Constrained No-Regret Learning

Ye Du* and Ehud Lehrer†

January 12, 2020

Abstract

We investigate a dynamic decision making problem with constraints. The decision maker is free to take any action as long as the empirical frequency of the actions played does not violate pre-specified constraints. In a case of violation the decision maker is penalized. We introduce the *constrained no-regret learning model*. In this model the set of alternative strategies, with which a dynamic decision policy is compared, is the set of stationary mixed actions that satisfy all the constraints. We show that there exists a strategy that satisfies the following properties: (i) It guarantees that after an unavoidable deterministic grace period, there are absolutely no violations; (ii) For an arbitrarily small constant $\epsilon > 0$, it achieves a convergence rate of $T^{-\frac{1-\varepsilon}{2}}$ which improves the $O(T^{-\frac{1}{3}})$ convergence rate of Mannor *et al.* [23].

JEL classification. C61,C72,D81,D83

Keywords. no-regret strategy, approachability, constrained no-regret, on-line learning algorithm

*Southwestern University of Finance and Economics, China. e-mail: henry.duye@gmail.com. Du's research was supported in part by NSFC through Grant #11501464 and #11761141007.

†School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel and INSEAD, Bd. de Constance, 77305 Fontainebleau Cedex, France. e-mail: lehrer@post.tau.ac.il. Lehrer's research was supported in part by ISF through Grants #963/15 and #2510/17.

1 Introduction

Dynamic decision making, also known as *online learning* or *(external) no-regret learning*, is one of the central topics studied in game theory, computer science, and machine learning. The dynamic decision making model considers two players: a decision maker (henceforth, DM) and an adversary. At each round, the DM chooses an action to play while at the same time, the adversary chooses a state of nature. The combination of both determines a payoff received by the DM. The DM can make decisions based on the historical states of nature as well as on his own previous actions. The objective of the DM is to have *no-regret* with respect to a set of alternatives. In other words, when the DM will examine the strategy he employed by comparing his performance to what he could achieve had he played certain alternative strategies instead, he will have no regret for not playing the latter. This is a non-Bayesian model: the DM has no prior probability distribution over the states of nature chosen by the adversary. Rather, he would like to play a strategy immunized against having regret, no matter how the adversary chooses the states.

We investigate a dynamic decision problem when the DM has exogenous constraints over the empirical frequency of actions actually played. The empirical frequency of actions must be kept within a pre-specified set, and in a case of violation the DM is penalized. It might be, for instance, that the DM is not allowed to play a certain action more than 50% of the time. In such a case, whatever the DM tried to achieve, he must do it without playing more than 50% of the time this action, otherwise he would be subject to a penalty.

This constrained model is motivated by quite a few real world applications. One motivation for this new model is the *Universal Portfolio Management Problem* (see [19]). Here, a portfolio manager dynamically allocates fund across a set of assets (e.g., stocks, bonds and commodities). This manager's objective is to perform at least as good as the best fixed portfolio in hindsight. Let us assume that market data (e.g., price and trading volume) is updated N times per trading year and the maximum number of assets that can be held in a portfolio is M. In principle, a portfolio manager can adjust his allocation every time when new market data comes in. In practice, however, he cannot do so due to trading constraints imposed by regulators or stock exchanges. For instance, stock exchanges, e.g. Shanghai Stock Exchange[1], require that stocks cannot be purchased and sold out on the same day. Thus, a portfolio

[1] http://english.sse.com.cn/tradmembership/rules/c/3977570.pdf.

manager can adjust his portfolio at most MK times per year, where K is the number of trading days in a year. In other words, the constraint in this setting is that a portfolio manager cannot adjust his portfolio more than $\frac{MK}{N}$ percent of time.

Another motivating example is the *market making via online learning* (see [1]). A market maker is a trading agency that provides liquidity to a market by quoting both the bid and asking prices of an asset simultaneousness. Due to the *inventory risk* or *information asymmetry risk* [24], a market maker may be reluctant to provide quotes at unfavorable market scenarios. Nevertheless, in order to ensure enough liquidity, trading exchanges may impose constraints on the minimal number of quotes per day[2] that a market maker must provide.

Beyond exploring a non-Bayesian dynamic decision making, the theory of no-regret strategies is motivated by the connection is has to the study of economic equilibrium. It is well known that in the context of strategic interaction, when players use strategies that guarantee no (internal) regret, the empirical distributions of their joint plays converge to the set of correlated equilibria (see [17]). This result tells us that the statistics of the plays by non-Bayesian players converges to a solution concept where players hold a prior (correlated) belief about other players' actions.

In this study the objective of the DM is to minimize his regret. For this purpose we introduce a model called *constrained no-regret learning model*, which generalizes the standard no-regret approach. In this learning model, there are finitely many linear constraints on the empirical frequency of actions actually played. A sequence of actions is considered *legal* if its empirical frequency is within the limits imposed by the constraints. At any time the sequence of actions played is not legal, the DM gets a penalty.

In some real world scenarios, when rules or regulations are violated, penalties are imposed on the DM. Motivated by such examples, we introduce a penalty function into our model. As long as the sequence of actions played is legal, the penalty is naturally assumed to be zero. However, if the sequence is illegal, a penalty is deducted from the payoff to the DM. The penalty function could be very general. It could depend on time, be constant or proportional to the level of violation; it could be even a nonlinear function.

A study of a no-regret strategy in a dynamic decision problems with constraints should focus on three important issues. The first is to introduce of a proper definition

[2]http://www.hkex.com.hk/Products/Listed-Derivatives/Market-Maker-Program/

of the set of alternative strategies. This set of strategies is the set to which the DM would eventually compare the performance of his own strategy. The second issue is whether there exists a strategy that guarantees having no-regret. In other words, is there a strategy that obeys all constraints and at the same time performs as well as any alternative strategy? The third issue is the extent of the regret as the time goes by. That is, how fast the gap between the DM's actual performance and that of the best among the alternative strategies, shrinks to zero?

In the standard no-regret learning model the choice of alternative strategies is a set of all stationary pure actions, where a pure action is constantly played all the time. This set of alternatives would be inappropriate in a constrained model. The reason is that employing one of these strategies would necessarily violate one or a few of the constraints. In the constrained model, the set of alternative strategies is defined to be the set of all stationary strategies that satisfy all the constraints. These stationary strategies are typically mixed, rather than pure.

Given the set of alternative strategies, we define the *constrained regret* that corresponds to a sequence of actions and states. It is defined along the spirit of the classical regret: the gap between the actual average payoff obtained (after penalties deducted) and the best payoff that achievable by constantly playing one of the alternative strategies. A strategy in the dynamic model is *constrained no-regret* if, no matter what the penalty function is, its regret diminishes to zero as time goes by. Our main result shows that such a strategy exists. In order to obtain no-regret independently of the penalty actually imposed, we design a strategy that guarantees that after a fixed grace period, absolutely no violation of any constraint occurs.

In the proof of the main result, we consider first an auxiliary dynamic problem. In this model the decision maker has a reduced set of pure actions: only those (mixed) actions in the original problem that obey the constraints. Suppose, for instance, that there is only one constraint dictating that the frequency of playing a particular action cannot exceed twenty percent of the time. In this case, the set of actions in the auxiliary model is the set of mixed actions in which the probability of playing that action is less than, or equal, to 0.2. Employing Blackwell's Approachability Theorem, we construct a no-regret strategy τ in the auxiliary problem, where the decision maker may use all actions available without any limit.

