Regret Guarantees for Adversarial Contextual Bandits with Delayed Feedback

Orin Levy[*]
Blavatnik School of Computer Science
Tel Aviv University
orinlevy@mail.tau.ac.il

Liad Erez[*]
Blavatnik School of Computer Science
Tel Aviv University
liaderez@mail.tau.ac.il

Yishay Mansour
Blavatnik School of Computer Science
Google Research
mansour.yishay@gmail.com

Abstract

In this paper we present regret minimization algorithms for the contextual multi-armed bandit (CMAB) problem in the presence of delayed feedback, a scenario where loss observations arrive with delays chosen by an adversary. We study two fundamental frameworks in terms of the function classes used to derive regret bounds for CMAB. Firstly, for a finite policy class Π, we establish an optimal regret bound of $O(\sqrt{KT\log|\Pi|}+\sqrt{D\log|\Pi|})$, where K is the number of actions, T is the number of rounds, and D is the sum of delays. Secondly, assuming a finite contextual loss function class $\mathcal{F}$ and access to an online least-square regression oracle $\mathcal{O}$ over $\mathcal{F}$, we achieve a regret bound of $\widetilde{O}(\sqrt{KT(\mathcal{R}_T(\mathcal{O})+\log(\delta^{-1}))}+\eta D+d_{max})$ that holds with probability at least $1-\delta$, where d_{max} is the maximal delay, $\mathcal{R}_T(\mathcal{O})$ is an upper bound on the oracle's regret and η is a stability parameter associated with the oracle.

1 Introduction

Multi-Armed Bandit (MAB) is one of the most fundamental and well-studied online learning settings (see, e.g., [23, 31]). MAB describes a sequential decision-making problem where in each round the learner chooses an action out of a finite set $\mathcal{A}$ containing K actions and suffers a loss for that choice. The learner's goal is to minimize the cumulative loss incurred throughout an interaction of T rounds. In this model, action selection strategies are referred to as a *policies*, and the learner ultimately aims to minimize *regret*, that is, the learner's cumulative loss is compared to that of the best action selection rule, i.e., the optimal policy.

MAB can describe various real-life online scenarios such as advertising, gaming, and healthcare. Notwithstanding, in many modern applications, there are exterior factors that affect the loss incurred by any choice of action. One such application is online advertising, where the reaction of a user to a presented advertisement (i.e., clicking or ignoring) is heavily dependent on the user's needs (e.g., if they would like to buy a new car), hobbies, and personal preferences. All of the previous can be encoded in the user's browsing history and cookies. Thus, the user's cookies can refer to the external factors that affect the user's implied loss. These examples (and many others) motivate the model of *Contextual Multi-Armed Bandits (CMAB)*, where the external information is referred to as the

[*] Equal contribution.

17th European Workshop on Reinforcement Learning (EWRL 2024).

context that determines the loss of each action. The context is revealed to the learner at the start of each round of the game. The context space, denoted by $\mathcal{X}$, is generally thought of as huge or even infinite. In the *adversarial CMAB* model ([6, 11]), the context in each round is chosen by a (possibly adaptive) adversary.

CMAB has been vastly studied, under various assumptions and different frameworks, which we will review later. Returning to the online advertising example, in such an application, delayed feedback is practically unavoidable. Consider the scenario where a sequence of users enters the application one by one. The algorithm then needs to present them with advertisements, even though the feedback of previous users has not arrived yet. As the application takes time to process each user's feedback, the feedback will arrive one by one, in a First-In-First-Out (FIFO) fashion, with an inherent and unavoidable delay. Such real-life applications motivate the setting of *MAB with delayed feedback*, which has also gained considerable attention in recent years, either when the environment is adversarial [7, 8, 15, 32, 38] or stochastic [14, 20, 34].

