Online decision making with history-average dependent costs

Vijeth Hebbar VHEBBAR2@ILLINOIS.EDU

Cédric Langbort LANGBORT@ILLINOIS.EDU

Coordinated Science Lab, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA.

Editors: A. Abate, K. Margellos, A. Papachristodoulou

Abstract

In many online sequential decision-making scenarios, a learner's choices affect not just their current costs but also the future ones. In this work, we look at one particular case of such a situation where the costs depend on the time average of past decisions over a history horizon. We first recast this problem with history dependent costs as a problem of decision making under stage-wise constraints. To tackle this, we then propose the novel Follow-The-Adaptively-Regularized-Leader (FTARL) algorithm. Our innovative algorithm incorporates *adaptive regularizers* that depend explicitly on past decisions, allowing us to enforce stage-wise constraints while simultaneously enabling us to establish tight regret bounds. We also discuss the implications of the length of history horizon on design of no-regret algorithms for our problem and present impossibility results when it is the full learning horizon.

Keywords: Sequential Decision Making, Online Optimization with Memory, Online Learning with Constraints.

1. Introduction

In the classical online optimization framework, one seeks to study a multi-stage decision making process where a learner faces a series of cost functions $\{l^t\}_{t=1}^T$ over a time horizon T. The learner has no prior knowledge about the sequence of cost functions they will face but has to make decisions x^t at stage t using only the information about the cost functions faced in past stages. They then incur a loss $l^t(x^t)$ for their decision. The goal of the learner is to make decisions that result in low *regret*, viz. the difference between the learner's cumulative cost and the cost of the best-in-hindsight decision. This setting has been widely studied and has been successfully applied in domains ranging from portfolio management [Cover (1991); Blum and Kalai (1997)] and auctioning [Bar-Yossef et al. (2002)] in economics to network routing [Awerbuch and Kleinberg (2008)] and control [Abbasi-Yadkori and Szepesvári (2011)] in engineering. Readers are referred to Hazan (2022) and Cesa-Bianchi and Lugosi (2006) for an extensive survey.

Deviating from this traditional setup, we seek to study the problem where the cost depends not only on the current action but also on the past actions. While such history dependence may in general show up in any number of ways, we restrict ourselves to the special case where costs (or payoffs) depend solely on the *time average* of *finitely* many past actions. This is motivated, for example, by the following scenarios.

Consider a manufacturing facility that selects a product from its range to produce at each time step, with the aim of producing the most profitable item. For ease of exposition, let us suppose that the facility produces one unit of this chosen good per time step. Now we make the assumption

that the manufacturer must 'commit' to produce the chosen good for a fixed term of H steps. Such 'commitment' can encapsulate a myriad of scenarios where decisions have long-term effect, for example, in training workers for a specific production line, who are then employed on an H-period contract to produce that good. Regardless of the specific way in which this commitment manifests itself, it implies that the production initiated at a time step continues for the next H time steps (in parallel with the production initiated in these later steps).

Therefore, the proportion of each product in the total production at any time is simply the time-average of the production initiated in the past H time steps. The manufacturer's reward – quantified as profit per unit produced – in a time step is tied to this average result of their past decisions. While the manufacturer may not know what profits they will enjoy for each product when initiating production, they can learn from historical data, and so the question arises

Q1: *How would a learning agent make decisions when the profit they receive depends on the long-term average of their choices?*

Let us now look at another scenario that provides an alternate view on the problem of learning with history average dependent payoffs. Consider a media outlet generating content with a goal of achieving high viewership, while not knowing a priori what content is most desirable to consumers. Any realistic consumer with a bounded memory will base their viewership decision not just on the current content, but on the finite history of outputs by the media outlet. Indeed, a media house known primarily for coverage of financial news announcing they will feature an hour long interview with the new World Chess Champion will not attract as many viewers as a sports news channel doing the same. Thus, the media outlet's reward – in this case their viewership – depends on their *reputation* i.e. the time-average of their past behaviour.

Viewing *reputation* as an averaged *state* induced by the sequence of actions taken by the learner, the learner can be seen as trying to learn what *reputation* to establish to enjoy high reward. In this view, the decision of the learner at a stage is constrained by their decision in past stages. Indeed the reputation of the media house in the eyes of a consumer with memory cannot be arbitrarily changed within short time frames. So, we ask the question

Q2: *How would a learning agent make decisions when the decision across stages are coupled by constraints (in particular, ones arising out of the averaging process)*?

