gamma := \eta/2$ is the implicit exploration factor and $\mathbb{I}_t(a)$ is the indicator function that is equal to 1 whenever $a = a_t$. Then, OMD with fixed-share update is used on $\widehat{\boldsymbol{\ell}}_t$ and η is set to $\sqrt{\ln(KT)/KT}$. Due to space constraints, we refer to Algorithm 3 in Appendix D for the complete algorithm.

We provide the regret bound attained by Algorithm 3 in the following theorem.

Theorem 5 *Let $\delta \in (0,1)$. With probability at least $1-6\delta$, Algorithm 3 attains:*

$$R_T \leq K\log_2(T)\ln\left(\frac{\log_2(T)K}{\delta}\right)+11\log_2(T)\ln\left(\frac{K\log_2(T)}{\delta}\right)\sqrt{KT\ln\left(\frac{TK}{\delta}\right)}+\frac{2C}{\rho}\log_2(T).$$

Theorem 5 shows that a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ is still attainable when the learner has bandit feedback on the losses. The analysis of Theorem 5 shares many similarities with the full feedback case. The key difference is that we prove no-switching-regret guarantees on moving decision spaces for OMD with implicit exploration, and we employ those guarantees to attain the final regret bound. Finally, both the bound and the analysis of the violation attained by Algorithm 3 are equivalent to the full feedback (on the losses) case. Thus, it is omitted for simplicity.

5. Learning With Bandit Feedback by Forcing Exploration

In this section, we focus on the *bandit feedback* setting, where the learner observes the loss and the constraint violation only for the chosen arm.

To handle bandit feedback, we require a slightly stronger assumption than Assumption 1. Specifically, we require the existence of a strictly feasible arm, as stated in the following.

Assumption 2 *There exists an arm $a^\circ \in [K]$ such that $\max_{t\in[T]} \bar{g}_{t,i}(a^\circ) < 0$ for all $i \in [m]$.*

We remark that this assumption is not novel in settings with adversarial constraints and bandit feedback. Specifically, it has already been employed in the analysis of the state-of-the-art algorithm for constrained MABs (Bernasconi et al., 2024) and the special case of bandits with knapsack (Immorlica et al., 2022; Castiglioni et al., 2022a). Throughout this section, we define the problem-specific *feasibility parameter* $\rho \in [0,1]$ according to Assumption 2, as follows $\rho := \max_{a\in[K]} \min_{t\in[T]} \min_{i\in[m]} -\bar{g}_{t,i}(a)$. We refer to $a^\diamond \in [K]$ as the are attaining the value ρ.

Similarly to the full feedback setting, we employ Lemma 1, which, in the bandit case, leads to $\max_{i\in[m]} \left|\widehat{g}_{t,i}(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a)\right| \leq 4\sqrt{\frac{1}{N_t(a)}\ln\left(\frac{TKm}{\delta}\right)} + \frac{C}{N_t(a)} + \frac{C}{T}$ for all $t \in [T], a \in [K]$, with probability at least $1-\delta$. Notice that, while the first term can be computed by the learning algorithm, in the bandit feedback setting, the confidence interval does *not* concentrate independently of the selected arms, resulting in the impossibility of applying the techniques of the previous section.

Algorithm 2 Explore and Optimize with Constrained OMD (`ExpOpt-ConOMD`)

Input: $T \in \mathbb{N}$, $K \in \mathbb{N}$, $\delta \in (0,1)$, $\beta \in [0,1]$

1 Initialize $T_0 \leftarrow K\lceil T^\beta \rceil$, $N_0(a) \leftarrow 0$ for all $a \in [K]$
2 Initialize $\boldsymbol{x}_{T_0+1} \leftarrow \boldsymbol{x}_{\mathcal{U}}$ where $\boldsymbol{x}_{\mathcal{U}}(a) \coloneqq 1/K$ for all $a \in [K]$
3 Initialize $\eta \leftarrow \sqrt{\ln(KT)/KT}, \gamma \leftarrow \eta/2$
4 **for** $a \in [K]$ **do**
5 Choose $a_t = a$ for $\lceil T^\beta \rceil$ rounds `// Exploration Phase`
6 Observe loss $\ell_t(a_t)$ and constraint violation $g_{t,i}(a_t)$ for all $i \in [m]$
7 Update counter $N_t(a_t)$ and empirical estimator $\widehat{g}_{t,i}(a_t)$ for all $i \in [m]$
8 **end**
9 **for** $t = T_0 + 1, \dots, T$ **do**
10 Choose $a_t \sim \boldsymbol{x}_t$ `// Optimization Phase`
11 Observe loss $\ell_t(a_t)$ and constraint violation $g_{t,i}(a_t)$ for all $i \in [m]$
12 Update counter $N_t(a_t)$ and empirical estimator $\widehat{g}_{t,i}(a_t)$ for all $i \in [m]$
13 Define $\boldsymbol{\xi}_t$ as $\xi_t(a) \coloneqq 4\sqrt{\frac{1}{N_t(a)} \ln\left(\frac{TKm}{\delta}\right)}$ for all $a \in [K]$
14 Build $\mathcal{X}_t \coloneqq \{\boldsymbol{x} \in \Delta_K : (\boldsymbol{g}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x} \leq 0\}$
15 Build $\widehat{\boldsymbol{\ell}}_t$ such that $\widehat{\ell}_t(a) \coloneqq \frac{\ell_t(a)}{x_t(a)+\gamma}\mathbb{I}_t(a)$ for all $a \in [K]$
16 Compute $\boldsymbol{x}_{t+1} \leftarrow \arg\min_{\boldsymbol{x} \in \mathcal{X}_t} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{x} + \frac{1}{\eta} D(\boldsymbol{x} || \boldsymbol{x}_t)$
17 **end**

5.1. Algorithm

In Algorithm 2, we provide the pseudocode of `ExpOpt-ConOMD`.

The algorithm splits the learning dynamic into two phases. In the first phase, which we call *exploration phase*, the learner uniformly chooses all the actions (Lines 4–8). Intuitively, this phase allows the algorithm to get a good estimate of the true feasible decision space. Notice that the length of the exploration T_0 is given as input through the parameter β. In the second phase, which we call *optimization phase*, the algorithm employs OMD with implicit exploration over the approximate feasible sets $\mathcal{X}_t$ (Lines 9–17). The following remarks are in order. First, $\mathcal{X}_t$ is estimated given the empirical mean and the confidence intervals computed under bandit feedback. Specifically, the counters $N_t(a)$ for all actions $a \in [K]$ are employed to compute both $\widehat{\boldsymbol{g}}_{t,i}$ and $\boldsymbol{\xi}_t$. Second, Algorithm 2 does *not* employ any fixed-share update. Indeed, as we show next, we do *not* require the optimization procedure to attain the sublinear switching regret property, since, when only bandit feedback is available, this property cannot be employed to bound the distance between the optimum and the best strategy inside $\mathcal{X}_t$. Thus, the fixed-share update is *not* helpful in this setting.

5.2. Analysis and Theoretical Guarantees

In this section, we provide the theoretical guarantees of Algorithm 2 in terms of regret and violation.

We start by providing some intuitions about the regret analysis. First, notice that the regret in the exploration phase can be easily bounded by $K(T^\beta + 1)$. Thus, we can simply focus on the regret in the second phase. By letting $\boldsymbol{x}^\diamond \in \Delta_K$ be the strategy that chooses arm $a^\diamond$ deterministically and $\boldsymbol{x}^*_{\alpha_t} \coloneqq (1-\alpha_t)\boldsymbol{x}^\diamond + \alpha_t \boldsymbol{x}^*$, it is possible to show that, with probability at least $1-\delta$:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} \leq -(1-\alpha_t)\rho + \alpha_t \cdot 2C \sum_{a \in [K]} \frac{x^*(a)}{N_t(a)}, \tag{3}$$

by following a reasoning similar to the one used for the full feedback case. Equation (3) highlights the intrinsic difficulty of our setting when only bandit feedback is available. Specifically, to recover the same regret rate as in the full feedback case, it is necessary to make sure that the mismatch between the optimal strategy $\boldsymbol{x}^*$ and the pulls performed by the algorithm concentrates. Intuitively, this *not* possible, since, when the losses are allowed to arbitrarily change across rounds, no convergence guarantees to the optimal strategy can be attained. We tackle this challenge by introducing the exploration phase, which forces an upper bound on such a mismatch. Specifically:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} \leq -(1-\alpha_t)\rho + \alpha_t \frac{2C}{T^\beta},$$

for all $t > T_0$. This in turn implies that the convex combination $\boldsymbol{x}^*_{\alpha_t}$ satisfies $(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x}^*_{\alpha_t} \leq 0$ and $\boldsymbol{x}^*_{\alpha_{T_0}} \in \mathcal{X}_t$ for all $t > T_0$, by setting $\alpha_t = \alpha_{T_0} := \frac{\rho}{\rho + 2C/T^\beta}$ for all $t > T_0$. As a result, it is possible to employ the no-regret guarantees of OMD against the comparator $\boldsymbol{x}^*_{\alpha_{T_0}}$, which is included in the decision space at every round of the second phase. The final result follows by bounding the distance between the optimal strategy $\boldsymbol{x}^*$ and $\boldsymbol{x}^*_{\alpha_{T_0}}$.

We provide the regret bound attained by Algorithm 2 in the following theorem.

Theorem 6 *Let $\delta \in (0,1)$. With probability at least $1 - 6\delta$, Algorithm 2 attains:*

$$R_T \leq K(T^\beta + 1) + 3\sqrt{KT\ln(KT)} + K\ln\left(\frac{K}{\delta}\right) + 5\sqrt{T\ln\left(\frac{TK}{\delta}\right)} + \sqrt{KT}\ln\left(\frac{K}{\delta}\right) + \frac{2C}{\rho}T^{1-\beta}.$$

Theorem 6 shows that our algorithm attains a regret of order $\widetilde{\mathcal{O}}(\max\{\sqrt{T}, T^\beta\} + CT^{1-\beta})$, which clearly highlights the tension between the exploration and the optimization phases. Specifically, when the exploration length is large, the corruption can be mitigated by a precise estimation of the decision space; nonetheless, the additional exploration is paid in the first component of the regret bound, which encompasses the number of rounds the algorithm is not properly minimizing the loss.

We are now ready to provide the violation bound attained by our algorithm.