In this paper, we consider the problem of *Adversarial CMAB with Delayed Feedback* under the two main frameworks studied in CMAB literature: (1) *Policy class learning* (see e.g., [5, 6, 10]), where the learner has access to a finite class $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$ of deterministic mappings from contexts to actions. The learner's goal is then to compete against the best policy in the class. (2) *Function approximation* (see, e.g., [4, 11, 13, 30]) where the learner has access to a loss function class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}\times\mathcal{A}}$ where each function defines a mapping from context and action to a loss value in $[0,1]$. The learner accesses the function class via a regression oracle, which is assumed to be efficient. In this setting the learner competes against the best contextual policy $\pi^\star : \mathcal{X} \to \mathcal{A}$ on the true loss function $f^\star \in \mathcal{F}$. In this setting, in addition to the standard least squares regret assumption required from the oracle, our approach will require a stability assumption that will be discussed later in the paper. Our goal is to derive regret minimization algorithms in both frameworks. Our main results are stated below.

Summary of our contributions. In this paper, we present delay-adapted algorithms in both CMAB frameworks and analyze the regret of the proposed methods. In more detail, our main results are summarized as follows:

- In the policy class learning framework, we present a delay-adapted version of the well-known EXP4 algorithm (Algorithm 1) with biased loss estimators that are specialized to the CMAB setting. For this approach we prove (see Theorem 1) that our algorithm has an expected regret bound of $O\left(\sqrt{KT\log|\Pi|} + \sqrt{D\log|\Pi|}\right)$, where K is the number of actions, T is the number of rounds, and D is the sum of delays, and has been shown in [8] to be optimal up to logarithmic factors.
- In the function approximation setting, we assume access to a finite contextual loss function class $\mathcal{F}$, which is accessed via an online least-square regression oracle $\mathcal{O}$ over $\mathcal{F}$. In this framework, we present a delay-adapted version of function approximation methods for CMAB, as specified in our algorithm DA-FA (Algorithm 2). This algorithm is a delay-adapted version of the algorithm OMG-CMDP! [26] that is specialized for CMAB (rather than contextual MDP). For this algorithm we prove (in Theorem 6) a regret bound of $\widetilde{O}(\sqrt{KT(\mathcal{R}_T(\mathcal{O}) + \log(\delta^{-1}))} + \eta D + d_{max})$ that holds with probability at least $1-\delta$, where d_{max} is the maximal delay, $\mathcal{R}_T(\mathcal{O})$ is an upper bound on the oracle's regret and η is a stability parameter associated with the oracle. To our knowledge, our work is the first to consider delayed feedback in adversarial CMAB in the fully general function approximation framework.

1.1 Additional related work

Contextual MAB. Contextual MAB has been vastly studied over the years, under various assumptions. Previous literature has two main lines of work. The first is policy class learning, starting from the well-known EXP4 algorithm for adversarial CMAB [6], to [5, 10] that study computationally efficient stochastic CMAB and obtain an optimal regret bound of $\widetilde{O}(\sqrt{TK\log(|\Pi|)})$.

The second line of work is the realizable function approximation setting, which has also been studied for stochastic CMAB, starting from Langford and Zhang [22] to [3, 30, 35] for which an optimal regret bound of $\widetilde{O}(\sqrt{TK\log(|\mathcal{F}|)})$ has been shown, where $\mathcal{F} \subseteq [0,1]^{\mathcal{X}\times\mathcal{A}}$ is a finite contextual reward/loss function class, accessed via an offline regression oracle. The adversarial variant of CMAB has also gained much attention recently, in the following significant line of works [11, 12, 13, 37],

where an online regression is being used to access the function class $\mathcal{F}$, with an optimal regret bound of $\widetilde{O}(\sqrt{KT\mathcal{R}_T(\mathcal{O})})$, where $\mathcal{R}_T(\mathcal{O})$ is the oracle's regret.

Regret guarantees for linear CMAB first studied by [2] and the SOTA algorithms are those of [1, 9]. Contextual MDPs (which are an extension of MAB, that has multiple states and dyanmics) have been studied under the function approximation framework [24, 25], with [26] being the most relevant to our setting as it studies adversarial CMDP, and inspired our algorithm and analysis for the function approximation setting.