The strategy τ is then adjusted to fit the dynamic problem with constraints. This adjustment is done as follows. The time line is divided into intervals, the length of each depends on the total lengths of its predecessors. Each interval is further divided into

three sub-intervals. In the first sub-interval, we ignore regret considerations and pay attention only to the constraint: we play a deterministic legal sequence of actions. In the second sub-interval, we follow τ as long as we run no risk of violating any constraint. In case there is a chance that we violate a constraint, we start playing the deterministic legal sequence again in the third sub-interval. This scheme is played repeatedly over and over in a carefully designed way so as to avoid any violation and at the same time minimize regret. In particular, we show that the probability of approaching a violation zone is slim, and therefore the probability to continue playing τ without any interruption and thereby guarantee no-regret, is high. This way we obtain a constrained no-regret strategy. In all, our constrained no-regret strategy has the following nice properties: (i) It guarantees that after an unavoidable deterministic grace period, there are absolutely no violations; (ii) For an arbitrarily small constant $\epsilon > 0$, it achieves a convergence rate of $T^{-\frac{1-\varepsilon}{2}}$ which improves the $O(T^{-\frac{1}{3}})$ convergence rate of Mannor *et al.* [23].

1.1 Related literature

No-regret strategies has introduced by Hanan [16]. Fudenburg *et al.* [13], Cesa-Bianchi *et al.* [5], Blum *et al.* [4], Dean *et al.* [11] as well as Perchet [25] have comprehensive discussions about no-regret learning and games. One well known dynamic decision strategy is the multiplicative weights (MW) strategy [2]. By the Blackwell's Approachability Theorem [3], one can show the MW strategy to a be no-regret one. No-regret strategy is closely related to the subject of equilibrium computation. For instance, Hart and Mas-Colell [18] showed that when all players employ (external) no-regret strategies, the statistics of play get infinitely close to the set of correlated equilibria (see also Foster and Vohra [10]). Freund *et al.* [12] show that the multiplicative weights strategy can solve a bimatrix zero-sum game efficiently.

Variants of the no-regret learning/approachability model are extensively investigated. Fudenburg *et al.* [14] studied the conditionally consistent strategy. Dekel *et al.* studied the bandit learning model with adaptive adversary [7] as well as a model with a switching cost [8]. Cesa-Bianchi *et al.* [6] looked into the problem of regret minimization with partial monitoring. Lehrer *et al.* studied no-regret learning/approachability with delayed information [20], with bounded computational capacity [21] and with bounded memory [22]. Wu *et al.* [27] investigated a multi-arm bandit problem requiring that the payoff of decision maker is above a fixed baseline,

uniformly over time.

The most relevant paper to our work is by Mannor *et al.* [23]. They investigated the no-regret learning with sample path constraints. They show that it is still possible to achieve no-regret given constraints on the empirical frequencies of actions played. In their model, no penalty[3] imposed when violation occurs and the constraints need to be satisfied only at the end of time, i.e., asymptotically. In the presence of a penalty, as introduced in our model, the regret of their strategy could be infinitely large. In contrast, our constrained no-regret strategy guarantees that after an unavoidable fixed grace period, all the constraints are satisfied. Furthermore, the convergence rate of the strategy in Mannor *et al.* [23] is in the order of $T^{-\frac{1}{3}}$. It is an open question to settle the rate of convergence when sample path constraints are imposed [23]. Fournier *et al.* [15] studied an approachability model with constraints on payoffs of Player 1 instead of constraints on empirical frequencies.

No-regret learning strategies have a various applications in practice, such as in the area of investments [1, 19, 26] and in asset pricing [9].

1.2 The structure of the paper

Section 2 introduces the general dynamic model with linear constraints on the frequency of the actions played and the notion of constrained no-regret strategy with respect to a set of constraints. We illustrate the possibility of obtaining a constrained no-regret strategy with a simple example. This example contains only one constraint: a requirement that a particular pure action would not be played more than a certain fraction of the time.

Section 3 provides a short review of the Blackwell's Approachability Theorem, which is the tool to get a no-regret strategy in the auxiliary model discussed above. In Section 4, we formally prove the existence of a constrained no-regret strategy. We conclude the paper in Section 5 by discussing related issues, including the convergence rate of the constrained no-regret dynamics.

[3]The "penalty" in Mannor *et al.* [23] is independent of sample path and constraints. It is more like a cost incurred when an action is played. In contrast, our penalty function depends the sequence of actions played and constraints, i.e. is non zero only when an illegal history of actions occurs.

2 The Model

2.1 A dynamic decision problem

Models of dynamic decision making, and more specifically, those related to online learning and no-regret learning, are central topics studied in game theory, operations research and machine learning. Consider a restricted dynamic decision making model with only two players: a decision maker (DM) and an adversary. At each round t, the DM chooses an action to play, while the adversary chooses a state of nature. The combination of both determines the payoff (or utility) received by the DM. The DM can make decisions based on the entire history of his own actions as well as on previously realized states. The objective of the DM is to achieve no-regret (with respect to a set of alternatives) that does not depend on the specific strategy employed by the adversary.

Definition 1. *A* dynamic decision problem *is given by*

- *Two players: one is a decision maker (DM) while the other is an adversary;*
- *A finite set, A, of the DM actions;*
- *A finite set, B, from which the adversary chooses a state;*
- *A payoff function, $f : A \times B \to [0, 1]$.*

A strategy of the decision maker is a function[4] $\sigma : \cup_{t=0}^{\infty}(A \times B)^t \to \Delta(A)$. That is, a strategy σ assigns a mixed action to every finite history. A strategy σ, along with a sequence $\overrightarrow{b} = (b_1, b_2, ...) \in B^{\mathbb{N}}$, induces a probability distribution $\mu_{\sigma,\overrightarrow{b}}$ over $(A \times B)^{\mathbb{N}}$, with the σ-field generated by the finite histories. We sometimes refer to a strategy as an *on-line algorithm*. Fix a sequence $\overrightarrow{b} = (b_1, b_2, ...) \in B^{\mathbb{N}}$. The regret of a sequence of actions $(a_1, ..., a_T)$ is defined as,

$$R(a_1, ..., a_T; \overrightarrow{b}) := \frac{1}{T}\Big(\sum_{t=1}^{T} f(a_t, b_t) - \max_{a \in A} \sum_{t=1}^{T} f(a, b_t)\Big).$$

$R(a_1, ..., a_T; \overrightarrow{b})$ is the difference between the actual (average) performance and what the DM could achieve by playing constantly the best mixed action against the empirical distribution of the states.

[4]For as finite set C, we denote by $\Delta(C)$ the set of probability distributions over C.

Definition 2. *A strategy (or an online algorithm) σ is* no-regret *if for every sequence* $\overrightarrow{b} = (b_1, b_2, ...)$

$$\liminf_{T\to\infty} \mathbb{E}(R(a_1, ..., a_T; \overrightarrow{b})) \geq 0, \tag{1}$$

where the expectation is w.r.t. $\mu_{\sigma,\overrightarrow{b}}$.

Eq. (1) means that the actual performance up to time T, $\frac{1}{T}\sum_{t=1}^{T} f(a_t, b_t)$, is asymptotically as good as the payoff achievable by any fixed action $a \in A$.

2.1.1 Constrained No-Regret

Let L be a natural number, $\mathbf{C}$ be a $L \times |A|$ matrix and $\mathbf{w} \in \mathbb{R}^L$. Define,[5] $\Delta(\mathbf{C}, \mathbf{w}) := \{p \in \Delta(A);\ \mathbf{C}p \leq \mathbf{w}\}$. We assume that the interior of $\Delta(\mathbf{C}, \mathbf{w})$ is non-empty. Let $\text{ext}(\mathbf{C}, \mathbf{w})$ be the set of the extreme points of $\Delta(\mathbf{C}, \mathbf{w})$. The set $\text{ext}(\mathbf{C}, \mathbf{w})$ is finite.

For a sequence $(a_1, ..., a_T) \in A^T$ we denote by $\overline{a}^T$ its empirical frequency. In particular, $\overline{a}^T \in \Delta(A)$.