Theorem 7 *Let $\delta \in (0,1)$. With probability at least $1 - 3\delta$, Algorithm 2 attains:*

$$V_T \leq 1 + K(T^\beta + 1) + C + KC + KC\ln(T) + 30\sqrt{KT\ln\left(\frac{TKm}{\delta}\right)}.$$

Theorem 7 shows that the violation does *not* get any benefit from exploration. Indeed, optimizing over $\mathcal{X}_t$ is sufficient to get the desired $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ bound, while the exploration phase only adds the $K(T^\beta + 1)$ term in the final result. The violation bound is proved similarly to Theorem 4, with the main exception that the confidence intervals concentrate with respect to the played action only.

We remark that by setting $\beta = 1/2$—thus exploring for $K\sqrt{T}$ rounds—we get, with high probability, the following regret and violation bounds $R_T \leq \widetilde{\mathcal{O}}(\sqrt{T} + C\sqrt{T})$, $V_T \leq \widetilde{\mathcal{O}}(\sqrt{T} + C)$.

6. Open Problem: Towards the Lower Bound For the Bandit Case

In this final section, we include a discussion on the technical challenges that prevented us from formally proving the lower bound for the *bandit* feedback case.

First, we state the result that we conjecture may be a valid lower bound for our setting.

Conjecture 8 *There exist two instances of the constrained MAB problem with bandit feedback, where $C = \Theta(\sqrt{T})$, such that any algorithm attaining $\mathcal{O}(\sqrt{T})$ positive violation in one of the instances must necessarily incur $\Omega(T^{\frac{1}{2}+\alpha})$ regret in the other instance, for a constant $\alpha > 0$.*

The following considerations are in order to find proper instances. On the one hand, we recall that, if we assume that the losses have bounded and small adversariality, *i.e.*, we define the corruption C as the maximum between the corruption of the losses C^ℓ and the corruption of the constraints C^g, then $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ regret and violation can be attained by the algorithm of Stradi et al. (2025c). Thus, at least in one of the instances, we need to enforce $C^\ell = \omega(\sqrt{T})$. To have an intuition on this aspect, notice that the last term of Equation (3) can be easily controlled in a setting with fully stochastic losses, where it is possible to converge to the optimal strategy. Similarly, when C^ℓ is small, it is still possible to employ uniform exploration techniques (*i.e.*, uniformly explore with small probability), to reduce the impact of C^g in the regret bound. On the other hand, our conjecture prescribes a corruption to be of order $\Theta(\sqrt{T})$; thus, the constraint distributions of one of the instances can be corrupted for $\Theta(\sqrt{T})$ rounds. This leads to our *first challenge*, that is, the instances are in a hybrid setting between fully stochastic and adversarial ones. This prevented us from employing standard arguments such as Pinsker's inequality and KL-decomposition, which are tailored for stochastic settings, to relate the instances. Similarly, the same reasoning holds for techniques tailored for fully adversarial lower bounds—generally used for impossibility results—as we need to show that the instances are hard to distinguish due to both the noise and the corruption.

As a second aspect, notice that, when full feedback is available to the learner, Algorithm 1 shows that $\widetilde{\mathcal{O}}(\sqrt{T}+C)$ regret and violation can be attained. Thus, bandit feedback jointly with adversarial losses must play a crucial role in the lower bound. This leads to our *second challenge*, that is, we need to effectively relate the corruption in one of the instances to the strategy played by the learner. In this way, we conjecture that it is possible to show that the corruption *perceived* by the learner $\widetilde{C}$ is *strictly larger* than the expected corruption injected in the instances by the opponent C, leading to the desired lower bound, after carefully noticing that the regret must scale at least as the perceived corruption. A bit more formally, we define the perceived corruption on an arm as the corruption observed on the arm normalized by the number of pulls. Intuitively, this is the quantity that really affects the estimation and the regret. For instance, suppose that we employ two instances: the first one is fully stochastic, while we inject the corruption into both losses and constraints for one action $\bar{a}$ of the second instance. We select the corruption on the constraint at time t as $P_t^g = f(x_t(\bar{a})) = x_t(\bar{a})$, where $x_t(\bar{a})$ is the probability of playing $\bar{a}$. Thus, the following chain of inequalities relates the perceived corruption $\widetilde{C}$ to the corruption C:

$$\widetilde{C} \approx T\frac{\sum_{t\in[T]} x_t(\bar{a})P_t^g}{\sum_{t\in[T]} x_t(\bar{a})} = T\frac{\sum_{t\in[T]} x_t^2(\bar{a})}{\sum_{t\in[T]} x_t(\bar{a})} \geq T\frac{\left(\sum_{t\in[T]} x_t(\bar{a})\right)^2}{T\sum_{t\in[T]} x_t(\bar{a})} = \sum_{t\in[T]} x_t(\bar{a}) = C.$$

The inequality holds *strictly*—up to Azuma inequality concentration terms—whenever the action is *not* chosen uniformly over T, which could be "enforced" by the adversariality of the losses.

Acknowledgments

This publication was funded with the contribution of Ministero dell'Università e della ricerca pursuant to D.D. n. 18010 of 12 November 2025 – BANDO FIS 3. Project FIS-2024-05736 (Starting

Grant), title: "Towards a trustworthy strategic use of data in machine learning pipelines " (STRATDATA). CUP: D53C25002380001

References

Shubhada Agrawal, Timothée Mathieu, Debabrota Basu, and Odalric-Ambrym Maillard. Crimed: Lower and upper bounds on regret for bandits with unbounded stochastic corruption. In *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 74–124. PMLR, 2024. URL https://proceedings.mlr.press/v237/agrawal24a.html.

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

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

Santiago R Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vxM49M5B4s.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 991–999. PMLR, 2021. URL https://proceedings.mlr.press/v130/bogunovic21a.html.

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*, pages 7396–7404, 2023.

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

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.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1562–1578. PMLR, 2019. URL https://proceedings.mlr.press/v99/gupta19a.html.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *J. ACM*, 69(6), November 2022. ISSN 0004-5411. doi: 10.1145/3557045. URL https://doi.org/10.1145/3557045.

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 2020. URL https://proceedings.mlr.press/v119/jin20c.html.

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

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

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

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 114–122, 2018. doi: 10.1145/3188745.3188918. URL https://doi.org/10.1145/3188745.3188918.

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.

Timothée Mathieu, Debabrota Basu, and Odalric-Ambrym Maillard. Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithms. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=oGIR0ic3jU.

Bianca Marin Moreno, Khaled Eldowa, Pierre Gaillard, Margaux Brégère, and Nadia Oudjane. Online episodic convex reinforcement learning. In *International Conference on Machine Learning*, pages 44775–44824. PMLR, 2025.

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

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.

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.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs: Handling stochastic and adversarial constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46692–46721. PMLR, 2024.

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

Francesco Emanuele Stradi, 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*, 2025b.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Taming adversarial constraints in CMDPs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c. URL https://openreview.net/forum?id=sT9nd1WQ76.

Xuchuang Wang, Maoli Liu, Jinhang Zuo, Xutong Liu, John C. S. Lui, and Mohammad Hajiesmaili. Stochastic bandits robust to adversarial attacks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vOFx8HDcvF. arXiv:2408.08859.

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 *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3274–3307. PMLR, 2022b.

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.

Lin Yang, Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John C. S. Lui, and Wing Shing Wong. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19943–19952. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/e655c7716a4b3ea67f48c6322fc42ed6-Abstract.html.

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.

Contents

Appendix A. Related Works

In this section, we provide a summary of the literature which is mainly related to our work. We first discuss the constrained online learning literature and then we survey the main results in the corruption robust online learning one.

Online learning under unknown constraints Online leaning with *unknown* constraints has been recently explored (see, *e.g.*, (Mannor et al., 2009; Liakopoulos et al., 2019; Pacchiano et al., 2021)). In such a setting, a well known impossibility result from (Mannor et al., 2009) prevents any algorithm from attaining both sublinear regret and sublinear violation when the constraints are adversarial and the optimal solution is feasible in hindsight, on average. Thus, the literature mainly focused on stochastic constrained setting (Chen et al., 2022; Pacchiano et al., 2021; Gangrade et al., 2024). Differently, some works focus on constrained online convex optimization settings (see, *e.g.*, (Mahdavi et al., 2012; Jenatton et al., 2016; Yu et al., 2017)). Notice that, even when full feedback is available non both losses and constraints, these works are not applicable to our setting for the two following reasons. First, we employ as a benchmark the optimum which satisfies the constraints on average, while constrained online convex optimization focus on benchmarks satisfying the constraints at each round. Second, they are not tailored to work in adversarial settings with noise. On the other hand, more recent contributions aim to provide best-of-both-worlds guarantees, allowing constraints to be either stochastic or adversarial (Castiglioni et al., 2022a,b; Bernasconi et al., 2024). These works typically guarantee $\widetilde{\mathcal{O}}(\sqrt{T})$ regret and violation in the stochastic regime, while in the adversarial regime they ensure sublinear violation together with sublinear α-regret, *i.e.*, regret measured against a fraction of the optimal reward.

Online learning in constrained Markov decision processes Our paper is strongly related with the constrained Markov decision processes (CMDPs) (Altman, 1999) literature, which we highlight in the following. Wei et al. (2018) study episodic CMDPs with known transitions, full-information feedback, adversarial losses, and stochastic constraints. Their algorithm guarantees $\tilde{\mathcal{O}}(\sqrt{T})$ upper bounds for both regret and constraint violations. Zheng and Ratliff (2020) consider stochastic losses and constraints with known transition dynamics under bandit feedback. They obtain a regret bound of $\tilde{\mathcal{O}}(T^{3/4})$ and ensure that cumulative constraint violations stay below a prescribed threshold with high probability. Bai et al. (2020) propose the first method achieving sublinear regret when transition probabilities are unknown, under the assumptions of deterministic rewards and structured stochastic constraints. Efroni et al. (2020) investigate unknown stochastic transitions, rewards, and constraints in the bandit setting. They design two algorithms that achieve sublinear regret and sublinear constraint violations by carefully balancing exploration and exploitation. Qiu et al. (2020) develop a primal-dual strategy inspired by *optimism in the face of uncertainty*. They show that, for episodic CMDPs with adversarial losses and stochastic constraints under full-information feedback, the approach attains sublinear regret and sublinear constraint violations. Liu et al. (2021) analyze stochastic rewards and constraints with sub-Gaussian noise, proving $\widetilde{O}(\sqrt{T})$ regret and zero violations when a strictly safe policy exists and is known. When such a policy is not known *a priori*, the algorithm guarantees bounded violations. Ding et al. (2021) introduce a primal-dual, no-regret policy optimization procedure for CMDPs with stochastic rewards and constraints. Wei et al. (2022b) present a model-free, simulator-free RL algorithm for CMDPs, obtaining $\widetilde{O}(T^{4/5})$ regret with zero constraint violations, provided that the number of episodes increases exponentially in $1/\rho$. Ding and Lavaei (2023), and Stradi et al. (2025c) study non-stationary rewards and con-

straints under bounded-variation assumptions. Notice that, both the aforementioned works do not attain sublinear regret when the rewards are fully adversarial. Stradi et al. (2025a) consider adversarial losses, stochastic constraints, and partial feedback, establishing sublinear regret together with sublinear positive constraint violations. Stradi et al. (2024) propose the first *best-of-both-worlds* algorithm for CMDPs with *full feedback* on rewards and constraints, and Stradi et al. (2025b) extend the guarantee to the *bandit* feedback scenario.