Online Learning with Delayed Bandit Feedback. Delayed feedback has been an area of considerable interest in various online MAB problems in the past few years, with the first work on adversarial MAB with a constant delay d by [8]. Subsequent results for adversarial MAB with arbitrary delays have been established by [7, 32], with [32] being the first work to introduce a technique known as *skipping*, which allows for obtaining nontrivial regret bounds even if the delay sequence contains a relatively small number of excessive delays. [38] proposed the first algorithm for adversarial MAB with arbitrary delays that is made fully adaptive and does not require any prior knowledge on the delays. MAB with delayed feedback has also been studied from a best-of-both-worlds perspective [28, 29] in which the suggested algorithms obtain desirable regret bounds when losses are either stochastic or adversarial.

The study of delayed feedback in MAB has also been extended in several works to more general learning settings. Such settings include linear bandits [18, 34], generalized linear bandits [17], combinatorial semi-bandits [33] and bandit convex optimization [27]. Another prevalent generalization of MAB, in which delayed feedback has been studied, is RL, specifically tabular MDPs [19, 21, 33], with the work of [19] who first suggested the use of biased delay-adapted loss estimators which inspired our loss estimators used in Algorithm 1.

In CMAB, delayed feedback is far less explored, with the work of [36] who consider stochastic delays and contexts in generalized linear contextual bandits, which is a special case of the general function approximation setting studied in this paper.

2 Problem Setup

We consider the problem of *adversarial contextual MAB (CMAB) with delayed bandit feedback*.

Contextual MAB. Formally, *CMAB* is defined by a tuple $(\mathcal{X}, \mathcal{A}, \ell)$ where $\mathcal{X}$ is the context space, which is assumed to be large or even infinite, and $\mathcal{A}$ is a finite action space. $\ell : \mathcal{X} \times \mathcal{A} \to [0,1]$ forms an expected loss function, meaning, for $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\ell(x,a) = \mathbb{E}[L(x,a) \mid x,a]$ where $L(x,a) \in [0,1]$ is sampled from an unknown distribution, related to the context x and the action a. Note that in CMAB, in contrast to MAB, each context is associated with a potentially different best action $a^\star_x$, as the loss function is context-dependent. In the *adversarial CMAB* setup, the learner is faced with a sequential decision-making game which is played for T rounds where she is tasked with repeatedly choosing actions from a finite set $\mathcal{A}$ of actions (or arms). We also denote $\mathcal{A} = \{1, 2, \ldots, K\}$. The context in each round is chosen by a (possibly adaptive) adversary. Thus, the interaction protocol is as follows. In each round $t = 1, 2, \ldots, T$, nature reveals a *context* $x_t \in \mathcal{X}$, to the learner. The learner chooses an action a_t and suffers loss $L(x_t, a_t)$. A *policy* π defines a mapping from context to a distribution over actions, i.e., $\pi : \mathcal{X} \to \Delta(\mathcal{A})$. The learner's cumulative performance is ultimately compared to that of the best (deterministic) *policy* $\pi^\star : \mathcal{X} \to \mathcal{A}$.

Delayed feedback. The learner observes delayed bandit feedback, where the sequence of delays can be arbitrary. Formally, delays are determined by an arbitrary sequence of numbers $d_1, \ldots, d_T \in \{0, 1, \ldots, T\}$. In each round t, after choosing an action a_t, the learner observes the pairs $(s, L(x_s, a_s))$ for all rounds $s \leq t$ with $s + d_s = t$; crucially, only the loss values are delayed, whereas the contexts x_t are each observed at the start of round t. We consider a setting where the sequence of delays $(d_t)_{t=1}^T$ as well as the contexts $(x_t)_{t=1}^T$ are generated by an adversary. In this paper we denote by D the sum of delays, that is $D = \sum_{t=1}^{d} d_t$, the the maximal delay is denoted by $d_{max} = \max_{t \in [T]} d_t$.