These examples underline the dual aspects of historical influence: the enduring impact of past decisions versus the limitations imposed by past decisions on present choices. This duality in viewpoint is precisely captured in the two equivalent questions **Q1** and **Q2** posed above. Finally, another natural question that arises from the consideration of the media outlet illustration above is how the length of the consumer's memory affects the media outlet's ability to learn about them. Simply put, if consumers remember everything, the media outlet might never overcome a 'bad' reputation caused by its earlier decisions. So we then ask

Q3: *How should the length of the past horizon that affects current loss scale with T to allow the learner to perform well?*

1.1. Our Contributions and Paper Roadmap

In this paper, we recast the problem of online sequential decision making with history-dependent payoffs as a problem of online optimization with stage-wise constraints. Specifically, in our work the dependence on history emerges solely through averaging and the cost in a stage relies explicitly on the average of past decisions. We also restrict ourselves to the class of online optimization

problems with linear stage costs and decisions being picked from a finite-dimensional simplex. In other words, we consider the widely studied setup of *Prediction-from-Expert-Advice* with the caveat that costs depend on the average of past decisions rather than the current decision alone. Section 2 details the formalization of our framework.

Drawing from the well-known Follow-The-Regularized-Leader (FTRL) [Shalev-Shwartz and Singer (2007)] class of algorithms, we propose the novel Follow-The-Adaptively-Regularized-Leader (FTARL) algorithm in Section 3. While the traditional role played by the regularizer function in FTRL style algorithms is to avoid over-fitting while picking decisions [Kalai and Vempala (2005); Shalev-Shwartz et al. (2012)], we curate our regularizer to also ensure our constraints are satisfied. This is achieved by allowing the regularizer to depend on past cost functions as well as past decisions, earning it the moniker of an *adaptive regularizer*. The usage of such *adaptive regularizers* is not a novel idea [McMahan (2017)] and they have been employed to obtain stronger regret bounds [Orabona (2019)] or to extend the applicability of certain online learning algorithms [Hazan et al. (2007)]. But to the best of our knowledge, this is the first such usage of history-dependent regularizers to enforce constraints over decisions across time steps. These regularizers provide a direct approach towards establishing regret bounds in this context, as we show in Theorem 3. The usage of such regularizers also offers a novel perspective on online learning with history-dependent costs and highlights the duality of question **Q1** and **Q2**.

With the aim of answering **Q3**, we first show in Section 4 that when the stage costs depend on the full history of past decisions, no algorithm can achieve sublinear regret (Claim 5). In the media outlet illustration from Section 1, this implies that if consumers have perfect memory, the media outlet cannot hope to achieve sublinear regret with respect to the viewership count. Using the tools developed in Section 3, we then show that when this history dependence is restricted to a shorter horizon, specifically when $H \in o(T)$, we can achieve $\mathcal{O}(\sqrt{TH})$ regret (Theorem 6). In other words, when consumers are forgetful (i.e., have $o(T)$ memory), the media outlet can still learn what content to output with sublinear regret. A simple extension of our approach also shows that our approach is H-agnostic so long as we have an upperbound $H < \Theta \in o(T)$. In this case, we establish a $\mathcal{O}(\sqrt{T\Theta})$ regret bound (Corollary 7).

1.2. Related Work

Our problem forms an instance of online learning with history-dependent costs [Arora et al. (2012); Cesa-Bianchi et al. (2013)] and is closest in spirit to the framework of Online Convex Optimization with Memory (OCO-M) as analyzed by Anava et al. (2015). In their work, Anava et al. (2015) consider adversarially generated convex cost functions that can depend arbitrarily on a finite number of past decisions. Of the two approaches they present, the one better suited for the Expert Advice type of problem yields a regret bound of $\mathcal{O}(\sqrt{HT\log(T)})$ in our case. By explicitly considering the nature of history dependence that shows up in our problem, our method offers an improvement by a logarithmic factor over theirs. A related line of work [Geulen et al. (2010); György and Neu (2014)] studies the Experts Advice problem when the cost of each expert also depends on their past actions. The challenge that arises then is that the cost incurred by the learner may be different from that of the picked expert as the two may have taken different actions in the past stages. In contrast, the experts in our setup have *memoryless* costs, and the learner's cost alone depends on past actions.