Corruption-robust online learning A related line studies *corruption-robust unconstrained* online learning, where the observed feedback is perturbed by adversarial or stochastic corruptions (e.g., budgeted or contamination-based), and the goal is to obtain regret bounds that explicitly scale with an appropriate corruption measure. In stochastic multi-armed bandits with adversarially corrupted rewards, Lykouris et al. (2018) obtain regret bounds of order $\widetilde{\mathcal{O}}(KC\sum_{a\neq a^*} 1/\Delta_a)$, where Δ_a is the sub-optimality gap of arm $a \in [K]$. Gupta et al. (2019) significantly sharpen this dependence getting an *additive* dependence on the corruption term, namely, $\widetilde{\mathcal{O}}(KC + \sum_{a\neq a^*} 1/\Delta_a)$. A complementary *attacker* model, where corruptions are chosen *after* observing the learner's action, is studied by Yang et al. (2020), who establish regret lower bounds and separations between budget-aware and budget-agnostic strategies; see also recent developments for stochastic bandits under attacks in Wang et al. (2025). Extensions beyond standard multi-armed bandits have been considered as well, *e.g.*, stochastic linear bandits under adversarial attacks (Bogunovic et al., 2021). More recently, stochastic contamination models allowing heavy-tailed and even unbounded corruptions have been analyzed, yielding instance-dependent lower bounds and asymptotically optimal algorithms (see, *e.g.*, Mathieu et al., 2024; Agrawal et al., 2024).

In episodic reinforcement learning, corruption-robust guarantees for MDPs with corrupted rewards and/or transitions under bandit feedback are developed in Lykouris et al. (2021) and subsequently improved by Chen et al. (2021); Wei et al. (2022a). Jin et al. (2024) study adversarial MDPs with corrupted (non-stationary) transitions providing $\widetilde{\mathcal{O}}(\sqrt{T} + C)$ regret guarantees. Notice that, while the above contributions relate regret to corruption/non-stationarity measures, they do not address constraints satisfaction and thus do not capture the dual objective of simultaneously controlling regret and (positive) constraint violation.

Appendix B. Omitted Proofs of Section 3

Lemma 9 *It holds:*

$$\max_{i\in[m]} \sum_{a\in[K]} \left| g_i^\circ(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \leq \frac{C}{T}.$$

Proof By definition of the uncorrupted vector, it holds:

$$\begin{aligned}
\max_{i\in[m]} \sum_{a\in[K]} \left| g_i^\circ(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| &= \max_{i\in[m]} \sum_{a\in[K]} \left| \frac{1}{T}\sum_{t=1}^{T} g_i^\circ(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \\
&= \max_{i\in[m]} \frac{1}{T} \sum_{a\in[K]} \left| \sum_{t=1}^{T} g_i^\circ(a) - \sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \\
&\leq \max_{i\in[m]} \frac{1}{T} \sum_{t=1}^{T} \sum_{a\in[K]} |g_i^\circ(a) - \bar{g}_{t,i}(a)|
\end{aligned}$$

$$= \max_{i\in[m]} \frac{C_i}{T}$$
$$= \frac{C}{T}.$$

This concludes the proof. ■

Lemma 10 *Let $\delta \in (0,1)$. When full feedback is available, it holds, with probability at least $1-\delta$:*

$$\max_{i\in[m]} |\widehat{g}_{t,i}(a) - g_i^\circ(a)| \leq 4\sqrt{\frac{1}{t}\ln\left(\frac{TKm}{\delta}\right)} + \frac{C}{t} \quad \forall t \in [T], a \in [K].$$

Similarly, when only bandit feedback is available, it holds, with probability at least $1-\delta$:

$$\max_{i\in[m]} |\widehat{g}_{t,i}(a) - g_i^\circ(a)| \leq 4\sqrt{\frac{1}{N_t(a)}\ln\left(\frac{TKm}{\delta}\right)} + \frac{C}{N_t(a)} \quad \forall t \in [T], a \in [K].$$

Proof We prove the results for the bandit feedback case. The same reasoning holds for the full feedback case, replacing $N_t(a)$ with t for all $a \in [K]$. Indeed, applying the triangle inequality, it holds:

$$\begin{aligned}\max_{i\in[m]} |\widehat{g}_{t,i}(a) - g_i^\circ(a)| &= \max_{i\in[m]} \left| \frac{1}{N_t(a)} \sum_{\tau\in[t]} g_{\tau,i}(a)\mathbb{I}_\tau(a) \pm \frac{1}{N_t(a)} \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) - g_i^\circ(a) \right| \\ &\leq \max_{i\in[m]} \left| \frac{1}{N_t(a)} \sum_{\tau\in[t]} g_{\tau,i}(a)\mathbb{I}_\tau(a) - \frac{1}{N_t(a)} \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) \right| \\ &\quad + \max_{i\in[m]} \left| \frac{1}{N_t(a)} \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) - g_i^\circ(a) \right|.\end{aligned}$$

We start bounding the first term. Let $\delta \in (0,1)$, we employ the Azuma inequality to get, with probability at least $1-\delta$:

$$\left| \sum_{\tau\in[t]} g_{\tau,i}(a)\mathbb{I}_\tau(a) - \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) \right| \leq 4\sqrt{N_t(a)\ln\left(\frac{TKm}{\delta}\right)},$$

which holds for all $t \in [T], a \in [K], i \in [m]$, by union bound. Thus, we can conclude that:

$$\max_{i\in[m]} \left| \frac{1}{N_t(a)} \sum_{\tau\in[t]} g_{\tau,i}(a)\mathbb{I}_\tau(a) - \frac{1}{N_t(a)} \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) \right| = 4\sqrt{\frac{1}{N_t(a)}\ln\left(\frac{TKm}{\delta}\right)}.$$

To bound the second term, we proceed as follows:

$$\max_{i\in[m]} \left| \frac{1}{N_t(a)} \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) - g_i^\circ(a) \right| = \max_{i\in[m]} \left| \frac{1}{N_t(a)} \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) - \frac{1}{N_t(a)} \sum_{\tau\in[t]} g_i^\circ(a)\mathbb{I}_\tau(a) \right|$$

$$
\begin{aligned}
&= \max_{i\in[m]} \frac{1}{N_t(a)} \left| \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a) - \sum_{\tau\in[t]} g_i^\circ(a)\mathbb{I}_\tau(a) \right| \\
&\leq \max_{i\in[m]} \frac{1}{N_t(a)} \sum_{\tau\in[t]} \mathbb{I}_t(a) \left| \bar{g}_{\tau,i}(a) - g_i^\circ(a) \right| \\
&\leq \max_{i\in[m]} \frac{C_i}{N_t(a)} \\
&= \frac{C}{N_t(a)}.
\end{aligned}
$$

Combining the previous results concludes the proof. ■

Lemma 11 *Let $\delta \in (0,1)$. When full feedback is available, with probability at least $1-\delta$ it holds:*

$$
\max_{i\in[m]} \left| \widehat{g}_{t,i}(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \leq 4\sqrt{\frac{1}{t}\ln\left(\frac{TKm}{\delta}\right)} + \frac{C}{t} + \frac{C}{T} \quad \forall t \in [T], a \in [K].
$$

Similarly, when only bandit feedback is available, with probability at least $1-\delta$ it holds:

$$
\max_{i\in[m]} \left| \widehat{g}_{t,i}(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \leq 4\sqrt{\frac{1}{N_t(a)}\ln\left(\frac{TKm}{\delta}\right)} + \frac{C}{N_t(a)} + \frac{C}{T} \quad \forall t \in [T], a \in [K].
$$

Proof Employing the triangle inequality, it holds:

$$
\begin{aligned}
\max_{i\in[m]} \left| \widehat{g}_{t,i}(a) - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| &= \max_{i\in[m]} \left| \widehat{g}_{t,i}(a) \pm g_i^\circ - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right| \\
&\leq \max_{i\in[m]} |\widehat{g}_{t,i}(a) - g_i^\circ| + \max_{i\in[m]} \left| g_i^\circ - \frac{1}{T}\sum_{t=1}^{T} \bar{g}_{t,i}(a) \right|.
\end{aligned}
$$

Applying Lemma 9 to bound the second term and Lemma 10 to bound the first term gives the result. ■

Appendix C. Omitted Proofs of Section 4

C.1. Preliminary Results

Lemma 12 *Let $\delta \in (0,1)$. With probability at least $1-\delta$, $\mathcal{X}_t$ is not empty at each round $t \in [T]$.*

Proof To prove the result, we show that any strategy $\bar{\boldsymbol{x}}$ so that $\bar{\boldsymbol{g}}_{t,i}^\top \bar{\boldsymbol{x}} \leq 0$ for all $i \in [m], t \in [T]$–notice that $\bar{\boldsymbol{x}}$ exists thanks to Assumption 1–is included with high probability in $\mathcal{X}_t$ at each round $t \in [T]$.