Definition 3. *Let* $(a_1, ..., a_T) \in A^T$. *We say that* $(a_1, ..., a_T)$ *is* $(\mathbf{C}, \mathbf{w})$-legal *if* $\mathbf{C}\overline{a}^T \leq \mathbf{w}$.

The sequence $(a_1, ..., a_T)$ is defined to be $(\mathbf{C}, \mathbf{w})$-*legal* if the empirical frequency of $(a_1, ..., a_T)$ meets all the constraints imposed by $\mathbf{C}$ and $\mathbf{w}$. When $(a_1, ..., a_T)$ is $(\mathbf{C}, \mathbf{w})$-*legal* we sometimes say that $\overline{a}^T$ is *legal* or a legal point in $\Delta(\mathbf{C}, \mathbf{w})$.

Let F be a penalty function defined over histories of actions $(a_1, ..., a_T)$. For a legal history $(a_1, ..., a_T)$, $F(a_1, ..., a_T) = 0$; otherwise, F could be an arbitrary real function. In particular, F could be unbounded. For instance, it might be increasing with T, implying that the penalties inflicted become increasing tough. A natural penalty function might take the form of a positive constant plus a term which is proportional to the size of violation. Fix a sequence $\overrightarrow{b} = (b_1, b_2, ...) \in B^{\mathbb{N}}$. The $(\mathbf{C}, \mathbf{w})$-regret of a sequence of actions $(a_1, ..., a_T)$ is defined as,

$$R(\mathbf{C}, \mathbf{w}, F)(a_1, ..., a_T; \overrightarrow{b}) := \min_{p\in\Delta(\mathbf{C},\mathbf{w})} \frac{1}{T}\Big(\sum_{t=1}^{T}\Big(f(a_t, b_t) - f(p, b_t) - F(a_1, ..., a_t)\Big)\Big) \tag{2}$$

, where $f(p, b_t) = \sum_{a\in A} p_a f(a, b_t)$ and p_a is the probability to play action a.

[5] We abuse the notation and use vectors both as rows and columns.

The inequality $R(\mathbf{C},\mathbf{w},F)(a_1,...,a_T;\overrightarrow{b})\geq 0$ means that the actual performance up to time T is as good as that of any $p\in\Delta(\mathbf{C},\mathbf{w})$, even if when violations of the constraints are penalized.

Definition 4. *A strategy σ is* no-regret with respect to $(\mathbf{C},\mathbf{w})$ *if for every penalty function F and every sequence $\overrightarrow{b}=(b_1,b_2,...)$,*

$$\liminf_{T\to\infty}\mathbb{E}(R(\mathbf{C},\mathbf{w},F)(a_1,...,a_T;\overrightarrow{b}))\geq 0, \tag{3}$$

where the expectation is w.r.t. $\mu_{\sigma,\overrightarrow{b}}$.

In the degenerate case with $\mathbf{C}=0$ and $\mathbf{w}=0$, a strategy which is no-regret with respect to $(\mathbf{C},\mathbf{w})$ is known as an *external no-regret strategy*. By definition, a strategy σ is no-regret with respect to $(\mathbf{C},\mathbf{w})$ if it guarantees that the regret is non-negative for *any* penalty function F. This definition is rather restrictive and therefore induces some nice properties of σ. For instance, when F is bounded away from zero on every illegal sequence (i.e., there is $\delta>0$ such that $F(a_1,...,a_T)>\delta$ for every illegal history $(a_1,...,a_T)$), it implies that the frequency of violations should diminish with time. Furthermore, when F is unbounded (e.g., when a repetitive violation may result in increasing penalties), then Eq. (3) implies that there must be only finitely many violations. Otherwise, if there are infinite number of violations, a penalty function such as $F(a_1,...,a_T)=T^2$ would imply that the regret approaches minus infinity. Therefore, a constrained no-regret strategy has a strong property: there is a deterministic time T_0 such that for all $T\geq T_0$, $(a_1,...,a_T)$ is $(\mathbf{C},\mathbf{w})$-legal with probability 1. In other words, absolutely no violations occur after T_0. We call T_0 the *grace period*.

We deal with a non-Bayesian model. The DM does not have a prior belief about the evolution of states. This is the reason why the DM's objective is not to maximize future payoffs. Rather, the DM's objective is to minimize regret, taking into account the penalty imposed in a case of violation. The DM compares between his actual performance and an imaginary one had he used an alternative strategy. We show that there is a strategy that if followed, avoids having regret.

2.2 An Illustrative example

Assume that the decision maker has two actions: $A=\{L,R\}$. The payoffs received when he plays L and R are, regardless of the strategy of the adversary, 0 and 1,

respectively. Suppose we have only one constraint: the action R cannot be played more than a half of the time. It is obvious that there is no legal strategy that can guarantee the decision maker an average payoff of more than $\frac{1}{2}$. However, if the decision maker could play the action R without any constraint, which is illegal, the average payoff would be 1. The previous notations adapted to this case would be the following: $\mathbf{C} = (0,1)$, $\mathbf{w} = \frac{1}{2}$, $\Delta(\mathbf{C},\mathbf{w}) = \{(p_1,p_2) \in \Delta(A);\ \langle \mathbf{C}, (p_1,p_2)\rangle \leq \mathbf{w}\} = \{(p_1,p_2) \in \Delta(A);\ p_2 \leq \frac{1}{2}\}$. The set of extreme points is $\mathrm{ext}(\mathbf{C},\mathbf{w}) = \{(1,0),(\frac{1}{2},\frac{1}{2})\}$.

Imagine a situation where the decision maker has only two pure actions: L and $\frac{1}{2}L \oplus \frac{1}{2}R$ (which corresponds to the mixed action $(\frac{1}{2},\frac{1}{2})$ in the original decision problem). Suppose that the payoff corresponding to the latter strategy is its respective expected payoff, $\frac{1}{2}$. A classical result [5] guarantees that there is an external no-regret algorithm. The strategy in this case plays stationarily the action $\frac{1}{2}L \oplus \frac{1}{2}R$. This is what we call the baseline strategy τ.

However, when a strategy analogous to τ is played in the original set-up, the pure action $\frac{1}{2}L \oplus \frac{1}{2}R$ is replaced by a mixed strategy that plays R and T with probability $\frac{1}{2}$ each. In this case, it might occur with positive probability that R is played with a frequency that exceeds 50% in which case the constraint is violated. In order to handle this violation, we design a new strategy σ as following: **(1)** The action L is played in an initial time segment of length $\sqrt{T}$; **(2)** from time $\sqrt{T}+1$ and on, a strategy analogous to τ is played as long as there is no violation; **(3)** if taking one more step may lead to a violation, action R is not being played anymore. Instead, action L is played with probability 1.

It is clear that the total payoff related to strategy σ could be lower than that related to strategy τ. We argue that the average payoffs when playing τ or σ are close to each other. The reason is that the probability of violation is small. The probabilistic argument is given in Lemma 2 below.

3 Vector-payoff games and approachability

The time horizon of the game is $\{1,2,...,T\}$. At each time step, the decision maker will choose an action a_t from an action set $A = \{1,2,...,N\}$, while the adversary will choose an action b_t from a set $B = \{1,2,...,M\}$. A vector-valued function $\mathbf{g} : A \times B \to \mathbb{R}^m$ is the vector-valued payoff function. Suppose that the realized pair of actions at time t is (a_t,b_t), then the payoff is $\mathbf{g}(a_t,b_t)$.

Definition 5. *We say that the set $S \subseteq \mathbb{R}^m$ is* approachable by player 1 *(or simply,* approachable*) if he has a strategy such that regardless of the strategy of the adversary,*

$$\lim_{T\to\infty} \mathbb{E}[d(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{g}(a_t, b_t), S)] = 0,$$

where $d(u, S) = \inf_{v\in S} \|u - s\|$.