Online optimization with history dependent costs has garnered increased attention amongst the control community in recent years with an important application area being Online Linear Control

(OLC) [Cohen et al. (2018); Abbasi-Yadkori and Szepesvári (2011)]. In this framework, the cost at each stage depends on the state of a linear dynamical system and so implicitly depends on all past actions. The primary goal here is to arrive at a linear feedback controller (or some modified version of it [Agarwal et al. (2019)]) at every stage to ensure low regret relative to the best-in-hindsight controller. On the other hand, in our problem, the meaningful notion of regret compares the performance of our algorithm with the performance of the best static action.

Bearing question **Q2** in mind, there are multiple lines of research that try to incorporate constraints into an online optimization framework. Some works consider constraints that are adversarially generated [Kveton et al. (2008); Mannor et al. (2009)], while some other consider a single long-term constraint connecting decisions across the learning horizon [Wang et al. (2021); Altschuler and Talwar (2018)]. In contrast, we consider a setup with stage-wise constraints that couple the decisions at each step with ones in the past. These constraints are known a-priori and we seek to design a decision making approach that explicitly accounts for them. In a similar spirit, Badiei et al. (2015) consider the problem of online optimization with ramp constraints i.e. known bounds on the magnitude of change in the decision across a step. To address this challenge, they consider a finite look ahead window on future costs. In contrast, we stick with the classical assumption in online optimization frameworks where only historical data is available when making decisions.

2. Problem Setup

Let us now formalize our problem statement. Let $\{v^t\}_{t=1}^T \in \mathbf{\Delta}_n$ denote the sequence of decisions made by the learner over a time horizon of length T. Here $\mathbf{\Delta}_n$ denotes the n-dimensional simplex and v_i^t corresponds to the weight given to action $i \in [n]$ at time t. In the case of Prediction from Expert Advice, v_i^t has the special interpretation of being the probability of choosing action i at time t. We then define the *time-averaged decision* x^t as

$$x^t = h^t(v^1, \ldots, v^t) \triangleq \begin{cases} \frac{1}{t}\sum_{\tau=1}^{t} v^\tau & t < H \\ \frac{1}{H}\sum_{\tau=t-H+1}^{t} v^\tau & t \geq H. \end{cases} \tag{1}$$

Thus, x_i^t corresponds to the average weight given to action i over the past horizon of length (at-most) H. We can view the decision v^t as an *input* and the time averaged decision x^t as the *state* at time t. This view makes explicit the idea that x^t is not independent of the past, but is instead generated through an update process. Formalizing this very idea we have

$$x^t = y^{t-1} + \beta^t v^t \quad \forall t \geq 1 \tag{2}$$

where $\beta^t = \frac{1}{\min\{t,H\}}$ and y^t is defined as

$$y^t \triangleq \begin{cases} \frac{1}{t+1}\sum_{\tau=1}^{t} v^\tau & 1 < t < H \\ \frac{1}{H}\sum_{\tau=t-H+2}^{t} v^t & t \geq H. \end{cases}$$

with $y^0 \triangleq \mathbf{0}$. This switch in the view from (1) to (2) captures precisely the change in viewpoint from **Q1** to **Q2** as x^t can now be viewed as being connected to the past (captured through y^t) by a constraint.

Let $\{g^t\}_{t=1}^T \subset \mathbb{R}^n_{\leq 0}$ be the sequence of non-positive cost vectors faced by the learner. These vectors may be generated adversarially but, if so, we assume that the adversary is oblivious, i.e. the

cost vectors are generated in advance before any learner decisions are revealed. The cost incurred by the learner at time t in our model is then simply $\langle g^t, x^t \rangle$. Naturally, the goal of the learner is to incur a low total cost $\sum_{t=1}^{T} \langle g^t, x^t \rangle$.