Let $\delta \in (0,1)$. Similarly to what was done in Lemma 10, we employ the Azuma inequality to get, with probability at least $1-\delta$:

$$\left|\sum_{\tau\in[t]} g_{\tau,i}(a) - \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\right| \leq 4\sqrt{t\ln\left(\frac{TKm}{\delta}\right)},$$

which holds for all $t \in [T], a \in [K], i \in [m]$, by union bound. Thus, we get:

$$\left|\frac{1}{t}\sum_{\tau\in[t]} g_{\tau,i}(a) - \frac{1}{t}\sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\right| = 4\sqrt{\frac{1}{t}\ln\left(\frac{TKm}{\delta}\right)} \quad \forall a \in [K], i \in [m], t \in [T].$$

From that, we get, with probability at least $1-\delta$:

$$\begin{aligned}(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \bar{\boldsymbol{x}} &= \left(\frac{1}{t}\sum_{\tau\in[t]} \boldsymbol{g}_{\tau,i} - \boldsymbol{\xi}_t\right)^\top \bar{\boldsymbol{x}} \\ &\leq \left(\frac{1}{t}\sum_{\tau\in[t]} \bar{\boldsymbol{g}}_{\tau,i}\right)^\top \bar{\boldsymbol{x}} \\ &\leq 0,\end{aligned}$$

where the last step holds by definition of $\bar{\boldsymbol{x}}$. This concludes the proof. ■

C.2. Regret

Lemma 13 *Let $\boldsymbol{x}^*_{\phi(1)}, \dots, \boldsymbol{x}^*_{\phi(T)}$ be a S-switch dynamic benchmark as defined in Definition 2. Thus, OMD with the following fixed share update:*

$$\widetilde{\boldsymbol{x}}_{t+1} \leftarrow \arg\min_{\boldsymbol{x}\in\mathcal{X}_t} \boldsymbol{\ell}_t^\top \boldsymbol{x}_t + \frac{1}{\eta} D(\boldsymbol{x}||\boldsymbol{x}_t)\ ;\ \boldsymbol{x}_{t+1} \coloneqq \left(1-\frac{1}{T}\right)\widetilde{\boldsymbol{x}}_{t+1} + \frac{1}{T}\boldsymbol{x}_{\mathcal{U}},$$

attains:

$$R_T(\{\boldsymbol{x}^*_{\phi(t)}\}_{t=1}^T) \leq \frac{2S}{\eta} + \frac{S\ln(KT)}{\eta} + \eta ST.$$

Proof To prove the result, we fix a phase $j \in [S]$, so that the baseline $\boldsymbol{x}^*_j$ used in the regret associated with that phase is fixed.

Thus, we recall the definition of the Bregman divergence

$$D(\boldsymbol{x}_1||\boldsymbol{x}_2) \coloneqq \sum_{a\in[K]} x_1(a)\ln\left(\frac{x_1(a)}{x_2(a)}\right) - \sum_{a\in[K]}(x_1(a)-x_2(a)),$$

which is equivalent to $D(\boldsymbol{x}_1||\boldsymbol{x}_2) \coloneqq \sum_{a\in[K]} x_1(a)\ln\left({}^{x_1(a)}/_{x_2(a)}\right)$ whenever $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are in the simplex and we define $\bar{\boldsymbol{x}}_{t+1}$ the vector whose components are $\bar{x}_{t+1}(a) = x_t(a)e^{-\eta\ell_t(a)}$ so that

$$\widetilde{\boldsymbol{x}}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}_t} D(\boldsymbol{x}||\bar{\boldsymbol{x}}_{t+1}).$$

We proceed by bounding the instantaneous regret attained by the algorithm at $t \in [T]$ s.t. $\phi(t+1) = j$ as:

$$
\begin{aligned}
&D(\boldsymbol{x}_j^* \| \boldsymbol{x}_t) - D(\boldsymbol{x}_j^* \| \boldsymbol{x}_{t+1}) + D(\boldsymbol{x}_t \| \bar{\boldsymbol{x}}_{t+1}) \\
&= \sum_{a \in [K]} \left[x_j^*(a) \ln \left(\frac{x_j^*(a)}{x_t(a)} \right) - x_j^*(a) + x_t(a) \right] \\
&\quad - \sum_{a \in [K]} \left[x_j^*(a) \ln \left(\frac{x_j^*(a)}{x_{t+1}(a)} \right) - x_j^*(a) + x_{t+1}(a) \right] \\
&\quad + \sum_{a \in [K]} \left[x_t(a) \ln \left(\frac{x_t(a)}{x_t(a) e^{-\eta \ell_t(a)}} \right) - x_t(a) + x_t(a) e^{-\eta \ell_t(a)} \right] \\
&= \sum_{a \in [K]} x_j^*(a) \ln \left(\frac{x_{t+1}(a)}{x_t(a)} \right) + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} \\
&= \sum_{a \in [K]} x_j^*(a) \ln \left(\frac{(1 - 1/T) \widetilde{x}_{t+1}(a) + \frac{1}{T} x_{\mathcal{U}}(a)}{(1 - 1/T) \widetilde{x}_t(a) + \frac{1}{T} x_{\mathcal{U}}(a)} \right) + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} \\
&\geq \sum_{a \in [K]} x_j^*(a) \ln \left(\frac{(1 - 1/T) \widetilde{x}_{t+1}(a)}{(1 - 1/T) \widetilde{x}_t(a) + \frac{1}{T} x_{\mathcal{U}}(a)} \right) + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} \\
&= \sum_{a \in [K]} x_j^*(a) \ln \left(\frac{\widetilde{x}_{t+1}(a)}{(1 - 1/T) \widetilde{x}_t(a) + \frac{1}{T} x_{\mathcal{U}}(a)} \right) - \ln \left(\frac{1}{1 - \frac{1}{T}} \right) + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} \\
&= D(\boldsymbol{x}_j^* \| \boldsymbol{x}_t) - D(\boldsymbol{x}_j^* \| \widetilde{\boldsymbol{x}}_{t+1}) + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} - \ln \left(\frac{1}{1 - \frac{1}{T}} \right) \\
&\geq D(\boldsymbol{x}_j^* \| \boldsymbol{x}_t) - D(\boldsymbol{x}_j^* \| \bar{\boldsymbol{x}}_{t+1}) + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} - \ln \left(\frac{1}{1 - \frac{1}{T}} \right) \\
&= \sum_{a \in [K]} x_j^*(a) \ln \left(\frac{x_j^*(a)}{x_t(a)} \right) - \sum_{a \in [K]} x_j^*(a) \ln \left(\frac{x_j^*(a)}{x_t(a) e^{-\eta \ell_t(a)}} \right) + 1 - \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} \\
&\quad + \eta \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - 1 + \sum_{a \in [K]} x_t(a) e^{-\eta \ell_t(a)} - \ln \left(\frac{1}{1 - \frac{1}{T}} \right) \\
&= \eta \boldsymbol{\ell}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_j^*) - \ln \left(\frac{1}{1 - \frac{1}{T}} \right),
\end{aligned}
$$

where we used that $D(\boldsymbol{x}_j^* || \bar{\boldsymbol{x}}_{t+1}) \geq D(\boldsymbol{x}_j^* || \widetilde{\boldsymbol{x}}_{t+1})$ by definition of $\widetilde{\boldsymbol{x}}_{t+1}$ and the fact that it is included in $\mathcal{X}_t$. Thus, we get the following bound on the instantaneous regret:

$$
\eta \boldsymbol{\ell}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_j^*) \leq D(\boldsymbol{x}_j^* || \boldsymbol{x}_t) - D(\boldsymbol{x}_j^* || \boldsymbol{x}_{t+1}) + D(\boldsymbol{x}_t || \bar{\boldsymbol{x}}_{t+1}) + \ln \left(\frac{1}{1 - \frac{1}{T}} \right).
$$

Now, we define the quantities $\mathcal{I}_j \coloneqq \{t \in [T] : \phi(t) = j\}$, $\underline{t}_j \coloneqq \min_{t\in\mathcal{I}_j} t$ and we proceed bounding the regret in phase j as follows:

$$
\begin{aligned}
&\eta \sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_j^*) \\
&\leq \sum_{t\in\mathcal{I}_j} \left[D(\boldsymbol{x}_j^*||\boldsymbol{x}_t) - D(\boldsymbol{x}_j^*||\boldsymbol{x}_{t+1})\right] + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) + \sum_{t\in\mathcal{I}_j} \ln\left(\frac{1}{1-\frac{1}{T}}\right) \\
&= D(\boldsymbol{x}_j^*||\boldsymbol{x}_{\underline{t}_j}) - D(\boldsymbol{x}_j^*||\boldsymbol{x}_{\underline{t}_{j+1}}) + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) + \sum_{t\in\mathcal{I}_j} \ln\left(\frac{1}{1-\frac{1}{T}}\right) \\
&\leq 2 + \sum_{a\in[K]} x_j^*(a) \ln\left(\frac{x_{\underline{t}_{j+1}}(a)}{x_{\underline{t}_j(a)}}\right) + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) \\
&\leq 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) \\
&= 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j} \left[\sum_{a\in[K]} x_t(a) \ln\left(\frac{x_t(a)}{x_t(a)e^{-\eta\ell_t(a)}}\right) - 1 + \sum_{a\in[K]} x_t(a) e^{-\eta\ell_t(a)}\right] \\
&= 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j} \left[\eta \sum_{a\in[K]} x_t(a)\ell_t(a) - 1 + \sum_{a\in[K]} x_t(a) e^{-\eta\ell_t(a)}\right] \\
&\leq 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j} \left[\eta \sum_{a\in[K]} x_t(a)\ell_t(a) - 1 + 1 - \eta \sum_{a\in[K]} x_t(a)\ell_t(a) + \sum_{a\in[K]} x_t(a)\eta^2\ell_t^2(a)\right] \\
&\leq 2 + \ln(KT) + \eta^2 T,
\end{aligned}
$$

where we used $\ln\left(\frac{1}{1-1/T}\right) \leq \frac{1/T}{1-1/T} = 1/T-1 \leq 2/T$, the definition of the fixed share update and the inequality $e^{-z} \leq 1 - z + z^2$ for $z \geq 0$.

Thus, rearranging, we obtain the following bound:

$$\sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_j^*) \leq \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta T.$$

Noticing that the regret can be rewritten as:

$$R_T(\{\boldsymbol{x}_{\phi(t)}^*\}_{t=1}^T) \coloneqq \sum_{t=1}^T \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}_{\phi(t)}^*\right] = \sum_{j\in[S]} \sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}_j^*),$$

concludes the proof. ■

Theorem 3 *Let $\delta \in (0,1)$. With probability at least $1 - 3\delta$, Algorithm 1 attains:*

$$R_T \leq 4\log_2(T)\sqrt{T\ln(KT)} + \frac{2C}{\rho}\log_2(T) + 4\sqrt{T\ln\left(\frac{TK}{\delta}\right)}.$$

Proof To prove the result, we first employ Lemma 12 to state that the update of Algorithm 1 admits a solution for all $t \in [T]$.

Thus, we study the decision space $\mathcal{X}_t$. Specifically, thanks to Lemma 1, the following bound holds for all $\boldsymbol{x} \in \Delta_K$, $t \in [T]$ and $i \in [m]$ with probability at least $1 - \delta$:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x} \leq \frac{1}{T}\sum_{t=1}^{T} \bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x} + C/T + C/t,$$

which in turn implies:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^* \leq C/T + C/t \quad \forall t \in [T], i \in [m].$$

On the other hand, by Assumption 1 and proceeding as in Lemma 12, it holds:

$$\begin{aligned}
(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x}^\diamond &= \left(\frac{1}{t}\sum_{\tau \in [t]} \boldsymbol{g}_{\tau,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^\diamond \\
&\leq \left(\frac{1}{t}\sum_{\tau \in [t]} \bar{\boldsymbol{g}}_{\tau,i}\right)^\top \boldsymbol{x}^\diamond \\
&\leq -\rho,
\end{aligned}$$

with probability at least $1 - \delta$.