The set S is called the *target set*. For a convex set to be approachable, we have the following characterization.

Theorem 1. *[5] Let S be closed and convex. Then, S is approachable if for every $\mathbf{x} \notin S$ there is a mixed action π such that for every $b \in B$,*

$$\langle \mathbf{x} - \mathbf{q}_S(\mathbf{x}), \boldsymbol{g}(\pi, b) - \mathbf{q}_S(\mathbf{x})\rangle \leq 0, \tag{4}$$

where $\mathbf{q}_S(\mathbf{x}) = \arg\min_{\boldsymbol{v}\in S} \|\boldsymbol{x} - \boldsymbol{v}\|$ is the closest point in S to $\mathbf{x}$. Moreover, there is a strategy that guarantees that[6]

$$\mathbb{E}[d(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{g}(a_t, b_t), S)] = O(\frac{1}{\sqrt{T}}). \tag{5}$$

4 Existence of a constrained no-regret strategy

Our main theorem is the following.

Theorem 2. *For every $\varepsilon > 0$ there is a strategy σ such that $\lim_{T\to\infty} \mathbb{E}_\sigma(R(\mathbf{C}, \mathbf{w}, F)) \geq 0$, and the regret is bounded below by $-O(T^{-\frac{1-\varepsilon}{2}})$.*

Lemma 1. *Denote $A' = \text{ext}(\mathbf{C}, \mathbf{w})$. There is a no-regret strategy τ with respect to $(\mathbf{C}, \mathbf{w})$ when Player 1 can use only actions (pure or mixed) in A'. Moreover, for every $\overrightarrow{b}$,*[7]

$$\mathbb{E}(R(\mathbf{C}, \mathbf{w}, F)(\mathbf{a'}_1, ..., \mathbf{a'}_T; \overrightarrow{b})) \geq -O(\frac{1}{\sqrt{T}}). \tag{6}$$

[6] $O(\frac{1}{\sqrt{T}})$ means that there is a constant $c > 0$ such that the LHS is bound from above by by $c\frac{1}{\sqrt{T}}$. In general $O(x)$ means that the term under consideration is bounded from above by a linear function of x.

[7] To avoid any ambiguity, $(\mathbf{a'}_1, ..., \mathbf{a'}_T)$ are the histories realized pure actions in $(A')^T$.

Proof. For the sake of simplicity, we enumerate $\text{ext}(\mathbf{C},\mathbf{w}) = \{\mathbf{p^1},...,\mathbf{p^K}\}$. We define a vector-valued game and use the approachability method as follows. Let $S = \{(x_1,...,x_K);\ \forall k, x_k \leq 0\}$ be the target set. For every $\pi = (\pi_1,...,\pi_K) \in \Delta(A')$ and $b \in B$, define $f(\pi,b)$ to be the expected value of f when player 1 mixes (over A') according to π. For $\pi \in \Delta(A')$ and $b \in B$, we define a vector-payoff $\mathbf{g}(\pi,b)$ in $\mathbb{R}^{\text{ext}(\mathbf{C},\mathbf{w})}$. The k-th coordinate of this vector is $\mathbf{g}_k(\pi,b) := f(\pi,b) - f(\mathbf{p}^k,b)$.

Fix $\mathbf{x} \notin S$ and define,

$$\pi := \frac{\mathbf{x} - \mathbf{q}_S(\mathbf{x})}{\|\mathbf{x} - \mathbf{q}_S(\mathbf{x})\|_1}$$

Since $\mathbf{x} \notin S$, π is a probability distribution over A', namely in $\Delta(A')$. It is easy to see, therefore, that $\pi \cdot \mathbf{q}_S(\mathbf{x}) = 0$. For any pure strategy $b \in B$ of the adversary we have,

$$\begin{aligned}\langle \mathbf{x} - \mathbf{q}_S(\mathbf{x}), \mathbf{g}(\pi,b) - \mathbf{q}_S(\mathbf{x})\rangle &= \langle \mathbf{x} - \mathbf{q}_S(\mathbf{x}), \mathbf{g}(\pi,b)\rangle = \|\mathbf{x} - \mathbf{q}_S(\mathbf{x})\|_1 \langle \pi, \mathbf{g}(\pi,b)\rangle \\ &= \|\mathbf{x} - \mathbf{q}_S(\mathbf{x})\|_1 \sum_{k=1}^{K} \pi_k \cdot (\sum_{h=1}^{K} \pi_h [f(\mathbf{p}^h,b) - f(\mathbf{p}^k,b)]) = 0.\end{aligned}$$

Thus, by Theorem 1, we obtain that S is approachable. That is, there is a strategy τ of player 1 such that for every sequence $b_1, b_1, ...$ in $B^{\mathbb{N}}$ and for every $\mathbf{p}^k \in \Delta(\mathbf{C},\mathbf{w})$,

$$\liminf_{T\to\infty} \mathbb{E}[\frac{1}{T}[\sum_{t=1}^{T} f(\mathbf{a'}_t, b_t) - \sum_{t=1}^{T} f(\mathbf{p_k}, b_t)] \geq 0, \tag{7}$$

where the expectation is taken w.r.t the probability induced by τ. Note again that $\mathbf{a'}_t \in A'$ is the realized action in A' at time t. This inequality is equivalent to Eq. (3). The rate of convergence, namely Eq. (6), is implied by Eq. (5). □

We denote by τ_t the history-dependent mixed action in $\Delta(\mathbf{C},\mathbf{w})$ to be played at time t. Note that the history here is a sequence of actions in $\text{ext}(\mathbf{C},\mathbf{w})$ (which are typically mixed actions in $\Delta(A)$). Since $\tau_t \in \Delta(\mathbf{C},\mathbf{w})$, it induces a mixed action over A. Thus, when using τ an action a_t in A is actually realized at time t. We abuse the notation and refer to τ_t as a history dependent mixed action over A, namely in $\Delta(A)$. For any t we have $\mathbb{E}[f(\mathbf{a'_t}, b_t)] = \mathbb{E}[f(a_t, b_t)]$. Thus, Eq. (7) is translated to

$$\liminf_{T\to\infty} \mathbb{E}[\frac{1}{T}[\sum_{t=1}^{T} f(a_t, b_t) - \sum_{t=1}^{T} f(\mathbf{p_k}, b_t)] \geq 0. \tag{8}$$

Recall that the matrix $\mathbf{C}$ has L rows, each representing one linear constraint. Fix $\ell \in \{1, ..., L\}$ and let $\mathbf{C}_\ell$ be the ℓ-th row of $\mathbf{C}$ and w_ℓ be the ℓ-th coordinate of $\mathbf{w}$. Assume w.l.o.g. that $\|\mathbf{C}_\ell\|_1 < 1$.

Consider time t and let $\mathbb{1}_{a_t} \in \mathbb{R}^{|A|}$ be the indicator of the action a_t played at time t. Define $X^0 = 0$, $Y_\ell^t = \mathbf{C}_\ell \mathbb{1}_{a_t}$ and $X_\ell^t = Y_\ell^t - \mathbb{E}[Y_\ell^t | \mathcal{F}_{t-1}]$, where $\mathcal{F}_{t-1}$ is the algebra generated by histories of length $t-1$.

Remark 1. *(i) The sequence $\sum_{i=0}^{T} X_\ell^i$, $T = 0, 1, 2, ...$, is a martingale.*
(ii) $|\sum_{i=0}^{t} X_\ell^i - \sum_{i=0}^{t-1} X_\ell^i| = |X_\ell^t| \leq 1$.
(iii) The strategy τ used only mixed actions in $\text{ext}(\mathbf{C}, \mathbf{w})$. Therefore, $\mathbb{E}[Y_\ell^t | \mathcal{F}_{t-1}] \leq w_\ell$.