There are two notions of regret we can consider. First, the standard notion of regret generally considered in the *memoryless* case (i.e. in the event we could choose the decision x^t at every step independent of the past) that is defined as

$$\mathcal{R}_T = \sum_{t=1}^{T} \langle g^t, x^t \rangle - \min_{x \in \mathbf{\Delta}_n} \langle G^T, x \rangle \text{ where } G^t \triangleq \sum_{\tau=1}^{t} g^\tau. \tag{3}$$

At a first glance, this appears like a very strong notion of regret for our setup since our decisions are connected by the constraints in (2). The second weaker notion of regret – one that is routinely employed in OCO-M literature [Anava et al. (2015); Arora et al. (2012)] – is that of *policy regret* and is defined for our setup as

$$\mathcal{R}_{T,\text{Pol}} \triangleq \sum_{t=1}^{T} \langle g^t, x^t \rangle - \min_{v \in \mathbf{\Delta}_n} \sum_{t=1}^{T} \langle g^t, h^t(v, \ldots, v) \rangle. \tag{4}$$

However, in the special case where dependence on past decisions is captured through an averaging process, we can state the following

Lemma 1 *When decisions $\{x^t\}_{t=1}^T$ satisfy the form in (1), $\mathcal{R}_T = \mathcal{R}_{T,Pol}$.*

Proof This follows simply from noting that $h^t(v, \ldots, v) = v$ from the definition in (1). ■

Lemma 1 highlights the crucial role played by our assumption on the averaging nature of history-dependence in allowing us to define regret as in (3). Indeed in the absence of such an averaging process, it is challenging to define meaningful notions of regret when costs depend on past decisions [Arora et al. (2012)]. Taking viewpoint **Q2**, a further consequence of being able to work with the regret $\mathcal{R}_T$ is that any sublinear regret bounds mean that despite the presence of constraints, in the long run we are able achieve the performance of the best static action under *no* constraints (or equivalently, *no* history dependence).

We make an additional note that when an algorithm picks decisions $\{x^t\}_{t=1}^T$ in a stochastic manner, the appropriate metric to evaluate its performance is *expected regret*. This is defined simply by taking expectations over the definition in (3) and they are taken with respect to the randomness in the algorithm. For the purposes of brevity, in this paper we will henceforth refer to 'expected regret' simply as 'regret' and assume it is understood from context which notion we are employing.

Our goal in this paper is to design an algorithm that allows the learner to generate decisions $\{x^t\}_{t=1}^T$ (that take the form in (1)) while ensuring that regret defined in (3) grows sub-linearly with T. To this end we will first develop some theory in the following section, which will motivate our algorithm and help us in analyzing its performance.

3. Going Beyond Follow-the-Leader

3.1. Preliminaries on Follow-The-Leader type algorithms

Let $\{l^t(\cdot)\}_{t=1}^T$ denote the sequence of the cost functions – mapping decision set $\mathcal{X} \subset \mathbb{R}^n$ to $\mathbb{R}$ – faced by the learner over a horizon of length T. We define the sub-sequence of cost functions

and learner's actions until time t as the history $\mathcal{H}^t = \{(l^\tau, x^\tau)\}_{\tau=1}^t$. In line with the standard online optimization framework, we assume that when making decision x^{t+1} the learner only has access to $\mathcal{H}^t$. Let us now look more closely at one class of algorithms that are routinely applied to solve online optimization problems: Follow-The-Leader (FTL) type algorithms. Consider first the canonical FTL algorithm that picks the decision x^t at time t such that

$$x^t \in \arg\min_{x\in\mathcal{X}} \sum_{\tau=1}^{t-1} l^\tau(x) \; \forall t > 1, \quad x^1 \in \mathcal{X}. \tag{5}$$

Note that we have not described either the class of cost functions or the decision set $\mathcal{X}$. Throughout this section, we only assume that the minimum in (5) exists and in the event of multiple minimizers, one is picked arbitrarily. For this algorithm we have the following well-known result bounding the regret of the FTL algorithm [Cesa-Bianchi and Lugosi (2006); Kalai and Vempala (2005)].

Theorem 2 *When $\{x^t\}$ is generated according to (5),*

$$\sum_{t=1}^{T} l^t(x^t) - \min_{x\in\mathcal{X}} \sum_{t=1}^{T} l^t(x) \leq \sum_{t=1}^{T} l^t(x^t) - l^t(x^{t+1}). \tag{6}$$

While the FTL algorithm captured in (5) leads to sublinear regret in some well structured online optimization problems, it can lead to linear regret even in some simple cases like when $l^t(\cdot)$ is linear [Shalev-Shwartz et al. (2012)]. Consequently, closely related methods like Follow-the-Regularized-Leader [Shalev-Shwartz and Singer (2007)] and Follow-The-Perturbed-Leader [Kalai and Vempala (2005); Hannan (1957)] were developed to ensure sublinear regret in a wide variety of online learning setups. Taking a page out of this book we propose the *Follow-The-Adaptively-Regularized-Leader* (FTARL) algorithm.