Thus, we build the following convex combination between $\boldsymbol{x}^*$ and $\boldsymbol{x}^\diamond$, parametrized given α_t, that is:

$$\boldsymbol{x}^*_{\alpha_t} = (1 - \alpha_t)\boldsymbol{x}^\diamond + \alpha_t \boldsymbol{x}^*.$$

Studying the per round optimistic violation attained by $\boldsymbol{x}^*_{\alpha_t}$, we get:

$$\begin{aligned}
\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} &= (1 - \alpha_t)\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^\diamond + \alpha_t \left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^* \\
&\leq -(1 - \alpha_t)\rho + \alpha_t \left(\frac{C}{T} + \frac{C}{t}\right) \\
&\leq -(1 - \alpha_t)\rho + \alpha_t \frac{2C}{t}.
\end{aligned}$$

Setting $\alpha_t = \frac{\rho}{\rho + 2C/t}$, we get $(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x}^*_{\alpha_t} \leq 0$, that is, $\boldsymbol{x}^*_{\alpha_t} \in \mathcal{X}_t$ for all $t \in [T]$.

We are now ready to employ Lemma 13. Notice that, since the guarantees of online mirror descent with fixed share scales linearly in the number of switches S, we employ a doubling trick approach. Specifically, in the analysis, the benchmark $\boldsymbol{x}^*_{\alpha_t}$ is updated $\log_2(T)$ times, namely, whenever the number of rounds doubles. Thus, we have the following result:

$$R_T(\{\boldsymbol{x}^*_{\alpha_{\phi(t)}}\}_{t=1}^T) \coloneqq \sum_{t=1}^{T}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{\phi(t)}}\right] \leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T \log_2(T),$$

where a new phase $j \in [S]$ starts whenever the number of rounds has doubled.

Now, we proceed by bounding the expected regret bound as follows:

$$\begin{aligned}
\sum_{t\in[T]}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] &= \sum_{t\in[T]}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t \pm \boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{\phi(t)}} - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \\
&\leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{t\in[T]}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{\phi(t)}} - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \\
&\leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{t\in[T]}\left(1-\alpha_{\phi(t)}\right)\boldsymbol{\ell}_t^\top \boldsymbol{x}^\diamond \\
&\leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{t\in[T]}\left(1-\alpha_{\phi(t)}\right) \\
&= \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{j\in[S]}(1-\alpha_j)\cdot|\mathcal{I}_j| \\
&= \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{j\in[S]}\frac{2C/\underline{t}_j}{\rho + 2C/\underline{t}_j}\cdot|\mathcal{I}_j| \\
&= \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{j\in[S]}\frac{2C}{\underline{t}_j\rho + 2C}\cdot|\mathcal{I}_j| \\
&= \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \sum_{i=1}^{k}\frac{2C}{2^{i-1}\rho + 2C}\cdot 2^{i-1} \\
&\leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + \eta T\log_2(T) + \frac{2C}{\rho}\log_2(T),
\end{aligned}$$

where $|\mathcal{I}_j|$ is the number of rounds of phase j and $\underline{t}_j \coloneqq \min_{t\in\mathcal{I}_j} t$ with $\mathcal{I}_j \coloneqq \{t\in[T] : \phi(t) = j\}$.

Setting $\eta = \frac{\sqrt{\ln(KT)}}{\sqrt{T}}$ we obtain, with probability at least $1-\delta$:

$$\sum_{t\in[T]}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \leq 4\log_2(T)\sqrt{T\ln(KT)} + \frac{2C}{\rho}\log_2(T).$$

Employing the Azuma-Höeffding inequality to bound $|\sum_{t\in[T]}\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \sum_{t\in[T]}\ell_t(a_t)|$ and $|\sum_{t\in[T]}\boldsymbol{\ell}_t^\top \boldsymbol{x}^* - \sum_{t\in[T]}\bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^*|$ with an additional union bound concludes the proof. ■

C.3. Violation

Theorem 4 *Let $\delta\in(0,1)$. With probability at least $1-\delta$, Algorithm 1 attains:*

$$V_T \leq 2 + 2C + C\ln(T) + 16\sqrt{T\ln\left(\frac{TKm}{\delta}\right)}.$$

Proof To prove the result, we first employ Lemma 12 to state that the update of Algorithm 1 admits a solution for all $t\in[T]$. Then, we bound the violation for a general constraint $i\in[m]$, which we

call

$$V_{T,i} \coloneqq \sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+,$$

for simplicity, which implies the same bound on $V_T \coloneqq \max_{i\in[m]} V_{T,i}$.

Thus, we proceed as follows:

$$\begin{aligned}
V_{T,i} &= \sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+ \\
&= \sum_{t\in[T]} \left[\left(1-\frac{1}{T}\right)\bar{\boldsymbol{g}}_{t,i}^\top \widetilde{\boldsymbol{x}}_t + \frac{1}{T}\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_{\mathcal{U}}\right]^+ \\
&\leq \sum_{t\in[T]} \left[\left(1-\frac{1}{T}\right)\bar{\boldsymbol{g}}_{t,i}^\top \widetilde{\boldsymbol{x}}_t\right]^+ + \sum_{t\in[T]} \left[\frac{1}{T}\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_{\mathcal{U}}\right]^+ \\
&= \left(1-\frac{1}{T}\right)\sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \widetilde{\boldsymbol{x}}_t\right]^+ + \frac{1}{T}\sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_{\mathcal{U}}\right]^+ \\
&\leq 1 + \sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \widetilde{\boldsymbol{x}}_t\right]^+,
\end{aligned}$$

where we simply used the definition of the fixed share update employed in Algorithm 1, the inequality $[a+b]^+ \leq [a]^+ + [b]^+$ and the fact that the maximum violation attainable in a single round is bounded by 1.

We proceed focusing on the violations attained by $\widetilde{\boldsymbol{x}}_t$. Thus, it holds:

$$\begin{aligned}
\sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \widetilde{\boldsymbol{x}}_t\right]^+ &= \bar{\boldsymbol{g}}_{1,i}^\top \widetilde{\boldsymbol{x}}_1 + \sum_{t=2}^{T} \left[\bar{\boldsymbol{g}}_{t,i}^\top \widetilde{\boldsymbol{x}}_t \pm \boldsymbol{g}_i^{\circ\top} \widetilde{\boldsymbol{x}}_t\right]^+ \\
&\leq 1 + \sum_{t=2}^{T} \|\bar{\boldsymbol{g}}_{t,i} - \boldsymbol{g}_i^\circ\|_1 + \sum_{t=2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \widetilde{\boldsymbol{x}}_t\right]^+ && \text{(4a)} \\
&\leq 1 + C + \sum_{t=2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \widetilde{\boldsymbol{x}}_t \pm (\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \widetilde{\boldsymbol{x}}_t\right]^+ && \text{(4b)} \\
&\leq 1 + C + \sum_{t=2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \widetilde{\boldsymbol{x}}_t - (\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \widetilde{\boldsymbol{x}}_t\right]^+ \\
&\quad + \sum_{t=2}^{T} \left[(\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \widetilde{\boldsymbol{x}}_t\right]^+ && \text{(4c)} \\
&= 1 + C + \sum_{t=2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \widetilde{\boldsymbol{x}}_t - (\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \widetilde{\boldsymbol{x}}_t\right]^+ && \text{(4d)} \\
&\leq 1 + C + \sum_{t=2}^{T} \left[\left(\boldsymbol{g}_i^\circ - \widehat{\boldsymbol{g}}_{t-1,i}\right)^\top \widetilde{\boldsymbol{x}}_t\right]^+ + \sum_{t=2}^{T} \left[\boldsymbol{\xi}_{t-1}^\top \widetilde{\boldsymbol{x}}_t\right]^+ && \text{(4e)}
\end{aligned}$$

$$\leq 1 + C + \sum_{t=2}^{T} \left[\boldsymbol{\sigma}_{t-1}^{\top} \widetilde{\boldsymbol{x}}_t\right]^+ + 2 \sum_{t=2}^{T} \left[\boldsymbol{\xi}_{t-1}^{\top} \widetilde{\boldsymbol{x}}_t\right]^+ \tag{4f}$$

$$= 1 + C + \sum_{t=2}^{T} \boldsymbol{\sigma}_{t-1}^{\top} \widetilde{\boldsymbol{x}}_t + 2 \sum_{t=2}^{T} \boldsymbol{\xi}_{t-1}^{\top} \widetilde{\boldsymbol{x}}_t \tag{4g}$$

$$\leq 1 + C + C(1 + \ln(T)) + 16\sqrt{T \ln\left(\frac{TKm}{\delta}\right)} \tag{4h}$$

$$= 1 + 2C + C \ln(T) + 16\sqrt{T \ln\left(\frac{TKm}{\delta}\right)},$$

where Inequality (4a) holds using $[a+b]^+ \leq [a]^+ + [b]^+$, the Hölder inequality and the fact that the violation attained in the first round is upper bounded by 1, Inequality (4b) holds by definition of C, Inequality (4c) follows from $[a+b]^+ \leq [a]^+ + [b]^+$, Equation (4d) holds since $(\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^{\top} \widetilde{\boldsymbol{x}}_t \leq 0$ for all $t \in [T]$ by definition of $\mathcal{X}_t$, Inequality (4e) follows from $[a+b]^+ \leq [a]^+ + [b]^+$, Inequality (4f) holds with probability at least $1-\delta$ employing Lemma 10 and defining $\boldsymbol{\sigma}_t \in \mathbb{R}^K$ such that $\sigma_t(a) = {}^{C}/_{t}$ for all $a \in [K]$, Equation (4g) follows from the fact that the quantities inside the $[\cdot]^+$ operator are positive and Inequality (4h) holds since

$$\sum_{t=2}^{T} \frac{1}{t-1} \leq 1 + \ln(T), \quad \sum_{t=2}^{T} \frac{1}{\sqrt{t-1}} \leq 2\sqrt{T},$$

and the fact that any strategy sums to 1.