Learning objective. We consider the objective of minimizing *regret*, which is the difference between the cumulative loss of the learner and that of the best-fixed policy $\pi^\star$, i.e.,

$$\mathcal{R}_T := \sum_{t=1}^{T} \ell(x_t, a_t) - \ell(x_t, \pi^\star(x_t)).$$

We consider two different learning settings, that affect the benchmark we compete against, i.e., affects the definition of $\pi^\star$. The first setting we consider is *Policy Class Learning*. In this setting, we assume access to a *finite policy class* $\Pi \subseteq \mathcal{A}^\mathcal{X}$. Then, the benchmark $\pi^\star$ is the best policy among the class, i.e., $\pi^\star \in \arg\min_{\pi\in\Pi} \sum_{t=1}^{T} \ell(x_t, \pi(x_t))$. The second setting we consider is *Online function approximation*. In this setting, we assume access to a *realizable contextual loss* class $\mathcal{F} \subseteq \mathcal{X} \times \mathcal{A} \to [0,1]$, where realizability means that there exists a function $f^\star \in \mathcal{F}$ such that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ it holds that $f^\star(x,a) = \ell(x,a)$. Then, the learner's goal is to compete against $\pi^\star(x) = \arg\min_{a\in\mathcal{A}} f^\star(x,a)$, for all $x \in \mathcal{X}$.

3 Policy Class Learning

In this section, we study a formulation of the CMAB problem which considers a finite but structureless policy class $\Pi \subseteq \mathcal{A}^\mathcal{X}$, indexed by $\Pi = \{\pi_1, \ldots, \pi_N\}$ where we denote $|\Pi| = N$. We remark that in this formulation, the loss vectors $(L(x_t, \cdot))_{t=1}^T$ may also be generated by an adversary.

3.1 Algorithm: EXP4 with Delay-Adapted Loss Estimators

In the following, we present a variant of the well-studied EXP4 algorithm [6], formally described in Algorithm 1, which incorporates delay-robust loss estimators specialized to the CMAB setting. On a high-level, the algorithm performs multiplicative weight updates over the N-dimensional simplex Δ_N, while using all of the feedback that arrives in each round t to construct loss estimators, denoted by $\hat{c}_t \in \mathbb{R}_+^N$. Our loss estimators are inspired by those suggested by [19], and are reminiscent of the standard importance-weighted loss estimators, with an additional term in the denominator which induces an under-estimation bias. More specifically, the standard (unbiased) importance-weighted loss estimators are defined by

$$\tilde{c}_{t,i} = \frac{L(x_t, a_t)\mathbb{I}[\pi_i(x_t) = a_t]}{Q_{t,a_t}} \quad \forall i \in [N], \tag{1}$$

where $Q_{t,a} = \sum_{i=1}^{N} p_{t,i}\mathbb{I}[\pi_i(x_t) = a]$ is the distribution over actions induced by $p_t \in \Delta_N$ and the context x_t. For our approach and analysis, it is crucial to introduce biased versions of the above estimators, defined as

$$\hat{c}_{t,i} = \frac{L(x_t, a_t)\mathbb{I}[\pi_i(x_t) = a_t]}{\max\left\{Q_{t,a_t}, \tilde{Q}_{t,a_t}^{t+d_t}\right\}} \quad \forall i \in [N], \tag{2}$$

where $\tilde{Q}_{t,a}^{t+d_t} = \sum_{i=1}^{N} p_{t+d_t,i}\mathbb{I}[\pi_i(x_t) = a]$ is the distribution over actions induced by the distribution p_{t+d_t} and the context x_t. Interestingly, these estimators exhibit a coupling between the context x_t, which arrives at a given round t, and the sampling distribution p_{t+d_t} from a *future* round, and can be thought of as a mechanism that incentivizes actions whose sampling probability has increased between rounds t and $t + d_t$, with respect to the context x_t. The main result for Algorithm 1 is given in the following theorem.

Theorem 1. *Algorithm 1 attains an expected regret bound of*

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log N}{\eta} + \eta KT + 2\eta D,$$

where the expectation is over the algorithm's stochasticity. For $\eta = \sqrt{\frac{\log N}{KT+D}}$ *we obtain*

$$\mathbb{E}[\mathcal{R}_T] \leq O\left(\sqrt{KT\log N} + \sqrt{D\log N}\right).$$

We remark that Algorithm 1 requires an upper bound on the sum of delays D, however, it can be made adaptive by utilizing a "doubling" mechanism, see Section 5 for more detail. In what follows, we highlight the main steps and technical challenges of the analysis towards proving Theorem 1. The full proof can be found in Appendix A.1.