Lemma 2. *Let $\mathbb{1}_{a_t}$ be generated by the strategy τ. Fix $\varepsilon > 0$ and $D > 0$. Then, $\forall \ell$*

$$P\left(\sum_{t=1}^{T} (\boldsymbol{C}_\ell \mathbb{1}_{a_t} - w_\ell) > DT^{\frac{1+\varepsilon}{2}}\right) \leq \exp\left(\frac{-T^\varepsilon D^2}{2}\right). \tag{9}$$

Lemma 2 states that while employing strategy τ the probability to exceed constraint ℓ by more than $DT^{\frac{1+\varepsilon}{2}}$ is small: it is bounded from above by $\exp\left(\frac{-T^\varepsilon D^2}{2}\right)$. The proofs of the lemmas appearing from here on can be found in the Appendix.

Concerning the probability to violate at least one among the L constraints, we have the following corollary according to the union bound.

Corollary 1. *Fix $\varepsilon > 0$ and $D > 0$. Then,*

$$P\left(\sum_{t=1}^{T} (\mathbb{1}_{a_t} \cdot \boldsymbol{C}_\ell - w_\ell) > DT^{\frac{1+\varepsilon}{2}}, \forall \ell = 1, 2, ..., L\right) \leq L \exp\left(\frac{-T^\varepsilon D^2}{2}\right). \tag{10}$$

We are now ready to prove Theorem 2.

4.1 The main idea of our proof

The main idea of our proof can be illustrated by Figure 1. Fix an interior point of $\Delta(\mathbf{C}, \mathbf{w})$, $\mathbf{m}$. We define a strict subset of $\Delta(\mathbf{C}, \mathbf{w})$ that contains $\mathbf{m}$ whose boundary is called the *critical boundary*. The time line is divided into intervals, the length of each depends on the total lengths of its predecessors. Each interval is further divided into three sub-intervals (a sub-interval is referred to as a block as well). During the first sub-interval, a sequence of actions that mimics the mixed strategy corresponding to $\mathbf{m}$ is played repeatedly. During the second one, a standard no-regret strategy is

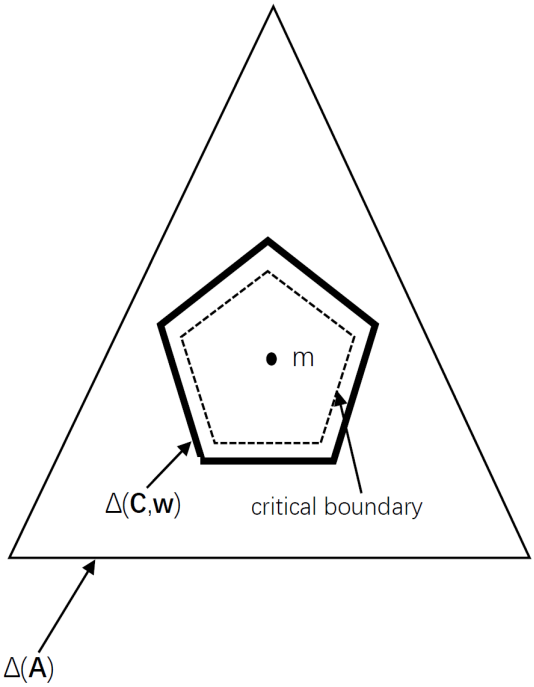

Figure 1: The legal set and the critical boundary

played as the empirical frequency does not hit the critical boundary. If it does hit, **m** will be played again in the third sub-interval.

The strategy thus designed has two desirable properties: (i) After the first sub-interval of the first interval is over, there is absolutely no constraints violations. In other words, penalty might be inflicted only during the first sub-interval of the first interval, meaning only finitely many times. Therefore, the average penalty paid diminishes with time. (ii) Since the length of the first sub-interval of each interval is relatively small and the probability of hitting the critical boundary is small as well, the regret is close to 0.

4.2 The proof of Theorem 2

Fix a small ε. Strategy σ is defined as follows. In order to construct strategy σ, we divide the time horizon $[1, \infty)$ into intervals and define the strategy separately in each interval. We will use the following notations. Let $\mathbf{x}^{[t_1,t_2]}$ be the empirical

frequency of the actions played between times t_1 and t_2 (inclusive) and $\mathbf{x}^t := \mathbf{x}^{[1,t]}$. For every $\mathbf{x} \in \Delta(A)$ define $d(\mathbf{x}) = \min_{\ell=1,...,L}(\mathbf{w}_\ell - \mathbf{C}_\ell \mathbf{x})$. Clearly, $d(\mathbf{x}) \geq 0$ if and only if $\mathbf{x} \in \Delta(\mathbf{C}, \mathbf{w})$ with strict inequality when $\mathbf{x}$ is in the interior of $\Delta(\mathbf{C}, \mathbf{w})$. Note that d is concave over $\Delta(\mathbf{C}, \mathbf{w})$. That is,

$$d\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) \geq \alpha d(\mathbf{x}) + (1-\alpha)d(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \Delta(\mathbf{C}, \mathbf{w}) \text{ and } \alpha \in (0,1). \tag{11}$$

Let $\mathbf{m} = (\frac{m_1}{M}, ..., \frac{m_{|A|}}{M})$ be a rational vector in the interior of $\Delta(\mathbf{C}, \mathbf{w})$, where $m_i \in \mathbb{N}, \forall i$ and $M \in \mathbb{N}$. Note that since $\mathbf{m}$ is in the interior of $\Delta(\mathbf{C}, \mathbf{w})$, $d(\mathbf{m}) > 0$.

The first interval:
The first interval is $[0, H \cdot M]$, where H is a large enough natural number to be determined later. The interval is further divided up to three blocks as illustrated in Figure 2. In the first block,[8] $[1, MH^{\frac{1+\varepsilon}{2}}]$, a sequence $(a_1, ..., a_M) \in A^M$ whose frequency

Figure 2: The structure of the first interval

coincides with $\mathbf{m}$ is repeatedly played for $H^{\frac{1+\varepsilon}{2}}$ times. The purpose of this block is to bring $\mathbf{x}^t$ to the interior of $\Delta(\mathbf{C}, \mathbf{w})$. This provides enough room for the following strategy to play without violating any constraint. Thus, $T_0 \leq MH^{\frac{1+\varepsilon}{2}}$.

The size of the second block is not deterministic. It depends on the realization of the following strategy. We denote $\mathbf{y}^t = \mathbf{x}^{[MH^{\frac{1+\varepsilon}{2}}+1,t]}$ the empirical frequency of actions during the second block up-to the t-th period. In the second block the decision maker follows strategy τ (derived from Lemma 1 above), as long as for every $\ell = 1, ..., L$,

$$t(\mathbf{C}_\ell \mathbf{y}^t - \mathbf{w}_\ell) \leq \frac{d(\mathbf{m})}{2} MH^{\frac{1+\varepsilon}{2}}. \tag{12}$$

. If Eq. is not violated, τ is played up to the end of the interval.

In the case where during the second block, while playing τ, Eq. (13) is violated, the third block starts. Here, the strategy plays a deterministic sequence of actions in a way

[8] Here and in the sequel, when we refer to a number which is not an integer, as period/time, such as $MH^{\frac{1+\varepsilon}{2}}$, we mean the largest integer smaller than that number.

that guarantees no violation of constraints. Specifically, the sequence $(a_1, ..., a_M) \in A^M$ whose frequency coincides with $\mathbf{m}$ is repeatedly played until the end of the first interval.

$$t(\mathbf{C}_\ell \mathbf{y}^t - \mathbf{w}_\ell) \leq \frac{d(\mathbf{m})}{2} M H^{\frac{1+\varepsilon}{2}}. \tag{13}$$

Otherwise, the sequence $(a_1, ..., a_M)$ is played repeatedly again to the end of the interval. We first show the following lemma.