3.2. Follow-The-Adaptively-Regularized-Leader Algorithm

Going beyond existing methods, our method picks a regularizer function that explicitly depends on the history $\mathcal{H}^t$ faced by the learner, hence the term *adaptive*. Mathematically, the learner makes decisions $\{x^t\}_1^T$ according to

$$x^{t+1} \in \arg\min_{x\in\mathcal{X}} \left(R^t(x, \mathcal{H}^t) + \sum_{\tau=1}^{t} l^\tau(x) \right) \; \forall t \geq 1, \tag{7a}$$

$$x^1 \in \arg\min_{x\in\mathcal{X}} R^0(x). \tag{7b}$$

While we introduce additionally restrictions on the nature of the regularizers $R^t(\cdot,\cdot)$ in later sections, for now we only assume that the regularizer ensures the existence of minimizers in (7). In (7) the second argument of the regularizer $R^t(\cdot,\cdot)$ highlights the dependence on history, but for brevity's sake we drop the argument henceforth and assume this history dependence is implicit. We now present a regret bound theorem for the FTARL algorithm presented in (7).

Theorem 3 ***(FTARL Regret Bound)*** *Let $\{x^t\}_1^T$ be picked according to the algorithm presented in (7), then for any $x' \in \mathcal{X}$*

$$\sum_{t=1}^{T} l^t(x^t) - l^t(x') \leq \sum_{t=1}^{T} l^t(x^t) - l^t(x^{t+1}) + \sum_{t=0}^{T-1} \left(R^t(x^{t+2}) - R^t(x^{t+1}) \right) + R^T(x') - R^T(x^{T+1}).$$

Indeed picking x' as the best-in-hindsight strategy gives us a regret bound.

Proof *Let us define $\tilde{l}^0(\cdot) = R^0(\cdot)$ and $\tilde{l}^t(\cdot) = l^t(\cdot) + R^t(\cdot) - R^{t-1}(\cdot)$ for $t \geq 1$. Then it is easy to see that running the FTARL algorithm in (7) is equivalent to running the FTL algorithm from (5) with cost functions $\tilde{l}$ at every time step starting from $t = 0$. Then as a direct corollary of Theorem 2 we have*

$$-\sum_{t=0}^{T} \tilde{l}^t(x') \leq -\min_{x \in \mathcal{X}} \sum_{t=0}^{T} \tilde{l}^t(x) \leq -\sum_{t=0}^{T} \tilde{l}^t(x^{t+1}) \tag{8}$$

The remainder of the proof is then an algebraic exercise involving expanding expressions for $\tilde{l}_t(\cdot)$ in both LHS and RHS of (8), rearranging terms and finally, adding $\sum_{t=1}^{T} l^t(x^t)$ to both sides. ■

By allowing the regularizer in (7) to also depend on the past decisions, we can use it to explicitly enforce constraints that relate past decisions to current decisions. In the following section, we will illustrate this ability by designing the regularizer function so that the decisions x^t taken by our FTARL algorithm at time t takes form described in (1) from Section 2. We will then employ Theorem 3 to analyze the performance of resulting algorithm.

4. FTARL for Average Dependent Costs

Let us begin by restating our objective as outlined at the end of Section 2. In doing so, we will stick with the viewpoint **Q2**. We want our learning agent to establish states $\{x^t\}_{t=1}^T$ with the aim of incurring regret, as defined in (3), that grows sub-linearly with T. Additionally, any algorithm $\mathcal{A}$ that the learner employs to achieve this goal must satisfy two properties.

1. First, it must ensure that x^t respects the relation in (2) at every stage.
2. Second, it must only rely on information $\mathcal{H}^{t-1}$ that is available to the learner at time t.

We will design such an algorithm based on the FTARL algorithm framework introduced in Section 3. Our first step then will be to design adaptive regularizers $\{R^t(\cdot)\}_{t=0}^T$ and for the purpose of this work, we will choose to work with randomized regularizers.