Algorithm 1 EXP4 with Delay-Adapted Loss Estimators (EXP4-DALE)

inputs:

- Finite policy class $\Pi \subseteq \mathcal{X} \to \mathcal{A}$ with $|\Pi| = N$,
- Upper bound on the sum of delays, D.
- Step size $\eta > 0$.

Initialize $p_1 \in \Delta_N$ as the uniform distribution over Π.

for round $t = 1, \ldots, T$ **do**

Receive context $x_t \in \mathcal{X}$.

Sample $\pi \sim p_t$ and play $a_t = \pi(x_t)$.

Observe feedback $(s, L(x_s, a_s))$ for all $s \leq t$ with $s + d_s = t$ and construct loss estimators

$$\hat{c}_{s,i} = \frac{L(x_s, a_s)\mathbb{I}[\pi_i(x_s) = a_s]}{\max\left\{Q_{s,a_s}, \tilde{Q}^t_{s,a_s}\right\}} \quad \forall i \in [N],$$

where we define $Q_{s,a} = \sum_{i=1}^N p_{s,i}\mathbb{I}[\pi_i(x_s) = a]$ and $\tilde{Q}^t_{s,a} = \sum_{i=1}^N p_{t,i}\mathbb{I}[\pi_i(x_s) = a]$.

Update

$$p_{t+1,i} \propto p_{t,i} \exp\left(-\eta \sum_{s:s+d_s=t} \hat{c}_{s,i}\right). \tag{3}$$

3.2 Analysis Overview and Technical Challenges

The main technical novelty of our approach is expressed in the use of delay-adapted loss estimators inspired by the work of [19]. While the natural delay-adapted version of EXP4 with the standard importance-weighted loss estimators may lead to an optimal regret bound, the standard analysis would result in a regret bound containing terms of the form $Q_{t+d_t,a}/Q_{t,a}$ which, in order to be able to bound appropriately, would require further involved analysis of the multiplicative stability of the algorithm. Our approach, however, alleviates the need to analyze the algorithm's multiplicative stability while instead incurring an additive bias term of the form $\sum_t \sum_a |Q_{t,a} - \tilde{Q}^{t+d_t}_{t,a}|$, which we bound in Lemma 13 in Appendix A.1 by an overall drift term which already appears in the regret bound due to the presence of delays.

With that in mind, we begin by decomposing the regret of Algorithm 1 as follows:

$$\begin{aligned}\mathcal{R}_T &= \sum_{t=1}^T c_t \cdot (p_t - p^\star) \\ &= \underbrace{\sum_{t=1}^T p_t \cdot (c_t - \hat{c}_t)}_{Bias_1} + \underbrace{\sum_{t=1}^T p^\star \cdot (\hat{c}_t - c_t)}_{Bias_2} + \underbrace{\sum_{t=1}^T (p_t - p_{t+d_t}) \cdot \hat{c}_t}_{Drift} + \underbrace{\sum_{t=1}^T (p_{t+d_t} - p^\star) \cdot \hat{c}_t}_{OMD},\end{aligned} \tag{4}$$

where $c_{t,i} = L(x_t, \pi_i(x_t))$ denotes the true loss of π_i on the context x_t. First, we note that the $Bias_2$ term is negative in expectation, since the delay-adapted loss estimators $\hat{c}_t$ are upper bounded by the standard unbiased estimators $\tilde{c}_t$. The OMD term can be bounded by standard OMD analysis as

$$\sum_{t=1}^T (p_{t+d_t} - p^\star) \cdot \hat{c}_t \leq \frac{\log N}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^N p_{t+d_t,i}\hat{c}^2_{t,i}\right].$$

Next, we bound both the $Bias_1$ and $Drift$ terms, the first of which arises from the bias in our delay-adapted loss estimators, and the second is affected by the discrepancy between the iterates p_{t+d_t} and p_t. Both of these terms can be bounded in expectation by a quantity which is governed by the stability of Algorithm 1 and the sequence of delays, namely $\sum_{t=1}^T \mathbb{E}\|p_{t+d_t} - p_t\|_1$. Thus, we obtain the following expected regret bound:

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log N}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t+d_t,i}\hat{c}_{t,i}^2\right] + 2\mathbb{E}\left[\sum_{t=1}^{T}\|p_{t+d_t} - p_t\|_1\right]. \tag{5}$$