Combining the previous equations concludes the proof. ■

Appendix D. Omitted Proofs of Section 4.3

Lemma 14 *Let $\boldsymbol{x}^*_{\phi(1)}, \dots, \boldsymbol{x}^*_{\phi(T)}$ be a S-switch dynamic benchmark as defined in Definition 2. Thus, OMD with implicit exploration and the following fixed share update:*

$$\widetilde{\boldsymbol{x}}_{t+1} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}_t} \widehat{\boldsymbol{\ell}}_t^{\top} \boldsymbol{x}_t + \frac{1}{\eta} D(\boldsymbol{x} || \boldsymbol{x}_t) \; ; \; \boldsymbol{x}_{t+1} \coloneqq \left(1 - \frac{1}{T}\right) \widetilde{\boldsymbol{x}}_{t+1} + \frac{1}{T} \boldsymbol{x}_{\mathcal{U}},$$

where $\widehat{\boldsymbol{\ell}}_t$ s.t. $\widehat{\ell}_t(a) \coloneqq \frac{\ell_t(a)}{x_t(a)+\gamma} \mathbb{I}_t(a)$ for all $a \in [K]$ attains, with probability at least $1-3\delta$:

$$R_T(\{\boldsymbol{x}^*_{\phi(t)}\}_{t=1}^T) \leq \frac{2S}{\eta} + \frac{S \ln(KT)}{\eta} + 2\eta SKT + SK \ln\left(\frac{SK}{\delta}\right) + S\sqrt{2T \ln\left(\frac{S}{\delta}\right)} + \frac{S \ln\left(\frac{KS}{\delta}\right)}{\eta}.$$

Proof To prove the result, we fix a phase $j \in [S]$, so that the baseline $\boldsymbol{x}^*_j$ used in the regret associated to that phase is fixed.

Thus, we recall the definition of the Bregman divergence

$$D(\boldsymbol{x}_1 || \boldsymbol{x}_2) \coloneqq \sum_{a \in [K]} x_1(a) \ln\left(\frac{x_1(a)}{x_2(a)}\right) - \sum_{a \in [K]} (x_1(a) - x_2(a)),$$

Algorithm 3 Constrained OMD with Fixed Share and Implicit Exploration (`ConOMD-FS IX`)

Input: $T \in \mathbb{N}$, $K \in \mathbb{N}$, $\delta \in (0,1)$

Initialize $\boldsymbol{x}_1 \leftarrow \boldsymbol{x}_{\mathcal{U}}$ where $\boldsymbol{x}_{\mathcal{U}}(a) \coloneqq 1/K \ \forall a \in [K]$

Initialize $\eta \leftarrow \sqrt{\ln(KT)/KT}, \gamma \leftarrow \eta/2$

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

  Play $a_t \sim \boldsymbol{x}_t$

  Observe loss $\ell_t(a_t)$ and constraints $\boldsymbol{g}_{t,i} \ \forall i \in [m]$

  Update empirical mean $\widehat{\boldsymbol{g}}_{t,i} \ \forall i \in [m]$

  Define $\boldsymbol{\xi}_t$ as $\xi_t(a) \coloneqq 4\sqrt{\frac{1}{t}\ln\left(\frac{TKm}{\delta}\right)}$ for all $a \in [K]$

  Build $\mathcal{X}_t \coloneqq \{\boldsymbol{x} \in \Delta_K : (\boldsymbol{g}_{t,i} - \boldsymbol{\xi}_t)^\top \boldsymbol{x} \leq 0\}$

  Build $\widehat{\boldsymbol{\ell}}_t$ such that $\widehat{\ell}_t(a) \coloneqq \frac{\ell_t(a)}{x_t(a)+\gamma}\mathbb{I}_t(a)$ for all $a \in [K]$

  Compute $\widetilde{\boldsymbol{x}}_{t+1} \leftarrow \arg\min_{\boldsymbol{x}\in\mathcal{X}_t} \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{x} + \frac{1}{\eta} D(\boldsymbol{x}||\boldsymbol{x}_t)$

  Select $\boldsymbol{x}_{t+1} \coloneqq (1-\frac{1}{T})\widetilde{\boldsymbol{x}}_{t+1} + \frac{1}{T}\boldsymbol{x}_{\mathcal{U}}$

end

which is equivalent to $D(\boldsymbol{x}_1||\boldsymbol{x}_2) \coloneqq \sum_{a\in[K]} x_1(a)\ln\left({}^{x_1(a)}/_{x_2(a)}\right)$ whenever $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are in the simplex and we define $\bar{\boldsymbol{x}}_{t+1}$ the vector whose components are $\bar{x}_{t+1}(a) = x_t(a)e^{-\eta\widehat{\ell}_t(a)}$ so that

$$\widetilde{\boldsymbol{x}}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}_t} D(\boldsymbol{x}||\bar{\boldsymbol{x}}_{t+1}).$$

We first decompose the regret in phase $j \in [S]$ as follows:

$$\sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) = \sum_{t\in\mathcal{I}_j} \widehat{\boldsymbol{\ell}}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) + \sum_{t\in\mathcal{I}_j} (\boldsymbol{\ell}_t - \widehat{\boldsymbol{\ell}}_t)^\top \boldsymbol{x}_t + \sum_{t\in\mathcal{I}_j} (\widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t)^\top \boldsymbol{x}_j^*.$$

We bound the three terms separately.

Bound on the first term We proceed similarly to Lemma 13 fixing $t \in [T]$ s.t. $\phi(t+1) = j$ and we get:

$$\eta\widehat{\boldsymbol{\ell}}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) \leq D(\boldsymbol{x}_j^*||\boldsymbol{x}_t) - D(\boldsymbol{x}_j^*||\boldsymbol{x}_{t+1}) + D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) + \ln\left(\frac{1}{1-\frac{1}{T}}\right).$$

Now, we define the quantities $\mathcal{I}_j \coloneqq \{t \in [T] : \phi(t) = j\}$, $\underline{t}_j \coloneqq \min_{t\in\mathcal{I}_j} t$ and we proceed bounding the first term in phase j as follows:

$$\begin{aligned}
&\eta \sum_{t\in\mathcal{I}_j} \widehat{\boldsymbol{\ell}}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) \\
&\leq \sum_{t\in\mathcal{I}_j} \left[D(\boldsymbol{x}_j^*||\boldsymbol{x}_t) - D(\boldsymbol{x}_j^*||\boldsymbol{x}_{t+1})\right] + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) + \sum_{t\in\mathcal{I}_j} \ln\left(\frac{1}{1-\frac{1}{T}}\right) \\
&= D(\boldsymbol{x}_j^*||\boldsymbol{x}_{\underline{t}_j}) - D(\boldsymbol{x}_j^*||\boldsymbol{x}_{\underline{t}_{j+1}}) + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) + \sum_{t\in\mathcal{I}_j} \ln\left(\frac{1}{1-\frac{1}{T}}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 + \sum_{a\in[K]} x_j^*(a) \ln\left(\frac{x_{\underline{t}_{j+1}}(a)}{x_{\underline{t}_j(a)}}\right) + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) \\
&\leq 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j} D(\boldsymbol{x}_t||\bar{\boldsymbol{x}}_{t+1}) \\
&= 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j}\left[\sum_{a\in[K]} x_t(a)\ln\left(\frac{x_t(a)}{x_t(a)e^{-\eta\widehat{\ell}_t(a)}}\right) - 1 + \sum_{a\in[K]} x_t(a)e^{-\eta\widehat{\ell}_t(a)}\right] \\
&= 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j}\left[\eta\sum_{a\in[K]} x_t(a)\widehat{\ell}_t(a) - 1 + \sum_{a\in[K]} x_t(a)e^{-\eta\widehat{\ell}_t(a)}\right] \\
&\leq 2 + \ln(KT) + \sum_{t\in\mathcal{I}_j}\left[\eta\sum_{a\in[K]} x_t(a)\widehat{\ell}_t(a) - 1 + 1 - \eta\sum_{a\in[K]} x_t(a)\ell_t(a) + \sum_{a\in[K]} x_t(a)\eta^2\widehat{\ell}_t(a)^2\right] \\
&= 2 + \ln(KT) + \eta^2\sum_{a\in[K]}\widehat{\ell}_t(a)^2 x_t(a),
\end{aligned}
$$

where we used $\ln\left(\frac{1}{1-1/T}\right) \leq \frac{1/T}{1-1/T} = 1/T-1 \leq 2/T$, the definition of the fixed share update and the inequality $e^{-z} \leq 1 - z + z^2$ for $z \geq 0$.

Thus, rearranging, we obtain the following bound for the second term:

$$
\begin{aligned}
\sum_{t\in\mathcal{I}_j}\widehat{\boldsymbol{\ell}}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) &\leq \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta\sum_{t\in\mathcal{I}_j}\sum_{a\in[K]}\widehat{\ell}_t(a)^2 x_t(a) \\
&= \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta\sum_{t\in\mathcal{I}_j}\sum_{a\in[K]}\frac{\ell_t(a)\mathbb{I}_t(a)}{x_t(a)+\gamma}\cdot\frac{\ell_t(a)\mathbb{I}_t(a)}{x_t(a)+\gamma}x_t(a) \\
&\leq \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta\sum_{t\in\mathcal{I}_j}\sum_{a\in[K]}\frac{\ell_t(a)\mathbb{I}_t(a)}{x_t(a)+\gamma}\cdot\frac{x_t(a)}{x_t(a)+\gamma} \\
&\leq \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta\sum_{t\in\mathcal{I}_j}\sum_{a\in[K]}\widehat{\ell}_t(a) \\
&\leq \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta\sum_{t\in\mathcal{I}_j}\sum_{a\in[K]}\ell_t(a) + \frac{\eta K\ln\left(\frac{K}{\delta}\right)}{2\gamma} \\
&\leq \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta TK + \frac{\eta K\ln\left(\frac{K}{\delta}\right)}{2\gamma},
\end{aligned}
$$

where we employed Corollary 1 of (Neu, 2015), which holds with probability at least $1-\delta$.