4.1. Designing the Adaptive Regularizer

Let $Z \in \mathbb{R}_+^n$ be a random vector with every element being picked i.i.d from an exponential distribution i.e. $Z_i \overset{i.i.d}{\sim} \exp(\epsilon)$. We then define the random sequence

$$v_*^t \in \arg\min_{v \in \mathbf{\Delta}_n} \langle G^{t-1} - Z, v \rangle \quad \forall t \geq 1 \tag{9}$$

where $G^0 = \mathbf{0}_n \in \mathbb{R}^n$. Owing to the stochasticity in Z, v_*^t is a random variable and is uniquely defined almost surely. In the event of non-uniqueness, we assume that one of the minimizers is picked arbitrarily. We then define our regularizer function as

$$R^0(x) = -\langle Z, x \rangle \tag{10a}$$

$$R^t(x) = \delta\left(\frac{1}{2}\|x - y^t\|_2^2 - \beta^{t+1}\langle v_*^{t+1}, x\rangle\right) - \langle G^t, x\rangle \quad \forall t \geq 1 \tag{10b}$$

for all $t \geq 1$ and for some $\delta > 0$. First, note that regularizers defined in (10) guarantee the existence of minimizers in (7) when $\mathcal{X} = \mathbf{\Delta}_n$ and $l^t(\cdot) = \langle g^t, \cdot \rangle$ and so, the FTARL algorithm is well-defined. Secondly, note that these regularizers depend only on the history $\mathcal{H}^t$ and so the causality property of the corresponding FTARL algorithm is satisfied. The dependence on past decisions is captured through y^t and the dependence on the cost functions is captured both explicitly, through G^t, and implicitly, through v_*^t. Finally, we propose

Lemma 4 *For all $t \geq 1$, the decisions $\{x^t\}_{t=1}^T$ induced by the FTARL algorithm in (7) (with regularizers as defined in (10)), satisfy the relation in (2).*

For proofs of this and other results readers are referred to Hebbar and Langbort (2023). A direct consequence of Lemma 4 is that taking decisions according to (7) is equivalent to picking $\{v_t\}_{t=1}^T$ according to (9) followed by generating decisions $\{x^t\}_{t=1}^T$ through the update step in (2). Thus, in practice, our algorithm can be implemented without any explicit consideration of the regularizers in (10), as highlighted in Algorithm 1. But, as we will see in Section 4.3, these regularizers play a critical role in our analytic approach for establishing regret bounds.

Algorithm 1 FTARL for History-Average Dependent Costs

Require: Learning rate $\epsilon > 0$

1: Draw perturbation $Z_i \overset{i.i.d}{\sim} \exp(\epsilon)$
2: **for** each round $t = 1, 2, \ldots, T$ **do**
3: Pick decision $v^t = \arg\min_{v \in \mathbf{\Delta}_n} \langle G^{t-1} - Z, v \rangle$
4: Update state $x^t = y^{t-1} + \beta^t v^t$
5: Observe cost vector g^t; Incur the loss $\langle g^t, x_t \rangle$
6: **end for**

4.2. The Challenge of Horizon Length

Before arriving at regret bounds for the proposed algorithm, we first present a result that highlights the limitations faced by a learner in our problem setup. We will constructively show that no algorithm can guarantee regret that grows sublinearly with H. In other words, if the horizon H over which decision averaging takes place scales linearly with T, we have no hope of designing an algorithm that achieves an $o(T)$ regret.

Claim 5 *Let $H \leq 0.8T$ be an multiple of 4 and define $T_s = T - \frac{H}{4}$. Consider the sequence of two-dimensional cost vectors $\{g^t\}_{t=1}^T$ such that*

$$g^t = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ \forall t \leq T_s, \quad g^t = \begin{cases} [-1 \quad 0]^T \ \ \forall t > T_s & w.p.\ 1/2 \\ [0 \quad -1]^T \ \ \forall t > T_s & w.p.\ 1/2 \end{cases}$$

Then for any (possibly randomized) algorithm $\mathcal{A}$ that generates $\{x^t\}_{t=1}^T$ in accordance with the process in (1) we have $\mathbb{E}[\mathcal{R}_T] \geq H/32$ where the expectation is taken over the randomness in the cost function (as well the randomness, if any, in the algorithm).