Lemma 3. *If at time t Eq. (13) is satisfied for every $\ell = 1, ..., L$, then all the constraints are kept at time t.*

We show next that the probability of getting a violation of Eq. (13) at time t is small. Specifically,

Lemma 4. *There is a constant $c > 0$ such that the probability of violating Eq. (13) at any time during the second block up to time $M \cdot H$ is bounded from above by $M \cdot H \cdot L \exp\left(\frac{-(cMH^{\frac{1+\varepsilon}{2}})^\varepsilon (\frac{d(m)}{2} M^{\frac{1-\varepsilon}{2}})^2}{2}\right)$.*

Note that the bound in Lemma 4 goes to zero as H goes to infinity.

We claim that during the third block no violation occurs with respect to any of the constraints. Formally,

Lemma 5. *For every time t'' during the third block, $d(\mathbf{x}^{t''}) > 0$.*

At the end of the first interval, the strategy guarantees that (a) $\forall t \in [MH^{\frac{1+\varepsilon}{2}}, MH]$, $d(\mathbf{x}^t) > 0$; and (b) the regret during the HM periods of the first interval is bounded below by

$$\begin{aligned}
-\frac{MH^{\frac{1+\varepsilon}{2}}}{HM} &- \left(1 - M \cdot H \cdot L \exp\left(\frac{-(cMH^{\frac{1+\varepsilon}{2}})^\varepsilon (\frac{d(\mathbf{m})}{2} M^{\frac{1-\varepsilon}{2}})^2}{2}\right)\right) O\left(\frac{(MH)^{\frac{1}{2}}}{HM}\right) \\
&- M \cdot H \cdot L \exp\left(\frac{-(cMH^{\frac{1+\varepsilon}{2}})^\varepsilon (\frac{d(\mathbf{m})}{2} M^{\frac{1-\varepsilon}{2}})^2}{2}\right) O(1) \\
&= -O\left(\frac{MH^{\frac{1+\varepsilon}{2}}}{MH}\right),
\end{aligned}$$

where the first term bounds the total regret in the first block. Let $T = HM$. Note that M is a constant, hence the regret is bounded by $-O(T^{-\frac{1-\varepsilon}{2}})$.

In the next, we proceed to the second, the third, up to the Nth block recursively. The length of each interval increases exponentially as illustrated in Figure 3. It is not hard to see that the convergences rate will be kept and the key is to make sure the empirical frequency is legal.

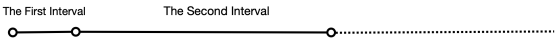

Figure 3: The structure of consecutive intervals

The N-th interval: Let $t_{N-1} = M\sum_{n=0}^{N-1} H^n$ and the N-th interval be $[t_{N-1}+1, t_N]$. Thus, the length of the N-th interval is MH^N. The first block is $[t_{N-1}+1, t_{N-1}+MH^{\frac{N(1+\varepsilon)}{2}}]$. In this block a sequence $(a_1, ..., a_M) \in A^M$ whose frequency coincides with $\mathbf{m}$ is repeatedly played for $H^{\frac{N(1+\varepsilon)}{2}}$ times. Following a similar argument to the one used for analyzing the third block of the last interval, we may conclude that $d(\mathbf{x}^t) > 0$ for every $t \in [t_{N-1}+1, t_{N-1}+MH^{\frac{N(1+\varepsilon)}{2}}]$.

The second block of the N-th interval follows strategy τ, as if the time starts just after the first block. It will stop at the the smallest time t in the second block where

$$t(\mathbf{C}_\ell \mathbf{y}^t - \mathbf{w}_\ell) > \frac{d(\mathbf{m})}{2} MH^{\frac{N(1+\varepsilon)}{2}} \tag{14}$$

for at least one $\ell = 1, ..., L$, where $\mathbf{y}^t = \mathbf{x}^{[t_{N-1}+MH^{\frac{N(1+\varepsilon)}{2}}+1, t]}$. If such a violation happens, the third block starts. In the third block the sequence $(a_1, ..., a_M) \in A^M$ whose frequency coincides with $\mathbf{m}$ is repeatedly played until the end of the N-th interval. Following the same argument as the (N-1)-th interval, we have $\forall t \in [t_{N-1}+1, t_N], d(\mathbf{x}^{[t_{N-1}+1,t]}) > 0$. We further get

$$d(\mathbf{x}^t) \geq \frac{t_{N-1}}{t} d(\mathbf{x}^{t_{N-1}}) + \frac{t - t_{N-1}}{t} d(\mathbf{x}^{[t_{N-1}+1,t]}) > 0.$$

Thus, no violation occurs in any period belonging to the N-th interval.

Next, we address the scope of the regret. We claim:

Lemma 6. *The regret at any time t during the N-th block is bounded below by $-O\left(t^{-\frac{1-\varepsilon}{2}}\right)$.*

This lemma completes the proof.

5 Conclusions and Open Problems

5.1 Contributions

The contribution of the paper is threefold. (i) We introduce a constrained no-regret model with penalties over violations. Our penalty function is general and reflects quite a few real world applications; (ii) Our constrained no-regret strategy can guarantee that after an unavoidable fixed (not random) grace period there are absolutely no violations; (iii) For an arbitrarily small constant $\epsilon > 0$, our strategy achieves a convergence rate of $T^{-\frac{1-\varepsilon}{2}}$ which significantly improves the $O(T^{-\frac{1}{3}})$ convergence rate of Mannor *et al.* [23].

Note that the key difference between our model and that of Mannor *et al.* [23] is that the constraints in their model only needs to be satisfied at the end of time, i.e., asymptotically. It implies those constraints could be violated in infinite number of times. In other words, at any time there is a possibility that the constraints get violated. In contrast, our model requires that after an unavoidable fixed grace period, all constraints are strictly satisfied. Thus, any action sequence generated by our constrained no-regret strategy in particular satisfies the constraints related to their model. Therefore, the set of strategies concerning our model is more restricted than that of Mannor *et al.* [23]. A no-regret strategy in our setting is interesting in its own right.

In our model the set of constraints are linear. This implies that the set of legal empirical frequency of a history of actions forms a polygon. The same proof technique can be used to prove similar results for any convex set of legal empirical frequencies, as long as this set has a non-empty interior.

In this paper we choose the set of alternative strategies to be the set of stationary mixed actions that satisfy finitely many linear constraints. It is important to note that our proof technique could be applied to any countable set of alternative choices,[9] as long as the inequality in Eq. (4) could be satisfied.

5.2 Open Problems

There are some problems that are worth further investigating.

Improvement of the regret bound The constant ε appears in the bound $T^{-\frac{1-\varepsilon}{2}}$

[9]The vector-payoff $\mathbf{g}(\pi, b)$ can be with a countable dimension – see Lehrer (2002).

bound. A removal of ε from this bound would make our constrained regret match the regret in the standard no-regret learning model.

Other choices of the set of alternative strategies In this paper the constraints imposed on the empirical frequency of actions played are linear. This makes the set $\text{ext}(\mathbf{C}, \mathbf{w})$ convex. If $\text{ext}(\mathbf{C}, \mathbf{w})$ is non-convex, what could be plausible alternative sets and constrained regret functions? Can one still be able to design a no-regret strategy?

References

[1] Abernethy, J. and S. Kale (2013),"Adaptive Market Making via Online Learning", *NIPS*, 2058-2066.