Bound on the second term We bound the term of interest as follows:

$$
\begin{aligned}
\sum_{t\in\mathcal{I}_j}\left(\boldsymbol{\ell}_t - \widehat{\boldsymbol{\ell}}_t\right)^\top\boldsymbol{x}_t &= \sum_{t\in\mathcal{I}_j}\left(\mathbb{E}[\widehat{\boldsymbol{\ell}}_t] - \widehat{\boldsymbol{\ell}}_t\right)^\top\boldsymbol{x}_t + \sum_{t\in\mathcal{I}_j}\left(\boldsymbol{\ell}_t - \mathbb{E}[\widehat{\boldsymbol{\ell}}_t]\right)^\top\boldsymbol{x}_t \\
&\leq \sqrt{2|\mathcal{I}_j|\ln\left(\frac{1}{\delta}\right)} + \sum_{t\in\mathcal{I}_j}\sum_{a\in[K]}\left(\ell_t(a) - \mathbb{E}[\widehat{\ell}_t(a)]\right)x_t(a)
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{2|\mathcal{I}_j| \ln\left(\frac{1}{\delta}\right)} + \sum_{t\in\mathcal{I}_j}\sum_{a\in[K]} \left(\ell_t(a) - \frac{\ell_t(a)x_t(a)}{x_t(a)+\gamma}\right) x_t(a) \\
&\leq \sqrt{2|\mathcal{I}_j| \ln\left(\frac{1}{\delta}\right)} + \gamma K|\mathcal{I}_j| \\
&\leq \sqrt{2T \ln\left(\frac{1}{\delta}\right)} + \gamma KT,
\end{aligned}
$$

where the first inequality holds with probability at least $1-\delta$ thanks to the Azuma-Höeffding inequality noticing that $\left|\left(\mathbb{E}[\widehat{\boldsymbol{\ell}}_t] - \widehat{\boldsymbol{\ell}}_t\right)^\top \boldsymbol{x}_t\right| \leq 1$ for all $t \in [T]$.

Bound on the third term To bound the last term we apply again Corollary 1 of (Neu, 2015) to get:

$$
\sum_{t\in\mathcal{I}_j} \left(\widehat{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\right)^\top \boldsymbol{x}_j^* \leq \frac{\ln\left(\frac{K}{\delta}\right)}{2\gamma},
$$

which holds with probability at least $1-\delta$.

Final bound Combining the previous results, we get with probability at least $1-3\delta$ by union bound:

$$
\sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) = \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + \eta TK + \frac{\eta K \ln\left(\frac{K}{\delta}\right)}{2\gamma} + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \gamma KT + \frac{\ln\left(\frac{K}{\delta}\right)}{2\gamma},
$$

which tacking $\eta = 2\gamma$ implies:

$$
\sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*) = \frac{2}{\eta} + \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta},
$$

Noticing that the regret can be rewritten as:

$$
R_T(\{\boldsymbol{x}_{\phi(t)}^*\}_{t=1}^T) \coloneqq \sum_{t=1}^T \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}_{\phi(t)}^*\right] = \sum_{j\in[S]}\sum_{t\in\mathcal{I}_j} \boldsymbol{\ell}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}_j^*),
$$

and taking a final union bound over the S phases concludes the proof. ■

Theorem 5 *Let $\delta \in (0,1)$. With probability at least $1-6\delta$, Algorithm 3 attains:*

$$
R_T \leq K\log_2(T)\ln\left(\frac{\log_2(T)K}{\delta}\right) + 11\log_2(T)\ln\left(\frac{K\log_2(T)}{\delta}\right)\sqrt{KT\ln\left(\frac{TK}{\delta}\right)} + \frac{2C}{\rho}\log_2(T).
$$

Proof To prove the result, we follow the analysis of Theorem 3.

Differently to Theorem 3, we employ the regret bound attained in Lemma 14 on the benchmark $\boldsymbol{x}^*_{\alpha_t}$, which is updated $\log_2(T)$ times, namely, whenever the number of rounds doubles. Thus, we have the following result, which holds with probability at least $1-3\delta$:

$$\begin{aligned}
R_T(\{\boldsymbol{x}^*_{\alpha_{\phi(t)}}\}_{t=1}^T) &:= \sum_{t=1}^T \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{\phi(t)}}\right] \\
&\leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + 2\eta KT\log_2(T) \\
&\quad + K\log_2(T)\ln\left(\frac{\log_2(T)K}{\delta}\right) + \log_2(T)\sqrt{2T\ln\left(\frac{\log_2(T)}{\delta}\right)} \\
&\quad + \frac{\log_2(T)\ln\left(\frac{K\log_2(T)}{\delta}\right)}{\eta}.
\end{aligned}$$

Following the analysis of Theorem 3, we get:

$$\begin{aligned}
\sum_{t\in[T]} \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] &= \sum_{t\in[T]} \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t \pm \boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{\phi(t)}} - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \\
&\leq \frac{2\log_2(T)}{\eta} + \frac{\log_2(T)\ln(KT)}{\eta} + 2\eta KT\log_2(T) \\
&\quad + K\log_2(T)\ln\left(\frac{\log_2(T)K}{\delta}\right) + \log_2(T)\sqrt{2T\ln\left(\frac{\log_2(T)}{\delta}\right)} \\
&\quad + \frac{\log_2(T)\ln\left(\frac{K\log_2(T)}{\delta}\right)}{\eta} + \frac{2C}{\rho}\log_2(T),
\end{aligned}$$

which holds with probability at least $1-4\delta$, by union bound.

Setting $\eta = \frac{\sqrt{\ln(KT)}}{\sqrt{KT}}$ we obtain, with probability at least $1-4\delta$:

$$\sum_{t\in[T]} \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \leq K\log_2(T)\ln\left(\frac{\log_2(T)K}{\delta}\right) + 7\log_2(T)\ln\left(\frac{K\log_2(T)}{\delta}\right)\sqrt{KT\ln(KT)} + \frac{2C}{\rho}\log_2(T).$$

Employing the Azuma-Höeffding inequality to bound $|\sum_{t\in[T]} \boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \sum_{t\in[T]} \ell_t(a_t)|$ and $|\sum_{t\in[T]} \boldsymbol{\ell}_t^\top \boldsymbol{x}^* - \sum_{t\in[T]} \bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^*|$ with an additional union bound concludes the proof. ■

Appendix E. Omitted Proofs of Section 5

E.1. Preliminary Results

Lemma 15 *Let $\delta \in (0,1)$. With probability at least $1-\delta$, $\mathcal{X}_t$ is not empty at each round $t \in [T]$.*

Proof To prove the result, we show that any action $\bar{a}$ so that $\bar{g}_{t,i}(\bar{a}) \leq 0$ for all $i \in [m], t \in [T]$–notice that $\bar{a}$ exists thanks to Assumption 2–is included with high probability in $\mathcal{X}_t$ at each round $t \in [T]$.

Let $\delta \in (0,1)$. Similarly to what was done in Lemma 10, we employ the Azuma inequality to get, with probability at least $1-\delta$:

$$\left|\sum_{\tau\in[t]} g_{\tau,i}(a)\mathbb{I}_\tau(a) - \sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a)\right| \leq 4\sqrt{N_t(a)\ln\left(\frac{TKm}{\delta}\right)},$$

which holds for all $t\in[T], a\in[K], i\in[m]$, by union bound. Thus, we get:

$$\left|\frac{1}{N_t(a)}\sum_{\tau\in[t]} g_{\tau,i}(a)\mathbb{I}_\tau(a) - \frac{1}{N_t(a)}\sum_{\tau\in[t]} \bar{g}_{\tau,i}(a)\mathbb{I}_\tau(a)\right| = 4\sqrt{\frac{1}{N_t(a)}\ln\left(\frac{TKm}{\delta}\right)},$$

for all $a\in[K], i\in[m], t\in[T]$.

From that, we get, with probability at least $1-\delta$:

$$\begin{aligned}
\widehat{g}_{t,i}(\bar{a}) - \xi_t(\bar{a}) &= \frac{1}{N_t(\bar{a})}\sum_{\tau\in[t]} g_{\tau,i}(\bar{a})\mathbb{I}_\tau(\bar{a}) - \xi_t(\bar{a}) \\
&\leq \frac{1}{N_t(\bar{a})}\sum_{\tau\in[t]} \bar{g}_{\tau,i}(\bar{a})\mathbb{I}_\tau(\bar{a}) \\
&\leq 0,
\end{aligned}$$

where the last step holds by definition of $\bar{a}$. This concludes the proof. ■

E.2. Regret

Theorem 6 *Let $\delta\in(0,1)$. With probability at least $1-6\delta$, Algorithm 2 attains:*

$$R_T \leq K(T^\beta+1)+3\sqrt{KT\ln(KT)}+K\ln\left(\frac{K}{\delta}\right)+5\sqrt{T\ln\left(\frac{TK}{\delta}\right)}+\sqrt{KT}\ln\left(\frac{K}{\delta}\right)+\frac{2C}{\rho}T^{1-\beta}.$$

Proof To prove the result, we first employ Lemma 15 to state that the update of Algorithm 2 admits a solution for all $t\in[T]$.

Then, we split the regret as follows:

$$\begin{aligned}
R_T &\coloneqq \sum_{t=1}^{T}\ell_t(a_t) - \sum_{t=1}^{T}\bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^* \\
&= \sum_{t\in[T_0]}\ell_t(a_t) - \sum_{t\in[T_0]}\bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^* + \sum_{t=T_0+1}^{T}\ell_t(a_t) - \sum_{t=T_0+1}^{T}\bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^* \\
&\leq K(T^\beta+1) + \sum_{t=T_0+1}^{T}\ell_t(a_t) - \sum_{t=T_0+1}^{T}\bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^*.
\end{aligned}$$

Thus, we study the decision space $\mathcal{X}_t$. Specifically, thanks to Lemma 1, the following bound holds for all $\boldsymbol{x} \in \Delta_K$, $t \in [T]$ and $i \in [m]$ with probability at least $1 - \delta$:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x} \leq \frac{1}{T}\sum_{t=1}^{T} \bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x} + C \sum_{a \in [K]} \frac{x(a)}{N_t(a)} + \frac{C}{T},$$

which in turn implies:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^* \leq C \sum_{a \in [K]} \frac{x^*(a)}{N_t(a)} + \frac{C}{T} \quad \forall t \in [T], i \in [m].$$

On the other hand, by Assumption 2 and proceeding as in Lemma 15, it holds:

$$\begin{aligned}
\widehat{g}_{t,i}(a^\diamond) - \xi_t(a^\diamond) &= \frac{1}{N_t(a^\diamond)} \sum_{\tau \in [t]} g_{\tau,i}(a^\diamond)\mathbb{I}_\tau(a^\diamond) - \xi_t(a^\diamond) \\
&\leq \frac{1}{N_t(a^\diamond)} \sum_{\tau \in [t]} \bar{g}_{\tau,i}(a^\diamond)\mathbb{I}_\tau(a^\diamond) \\
&\leq -\rho,
\end{aligned}$$

with probability at least $1 - \delta$.