Proof Effectively, at $t = T_s$, a fair coin is flipped once and the cost vector for all subsequent time steps is determined based on the result of this flip. Let us denote by random variable θ the outcome

of the coin flip. Indeed, if the learner could see this coin flip before hand they could easily achieve the best-in-hindsight cost of

$$\min_{x \in \mathbf{\Delta}_2} \langle G^T, x \rangle = \frac{-H}{4} \tag{11}$$

by playing the best response to the costs incurred after the coin flip for all T stages. Recalling that $x_1^{T_s} + x_2^{T_s} = 1$, let us first assume that the learning algorithm generates $x_1^{T_s} \leq 1/2$. With this assumption, from (1), $\forall t > T_s$ we have

$$x_1^t = x_1^{T_s} + \frac{1}{H}\left(\sum_{\tau=T_s+1}^{t} v_1^\tau - \sum_{\tau=T_s-H+1}^{t-H} v_1^t \right) \leq \frac{1}{2} + \frac{(t-T_s)}{H} \leq \frac{3}{4}. \tag{12}$$

Then taking expectations both over the coin flip and the randomization in the learning algorithm, the expected cost incurred by the learner after the coin flip is

$$\begin{aligned}\mathbb{E}_{\theta,\mathcal{A}}\left[\sum_{t=T_s+1}^{T} \langle g^t, x^t \rangle \middle| x_1^{T_s} \leq 0.5 \right] &= \frac{-1}{2} \sum_{T_s+1}^{T} \left(\mathbb{E}_{\mathcal{A}}[x_1^t | x_1^{T_s} \leq 0.5] + \mathbb{E}_{\mathcal{A}}[x_2^t | x_1^{T_s} \leq 0.5] \right) \\ &\overset{(a)}{\geq} \frac{7}{8}(T_s - T) = \frac{-7H}{32}.\end{aligned}$$

where the inequality (a) results from (12) and from noting that $x_2^t \leq 1$ for all t. An identical lower bound can be obtained if we instead assumed $x_2^{T_s} < 1/2$ and consequently, by law of total expectation we have

$$\mathbb{E}_{\theta,\mathcal{A}}\left[\sum_{t=1}^{T} \langle g^t, x^t \rangle \right] \geq \frac{-7H}{32}$$

Comparing this with the cost of best-in-hindsight action from (11) gives us the required lower bound on regret. ■

Note that in Claim 5, the assumption $H \leq 0.8T$ is made only for technical reasons and it is possible to generate similar examples where $\Omega(H)$ regret is guaranteed for any $H \leq T$. Nevertheless, this claim lends us the insight that when $H \in \Omega(T)$ the learner cannot hope to achieve regret sublinear in T. With the hope of arriving at sublinear regret algorithms for our problem, the natural regime to explore then is when H scales sub-linearly with T.

In concluding this section, we make a note comparing the lower bound in Claim 5 to similar impossibility results that exist in the literature. While we assume that stage cost functions depend solely on the time average of past learner decisions, Arora et al. (2012) consider a setup where an *adaptive adversary* can generate cost functions that depend arbitrarily on past decisions. To show that it is impossible to learn when facing such an adversary they then construct specific forms of cost functions. Effectively, our result shows that such impossibility still holds against a *weaker* adversary, one that is restricted to devise cost functions that depend exclusively on the average of past decisions. In a similar spirit to the lower bounds obtained by Cesa-Bianchi et al. (2013) for the setting when the adversary can impose switching costs, our bound considers a setup with a type of *deterministically adaptive adversary* and obtain lower bounds for it.

4.3. Regret Bounds

Inline with the discussion in the previous section, henceforth in this paper, we will assume $H \in o(T)$. We now present the main result of our work

Theorem 6 *Running algorithm (7) with regularizers defined in (10) (or, equivalently, Algorithm 1) with* $\epsilon = \sqrt{\frac{4(\log(n)+1)}{M^2(T-H)(2+H)}}$ *results in an expected regret*

$$\mathbb{E}[\mathcal{R}_T] \leq 5MH + 4M\sqrt{(T-H)(H+2)(\log(n)+1)},$$

where the expectation is taken over the distribution of Z *employed in (9) and* M *is a bound on* $\|g^t\|_\infty$.

Some comments on this result are in order. First, note that our regret bound is linear in H. This was expected and agrees with the $\Omega(H)$ regret we obtained for the constructive example in Section 4.2. Accordingly, we began this section with the underlying assumption that $H \in o(T)$ which indicates that our regret bound is $\mathcal{O}(M\sqrt{TH\log(n)})$. Indeed when $H = 1$, our problem reduces to the Expert Advice problem and we recover the well established and tight $\mathcal{O}(M\sqrt{T\log(n)})$ regret bound [Cesa-Bianchi and Lugosi (2006)].