[2] Arora, S., E. Hazan, and S. Kale (2012), "The Multiplicative Weights Update Method: a Meta-Algorithm and Applications" , *Theory of Computing*, **8**, 121-164,

[3] Blackwell, D. (1956), "An Analog of the Minimax Theorem for Vector Payoffs", *Pacific Journal of Mathematics*, **6**, 1-8.

[4] Blum, A. and Y. Mansour (2007), "Learning, Regret minimization, and Equilibria" , *in Algorithmic Game Theory*, Cambridge, UK: Cambridge University Press.

[5] Cesa-Bianchi, N. and G. Lugosi (2006), "Prediction, Learning, and Games", *Cambridge University Press.*

[6] Cesa-Bianchi, N., G. Lugosi, and G. Stoltz (2006) "Regret Minimization under Partial Monitoring," *Mathematics of Operations Research*, **31**, 562-580.

[7] Dekel, O., A. Tewari and R. Arora (2012), "Online Bandit Learning against an Adaptive Adversary: from Regret to Policy Regret", *ICML.*

[8] Dekel, O., J. Ding, T. Koren and Y. Peres (2014), "Bandits with Switching Costs: $T^{2/3}$ Regret", *STOC*, 459-467.

[9] DeMarzo, P., I. Kremer, and Y. Mansour (2016), "Robust Option Pricing: Hannan and Blackwell meet Black and Scholes", *Journal of Economic Theory*, **163**, 410-434.

[10] Foster, D. and R. Vohra (1997), "Calibrated Learning and Correlated Equilibrium," *Games and Economics Behavior*, **21**, 40-55.

[11] Foster, D. and R. Vohra (1999), "Regret in the On-Line Desicion Problem," *Games and Economics Behavior*, **29**, 7-36.

[12] Freund, Y. and R. Schapire (1999), "Adaptive Game Playing using Multiplicative Weights", *Games and Economic Behavior*, **29**, 79-103.

[13] Fudenberg, D. and D. Levine (1998), "The Theory of Learning in Games", *MIT press*.

[14] Fudenberg D. and D. Levine (1999), "Universal Conditional Consistency," *Games and Economics Behavior*, **29**, 104-130.

[15] Fournier, G., E. Kuperwasser, O. Munk, E. Solan and A. Weinbaum (2017), "Approachability with Constraints", *manuscript*.

[16] Hannan, J. (1957), "Approximation to Bayes Risk in Repeated Plays", in *Contribution to the Theory of Games*, **3**, 97-139. Princeton, NJ: Princeton University Press.

[17] Hart, S. and A. Mas-Colell (2000), "A Simple Adaptive Procedure Leading to Correlated Equilibrium" *Econometrica*, **68**, 1127-1150.

[18] Hart, S. and A. Mas-Colell (2001), "A General Class of Adaptive Strategies," *Journal of Economic Theory*, **98**, 26-54.

[19] Hazan, E. and S. Kale (2009), "On Stochastic and Worst-case Models for Investing", *NIPS*, 709-717.

[20] Lagziel, D. and E. Lehrer (2012), "Approachability with Delayed Information", *Journal of Economic Theory*, **157**, 425-444.

[21] Lehrer, E. and E. Solan (2006), "Excludability and Bounded Computational Capacity Strategies", *Mathematics of Operations Research*, **31**, 637-648.

[22] Lehrer, E. and E. Solan (2009), "Approachability with Bounded Memory", *Games Econ. Behavior*, **66**, 995-1004.

[23] Mannor, S., J. Tsitsiklis and J. Yu (2009), "Online Learning with Sample Path Constraints", *Journal of Machine Learning Research*, **10**, 569-590.

[24] O'Hara, M. (1998), "Market Microstructure Theory", *Wiley.*

[25] Perchet, V. (2014), "Approachability, regret and calibration: Implications and equivalences", *Journal of Dynamics and Games*, **1(2)**, 181-254.

[26] Stoltz, G. and G. Lugosi (2005), "Internal Regret in On-Line Portfolio Selection," *Machine Learning*, **59**, 125-159.

[27] Wu, Y., R. Shariff, T. Lattimore, and C. Szepesvari (2016). "Conservative bandits", *In Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

6 Appendix

Proof of Lemma 2.

$$\begin{aligned} & P\left(\sum_{t=1}^{T}(\mathbf{C}_\ell \mathbb{1}_{a_t} - w_\ell) > DT^{\frac{1+\varepsilon}{2}}\right) \leq P\left(\sum_{t=1}^{T}\left(Y_\ell^t - \mathbb{E}[Y_\ell^t|\mathcal{F}_{t-1}]\right) > DT^{\frac{1+\varepsilon}{2}}\right) \\ = \quad & P\left(\sum_{t=1}^{T} X_\ell^t > DT^{\frac{1+\varepsilon}{2}}\right) \leq \exp\left(\frac{-(DT^{\frac{1+\varepsilon}{2}})^2}{2T}\right) = \exp\left(\frac{-T^\varepsilon D^2}{2}\right). \end{aligned}$$

The first inequality is by Remark 1(iii). The last inequality is by Remark 1(i)-(ii) that allow the use of Azuma inequality. □

Proof of Lemma 3. Note that for every $\ell = 1, ..., L$,

$$t(\mathbf{C}_\ell \mathbf{y}^t - \mathbf{w}_\ell) \leq \frac{d(\mathbf{m})}{2}(MH)^{\frac{1+\varepsilon}{2}} \leq \frac{(\mathbf{w}_\ell - \mathbf{C}_\ell \mathbf{m})}{2} MH^{\frac{1+\varepsilon}{2}}. \tag{15}$$

Recall that t is the t-th period counted from the start of the second block and is therefore the $t + MH^{\frac{1+\varepsilon}{2}}$ period from the beginning. Note that,

$$t\mathbf{y}^t = \sum_{r=MH^{\frac{1+\varepsilon}{2}}+1}^{MH^{\frac{1+\varepsilon}{2}}+t} \mathbb{1}_{a_r}.$$

Eq. (15) implies for every $\ell = 1, ..., L$,

$$\sum_{r=1}^{MH^{\frac{1+\varepsilon}{2}}+t} (\mathbf{C}_\ell \mathbb{1}_{a_r} - \mathbf{w}_\ell) = (\mathbf{C}_\ell \mathbf{m} - \mathbf{w}_\ell) MH^{\frac{1+\varepsilon}{2}} + t(\mathbf{C}_\ell \mathbf{y}^t - \mathbf{w}_\ell) \quad \leq \tag{16}$$

$$(\mathbf{C}_\ell \mathbf{m} - \mathbf{w}_\ell) MH^{\frac{1+\varepsilon}{2}} + \frac{\mathbf{w}_\ell - \mathbf{C}_\ell \mathbf{m}}{2} MH^{\frac{1+\varepsilon}{2}} = \frac{\mathbf{C}_\ell \mathbf{m} - \mathbf{w}_\ell}{2} MH^{\frac{1+\varepsilon}{2}} \quad < \quad 0.$$

In other words,

$$d(\mathbf{x}^{t'}) \geq \frac{d(\mathbf{m})}{2} \frac{MH^{\frac{1+\varepsilon}{2}}}{t'} > 0, \tag{17}$$

where $t' = MH^{\frac{1+\varepsilon}{2}} + t$. This shows that all constraints are indeed kept at time t'. □