Thus, we build the following convex combination between $\boldsymbol{x}^*$ and $\boldsymbol{x}^\diamond$—where, in this case, we define $\boldsymbol{x}^\diamond$ as the strategy the play $a^\diamond$ deterministically—parametrized given α_t, that is:

$$\boldsymbol{x}^*_{\alpha_t} = (1 - \alpha_t)\boldsymbol{x}^\diamond + \alpha_t \boldsymbol{x}^*.$$

Studying the per round optimistic violation attained by $\boldsymbol{x}^*_{\alpha_t}$, we get:

$$\begin{aligned}
\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} &= (1 - \alpha_t)\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^\diamond + \alpha_t \left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^* \\
&\leq -(1 - \alpha_t)\rho + \alpha_t \left(C \sum_{a \in [K]} \frac{x^*(a)}{N_t(a)} + \frac{C}{T}\right) \\
&\leq -(1 - \alpha_t)\rho + \alpha_t \cdot 2C \sum_{a \in [K]} \frac{x^*(a)}{N_t(a)}.
\end{aligned}$$

Now, we focus on the case $t > T_0$, that is, Algorithm 2 has concluded the forced exploration phase. In such a case, the algorithm guarantees $N_t(a) \geq T^\beta$ for all $a \in [K]$. From that, we have:

$$\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} \leq -(1 - \alpha_t)\rho + \alpha_t \frac{2C}{T^\beta},$$

for all $t > T_0$. Setting $\alpha_t = \alpha_{T_0} := \frac{\rho}{\rho + 2C/T^\beta}$, we get $\left(\widehat{\boldsymbol{g}}_{t,i} - \boldsymbol{\xi}_t\right)^\top \boldsymbol{x}^*_{\alpha_t} \leq 0$, that is, $\boldsymbol{x}^*_{\alpha_{T_0}} \in \mathcal{X}_t$ for all $t > T_0$. We are now ready to employ the guarantees of online mirror descent with negative entropy regularizer. Indeed, by standard analysis (e.g., following the analysis Lemma 14 without the fixed share update) and setting $\gamma = \eta/2$, we have the following result for all $\boldsymbol{x} \in \cap_{t=T_0+1}^{T} \mathcal{X}_t$:

$$\sum_{t=T_0+1}^{T} \left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}\right] \leq \frac{\ln(KT)}{\eta} + 2\eta T K + K \ln\left(\frac{K}{\delta}\right) + \sqrt{2T \ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta},$$

which holds with probability at least $1-3\delta$.

Now, we are ready to bound the expected regret bound as follows:

$$\begin{aligned}
&\sum_{t=T_0+1}^{T}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \\
&\quad= \sum_{t=T_0+1}^{T}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t \pm \boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{T_0}} - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \\
&\quad\le \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta} + \sum_{t=T_0+1}^{T}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}^*_{\alpha_{T_0}} - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \\
&\quad\le \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta} + \sum_{t=T_0+1}^{T}(1-\alpha_{T_0})\,\boldsymbol{\ell}_t^\top \boldsymbol{x}^\diamond \\
&\quad\le \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta} + \sum_{t=T_0+1}^{T}(1-\alpha_{T_0}) \\
&\quad\le \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta} + (1-\alpha_{T_0})\cdot T \\
&\quad= \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta} + \frac{2CT}{\rho T^\beta + 2C} \\
&\quad\le \frac{\ln(KT)}{\eta} + 2\eta TK + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \frac{\ln\left(\frac{K}{\delta}\right)}{\eta} + \frac{2C}{\rho}T^{1-\beta},
\end{aligned}$$

which holds with probability at least $1-4\delta$ by union bound.

Setting $\eta = \frac{\sqrt{\ln(KT)}}{\sqrt{KT}}$ we obtain, with probability at least $1-4\delta$:

$$\sum_{t=T_0+1}^{T}\left[\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \boldsymbol{\ell}_t^\top \boldsymbol{x}^*\right] \le 3\sqrt{KT\ln(KT)} + K\ln\left(\frac{K}{\delta}\right) + \sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \sqrt{KT}\ln\left(\frac{K}{\delta}\right) + \frac{2C}{\rho}T^{1-\beta}.$$

Employing the Azuma-Höeffding inequality to bound $|\sum_{t=T_0+1}^{T}\boldsymbol{\ell}_t^\top \boldsymbol{x}_t - \sum_{t=T_0+1}^{T}\ell_t(a_t)|$ and $\sum_{t=T_0+1}^{T}\boldsymbol{\ell}_t^\top \boldsymbol{x}^* - \sum_{t=T_0+1}^{T}\bar{\boldsymbol{\ell}}_t^\top \boldsymbol{x}^*|$ with an additional union bound and adding the regret incurred in the exploration phase concludes the proof. ■

E.3. Violation

Theorem 7 *Let $\delta \in (0,1)$. With probability at least $1-3\delta$, Algorithm 2 attains:*

$$V_T \le 1 + K(T^\beta + 1) + C + KC + KC\ln(T) + 30\sqrt{KT\ln\left(\frac{TKm}{\delta}\right)}.$$

Proof To prove the result, we first employ Lemma 15 to state that the update of Algorithm 2 admits a solution for all $t \in [T]$. Then, we bound the violation for a general constraint $i \in [m]$, which we call

$$V_{T,i} \coloneqq \sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+,$$

for simplicity, which implies the same bound on $V_T \coloneqq \max_{i\in[m]} V_{T,i}$.

Thus, we proceed splitting the violation as follows:

$$\begin{aligned} V_{T,i} &= \sum_{t\in[T]} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+ \\ &\leq K(T^\beta + 1) + \sum_{t=T_0+1}^{T} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+, \end{aligned}$$

where we simply used the fact that the maximum violation attainable in a single round is bounded by 1.

We proceed bounding the violation attained in the second phase of the algorithm. Thus, it holds:

$$\begin{aligned} \sum_{t=T_0+1}^{T} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t\right]^+ &= \bar{\boldsymbol{g}}_{T_0+1,i}^\top \boldsymbol{x}_{T_0+1} + \sum_{t=T_0+2}^{T} \left[\bar{\boldsymbol{g}}_{t,i}^\top \boldsymbol{x}_t \pm \boldsymbol{g}_i^{\circ\top} \boldsymbol{x}_t\right]^+ \\ &\leq 1 + \sum_{t=T_0+2}^{T} \|\bar{\boldsymbol{g}}_{t,i} - \boldsymbol{g}_i^\circ\|_1 + \sum_{t=T_0+2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \boldsymbol{x}_t\right]^+ && \text{(5a)} \\ &\leq 1 + C + \sum_{t=T_0+2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \boldsymbol{x}_t \pm (\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \boldsymbol{x}_t\right]^+ && \text{(5b)} \\ &\leq 1 + C + \sum_{t=T_0+2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \boldsymbol{x}_t - (\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \boldsymbol{x}_t\right]^+ \\ &\quad + \sum_{t=T_0+2}^{T} \left[(\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \boldsymbol{x}_t\right]^+ && \text{(5c)} \\ &= 1 + C + \sum_{t=T_0+2}^{T} \left[\boldsymbol{g}_i^{\circ\top} \boldsymbol{x}_t - (\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \boldsymbol{x}_t\right]^+ && \text{(5d)} \\ &\leq 1 + C + \sum_{t=T_0+2}^{T} \left[\left(\boldsymbol{g}_i^\circ - \widehat{\boldsymbol{g}}_{t-1,i}\right)^\top \boldsymbol{x}_t\right]^+ + \sum_{t=T_0+2}^{T} \left[\boldsymbol{\xi}_{t-1}^\top \boldsymbol{x}_t\right]^+ && \text{(5e)} \\ &\leq 1 + C + \sum_{t=T_0+2}^{T} \left[\boldsymbol{\sigma}_{t-1}^\top \boldsymbol{x}_t\right]^+ + 2 \sum_{t=T_0+2}^{T} \left[\boldsymbol{\xi}_{t-1}^\top \boldsymbol{x}_t\right]^+ && \text{(5f)} \\ &= 1 + C + \sum_{t=T_0+2}^{T} \boldsymbol{\sigma}_{t-1}^\top \boldsymbol{x}_t + 2 \sum_{t=T_0+2}^{T} \boldsymbol{\xi}_{t-1}^\top \boldsymbol{x}_t && \text{(5g)} \end{aligned}$$

$$
\begin{aligned}
&\leq 1 + C + \sum_{t=T_0+2}^{T} \sum_{a \in [K]} \sigma_{t-1}(a) \mathbb{I}_t(a) \\
&\quad + 2 \sum_{t=T_0+2}^{T} \sum_{a \in [K]} \xi_{t-1}(a) \mathbb{I}_t(a) + 6\sqrt{T \ln\left(\frac{TK}{\delta}\right)} && \text{(5h)} \\
&\leq 1 + C + CK(1 + \ln(T)) + 24\sqrt{KT \ln\left(\frac{TKm}{\delta}\right)} + 6\sqrt{T \ln\left(\frac{TK}{\delta}\right)} && \text{(5i)} \\
&\leq 1 + C + KC + KC \ln(T) + 30\sqrt{KT \ln\left(\frac{TKm}{\delta}\right)},
\end{aligned}
$$

where Inequality (5a) holds using $[a+b]^+ \leq [a]^+ + [b]^+$, the Hölder inequality and the fact that the violation for each round is upper bounded by 1, Inequality (5b) holds by definition of C, Inequality (5c) follows from $[a+b]^+ \leq [a]^+ + [b]^+$, Equation (5d) holds since $(\widehat{\boldsymbol{g}}_{t-1,i} - \boldsymbol{\xi}_{t-1})^\top \boldsymbol{x}_t \leq 0$ for all $t \in [T]$ by definition of $\mathcal{X}_t$, Inequality (5e) follows from $[a+b]^+ \leq [a]^+ + [b]^+$, Inequality (5f) holds with probability at least $1-\delta$ employing Lemma 10 and defining $\boldsymbol{\sigma}_t \in \mathbb{R}^K$ such that $\sigma_t(a) = {}^{C}/_{N_t(a)}$ for all $a \in [K]$, Equation (5g) follows from the fact that the quantities inside the $[\cdot]^+$ operator are positive, Inequality (5h) holds with probability at least $1-2\delta$ employing the Azuma-Höeffding inequality after noticing that the confidence intervals can be capped to 1 still making Lemma 10 valid and a union bound and Inequality (5i) holds since:

$$
\sum_{t=1}^{T} \sum_{a \in [K]} \frac{\mathbb{I}_t(a)}{N_{t-1}(a)} \leq K(1 + \ln(T)), \quad \sum_{t=1}^{T} \sum_{a \in [K]} \frac{\mathbb{I}_t(a)}{\sqrt{N_{t-1}(a)}} \leq 3\sqrt{KT},
$$

given that $\sum_{a \in [K]} \sqrt{N_T(a)} \leq \sqrt{K \sum_{a \in [K]} N_T(a)} \leq \sqrt{KT}$.

Combining the previous equations with a final union bound concludes the proof. ■