Recall that the notion of regret we consider compares the performance of our algorithm with that of the best-in-hindsight decision under no history dependence as highlighted in Section 2. The bounds in Theorem 6 then indicate that in reconciling with this history dependence the penalty we pay appears as a multiplicative factor that scales as $\sqrt{H}$.

Finally, note that in Theorem 6, the value of ϵ we picked depended explicitly on the H which translates to our algorithm (7) also depending on H. But knowledge of H may be an unrealistic assumption in many cases - in the illustrative scenario from Section 1 for instance, the media house may not have an estimate of the length of memory amongst their consumers. Fortunately, it suffices to know a suitable upper bound Θ on H to make the following

Corollary 7 *Running algorithm (7) with regularizers defined in (10) (or, equivalently, Algorithm 1) with* $\epsilon = \sqrt{\frac{4(\log(n)+1)}{TM^2\Theta}}$ *allows us to bound the expected regret as*

$$\mathbb{E}[\mathcal{R}_T] \leq 5M\Theta + 4M\sqrt{T\Theta(\log(n)+1)}.$$

5. CONCLUSIONS & FUTURE WORK

In this paper, we considered a specific case of online decision making where the stage costs depend on the average of past decisions. By converting it into a online learning problem with stage-wise constraints, we were able to apply our novel FTARL algorithm to obtain tight regret bounds for this problem. The success of our approach in being able to handle constraints in online learning problems brings up a natural direction of future work. It remains to be seen how our algorithm can be adapted to handle constraints that crop up in other online learning scenarios.

We also restricted ourselves to learning decisions that lie in a simplex with signed cost functions. Broadening our scope to include general convex decision set and non-linear convex cost function is an important avenue for future work. In our work, we also weigh every decision in the finite history horizon equally. However, in many realistic scenarios, past decisions may have less influence on the current cost compared to the current decision. Exploring weighted averaging within our framework is another interesting line of study to explore.

Acknowledgments

This work was supported by the ARO MURI grant W911NF-20-0252 (76582 NSMUR). We would like to thank the anonymous reviewers for their thoughtful comments and for helping us in improving our manuscript.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In Sham M. Kakade and Ulrike von Luxburg, editors, *Conference on Learning Theory*, volume 19, pages 1–26. PMLR, 2011.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M. Kakade, and Karan Singh. Online control with adversarial disturbances. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, 2019.

Jason Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. In *Conference On Learning Theory*, pages 1569–1573. PMLR, 2018.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: Price of past mistakes. In *Advances in Neural Information Processing Systems*, volume 2015-January, 2015.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: From regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, volume 2, 2012.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Masoud Badiei, Na Li, and Adam Wierman. Online convex optimization with ramp constraints. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 6730–6736. IEEE, 2015.

Ziv Bar-Yossef, Kirsten Hildrum, and Felix Wu. Incentive-compatible online auctions for digital goods. In *SODA*, volume 2, pages 964–970, 2002.

Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 309–313, 1997.

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

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26, 2013.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1029–1038. PMLR, 2018.

Thomas M Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

Sascha Geulen, Berthold Vöcking, and Melanie Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In *COLT 2010 - The 23rd Conference on Learning Theory*, 2010.

András György and Gergely Neu. Near-optimal rates for limited-delay universal lossy source coding. *IEEE Transactions on Information Theory*, 60(5), 2014. ISSN 00189448. doi: 10.1109/TIT.2014.2307062.

James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Elad Hazan. *Introduction to online convex optimization*, chapter 5. The MIT Press, 2 edition, 2022.

Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in neural information processing systems*, 20, 2007.

Vijeth Hebbar and Cedric Langbort. Online decision making with history-average dependent costs (extended). *arXiv preprint arXiv:2312.06641*, 2023.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Branislav Kveton, Jia Yuan Yu, Georgios Theocharous, and Shie Mannor. Online learning with expert advice and finite-horizon constraints. In *AAAI*, pages 331–336, 2008.

Shie Mannor, John N Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(3), 2009.

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69:115–142, 2007.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Guanghui Wang, Yuanyu Wan, Tianbao Yang, and Lijun Zhang. Online convex optimization with continuous switching constraint. *Advances in Neural Information Processing Systems*, 34:28636–28647, 2021.