Proof of Lemma 4. In order to violate Eq. (13), $\sum_{r=MH^{\frac{1+\varepsilon}{2}}+1}^{MH^{\frac{1+\varepsilon}{2}}+t} (\mathbf{C}_\ell \mathbb{1}_{a_r} - \mathbf{w}_\ell)$ needs to be strictly greater than $\frac{d(\mathbf{m})}{2} MH^{\frac{1+\varepsilon}{2}}$. By Lemma 2, recalling that $t \leq MH$,

$$P\left(\sum_{r=1}^{t} (\mathbb{1}_{a_r} \cdot \mathbf{C}_\ell - w_\ell) > \frac{d(\mathbf{m})}{2} MH^{\frac{1+\varepsilon}{2}}, \forall \ell = 1, 2, ..., L \right) \leq \tag{18}$$

$$P\left(\sum_{r=1}^{t} (\mathbb{1}_{a_r} \cdot \mathbf{C}_\ell - w_\ell) > \frac{d(\mathbf{m})}{2} M^{\frac{1-\varepsilon}{2}} t^{\frac{1+\varepsilon}{2}}, \forall \ell = 1, 2, ..., L \right) \leq L \exp\left(\frac{-t^\varepsilon (\frac{d(\mathbf{m})}{2} M^{\frac{1-\varepsilon}{2}})^2}{2} \right).$$

Denote

$$c := \begin{cases} \frac{d(\mathbf{m})}{\max_{\ell, x \in \Delta(A)} 2(\mathbf{C}_\ell x - \mathbf{w}_\ell)}, & \text{if } \max_{\ell, x \in \Delta(A)} 2(\mathbf{C}_\ell x - \mathbf{w}_\ell) > 0 \\ 1 & \text{otherwise.} \end{cases}$$

Note that during the first $cMH^{\frac{1+\varepsilon}{2}}$ rounds of the second block it is impossible to violate Eq. (13). We therefore obtain from Eq. (18) and the union bound, that the probability of violating Eq. (13) (and therefore stopping playing τ) at any time during the second block is bounded above by

$$\sum_{t=cMH^{\frac{1+\varepsilon}{2}}}^{MH} L \exp\left(\frac{-t^\varepsilon (\frac{d(\mathbf{m})}{2} M^{\frac{1-\varepsilon}{2}})^2}{2} \right) \leq M \cdot H \cdot L \exp\left(\frac{-(cMH^{\frac{1+\varepsilon}{2}})^\varepsilon (\frac{d(\mathbf{m})}{2} M^{\frac{1-\varepsilon}{2}})^2}{2} \right),$$

which goes to zero with H. □

Proof of Lemma 5. Suppose that the deterministic sequence $(a_1, ..., a_M)$ has been played repeatedly until stopped for s periods. That is, $t'' = t' + s$ (as in Eq. (17), t' is the first time τ has been stopped). Due to Eq. (11),

$$d(\mathbf{x}^{t''}) \geq \frac{t'}{t'+s} d(\mathbf{x}^{\mathbf{t}'}) + \frac{s}{t'+s} d(\mathbf{x}^{[t'+1, t'+s]}).$$

When s is a multiple of M (meaning that a few cycles of playing $(a_1, ..., a_M)$ have been completed), $d(\mathbf{x}^{[t'+1,t'+s]}) = d(\mathbf{m})$. Due to Eq. (17) and $d(\mathbf{x}^{t'}) < \frac{d(\mathbf{m})}{2}\frac{MH^{\frac{1+\varepsilon}{2}}}{t'} < d(\mathbf{m})$, we have $d(\mathbf{x}^{t''}) > d(\mathbf{x}^{\mathbf{t}'})$. In other words, $d(\mathbf{x}^{t''})$ cannot get lower than $d(\mathbf{x}^{\mathbf{t}'}) > 0$.

When s is not a multiple of M, t'' is in the middle of the cycle. In this case, $d(\mathbf{x}^{t''}) \geq \frac{t'}{t'+s}d(\mathbf{x}^{\mathbf{t}'}) - \frac{s}{t'+s}c_1$, where $c_1 = \max_{\mathbf{x}\in\Delta(A)} |d(\mathbf{x})| \geq 0$, and $s < M$. Thus, again due to Eq. (17), if H is chosen to be much larger than c_1, we have

$$d(\mathbf{x}^{t''}) > \frac{t'}{t'+s}d(\mathbf{x}^{\mathbf{t}'}) - \frac{s}{t'+s}c_1 \geq \frac{t'}{t'+s}\frac{d(\mathbf{m})}{2}\frac{MH^{\frac{1+\varepsilon}{2}}}{t'} - \frac{s}{t'+s}c_1.$$

If $\frac{d(\mathbf{m})}{2}MH^{\frac{1+\varepsilon}{2}} > Mc_1$ (i.e., H is large enough), we have $d(\mathbf{x}^{t''}) > 0$, which completes the proof. □

Proof of Lemma 6. Note that H is a constant, the number of periods preceding the N-th interval is $MH + MH^2 + ... + MH^{N-1} = \Omega(MH^{N-1}) = \Omega(H^{N\pm c})$ for any constant $c > 0$. We can therefore state the following.

(a) The regret up to the end of the N-th interval is

$$\mathbb{E}(R(\mathbf{C}, \mathbf{w})(\mathbf{a'}_1, ..., \mathbf{a'}_T; \overrightarrow{b})) \geq -\frac{O(MH^{\frac{1+\varepsilon}{2}}) + ... + O(MH^{\frac{N(1+\varepsilon)}{2}})}{M(H + H^2 + ... + H^N)} = -O\left(\frac{H^{\frac{N(1+\varepsilon)}{2}}}{H^N}\right), \tag{19}$$

where $T = MH^N$. The regret bound is therefore, $-O(\frac{T^{\frac{1+\varepsilon}{2}}}{T}) = -O(T^{-\frac{1-\varepsilon}{2}})$.

(b) For any t in the first block of the N-th interval, i.e., $t \in [t_{N-1}+1, t_{N-1}+MH^{\frac{N(1+\varepsilon)}{2}}]$, the regret is bounded by

$$\mathbb{E}(R(\mathbf{C}, \mathbf{w})(\mathbf{a'}_1, ..., \mathbf{a'}_t; \overrightarrow{b})) \geq -\left(\frac{O((H^{N-1})^{\frac{1+\varepsilon}{2}})}{t} + O(\frac{t - t_{N-1}}{t})\right) \tag{20}$$

$$\geq -\left(\frac{O((H^N)^{\frac{1+\varepsilon}{2}}) + O(H^{\frac{N(1+\varepsilon)}{2}})}{t}\right) = -O\left(\frac{t^{\frac{1+\varepsilon}{2}}}{t}\right), \tag{21}$$

where the term $-O(\frac{t-t_{N-1}}{t})$ is the bound on total regret during the first block. The inequality is due to the fact that $t - t_{N-1} \leq MH^{\frac{N(1+\varepsilon)}{2}}$ and $t = O(H^N)$.

(c) For any time t in the second and third block of the N-th interval, i.e., $t \in$

$[t_{N-1} + MH^{\frac{N(1+\varepsilon)}{2}} + 1, t_N]$, the regret up to time t is bounded by

$$\begin{aligned}
& \mathbb{E}(R(\mathbf{C}, \mathbf{w})(\mathbf{a'}_1, ..., \mathbf{a'}_t; \overrightarrow{b})) && (22)\\
\geq\ & -\left(\frac{O((H^{N-1})^{\frac{1+\varepsilon}{2}})}{t} + \frac{MH^{\frac{N(1+\varepsilon)}{2}}}{t} + \frac{O((t - (t_{N-1} + MH^{\frac{N(1+\varepsilon)}{2}}))^{\frac{1+\varepsilon}{2}})}{t}\right) && (23)\\
\geq\ & -\left(\frac{O(H^{\frac{N(1+\varepsilon)}{2}}) + O(t^{\frac{1+\varepsilon}{2}})}{t}\right) \geq -O\left(\frac{t^{\frac{1+\varepsilon}{2}}}{t}\right), && (24)
\end{aligned}$$

which follows from the fact that $t = O(H^N)$. Thus, the regret bound holds not only at the end of each interval, but also in all times. □