Given this result, Theorem 1 follows by bounding the first two terms (arising from the OMD term) by ηKT, and the third term by $\eta(D+T)$. Intuitively, the drift term can be controlled by using the fact that Algorithm 1 is an instantiation of mirror descent with a negative-entropy regularization, which is strongly convex with respect to the L_1 norm. This ensures that the algorithm's updates will be stable in the sense that

$$\mathbb{E}\|p_{t+1} - p_t\|_1 \leq \eta m_t,$$

where m_t is the number of observations that arrive on round t. Thus, we can bound the third term in Eq. (5) by $\eta\sum_{t=1}^{T}(\#$ of observations that arrive between rounds t and $t+d_t)$, which is shown to be bounded by $\eta(D+T)$. Then, using the specific form of our loss estimators given in Eq. (2), we can bound the second term in Eq. (5) by ηKT by making use of the specific form of the loss estimators given in Eq. (2), which gives us the desired bound in Theorem 1.

4 Online Function Approximation

In this section, we provide regret guarantees for CMAB with delayed feedback under the framework of online function approximation [11, 13]. In this setting, the learner has access to a class of loss functions $\mathcal{F} \subseteq \mathcal{X} \times \mathcal{A} \to [0,1]$, where each function $f \in \mathcal{F}$ maps context $x \in \mathcal{X}$ and an action $a \in \mathcal{A}$ to loss $\ell \in [0,1]$. We use $\mathcal{F}$ to approximate the context-dependent expected loss of any action $a \in \mathcal{A}$ for any context $x \in \mathcal{X}$. We access $\mathcal{F}$ using an online least-squares regression (OLSR) oracle that will operate under the following standard realizability assumption.

Assumption 2. *There exists a function $f^\star \in \mathcal{F}$ such that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, $f^\star(x,a) = \ell(x,a)$.*

We assume access to a classical, non-delayed, online regression oracle with respect to the square loss function $h_{sq}(\hat{y}, y) = (\hat{y} - y)^2$. The oracle, which we denote by $\mathcal{O}_{\text{sq}}^{\mathcal{F}}$, is given as input at each round t the past observations $(x_s, a_s, L_s(x_s,a_s))_{s=1}^{t-1}$ and outputs a function $\hat{f}_t \in \mathcal{F}$. A general formulation of the online oracle model is discussed in [11]. We make use of the following standard online least-squares regret assumption of the oracle which is also considered in [11].

Assumption 3 (Least-Squares Oracle Regret). *The oracle $\mathcal{O}_{sq}^{\mathcal{F}}$ guarantees that for every sequence $\{(x_t, a_t, L_t)\}_{t=1}^{T}$, regret is bounded as*

$$\sum_{t=1}^{T}(\hat{f}_t(x_t,a_t) - L_t)^2 - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}(f(x_t,a_t) - L_t)^2 \leq \mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F}}).$$

The oracle also satisfies the following concentration guarantee, as implied by Lemma 2 in [11].

Lemma 4 (Concentration of non-delayed OLSR oracle). *Under Assumption 2 and Assumption 3, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$.*

$$\sum_{t=1}^{T}\mathbb{E}_{a_t\sim p_t}\left[\left(\hat{f}_t(x_t,a_t) - \ell(x_t,a_t)\right)^2\right] \leq 2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F}}) + 16\log(2/\delta).$$

Assumption 2 and Assumption 3 (or variants of it for other loss functions) are necessary to derive regret bounds for adversarial CMAB and appear in previous literature, e.g., [11, 12]. However, a general implementation of online least-square regression oracle for a general function class might be unstable. In more detail, let $\{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_T\}$ denote the sequence of functions outputted by the (non-delayed) oracle on the observation sequence $\{(x_t, a_t, L_t)\}_{t=1}^{T}$. Then, $\|\hat{f}_i(x,\cdot) - \hat{f}_{i+1}(x,\cdot)\|_\infty$ might be of high magnitude, for any $x \in \mathcal{X}$ and $i \in [T]$. In the standard adversarial CMAB setting, this is a non-issue as that difference is absorbed by the oracle's regret on the sequence of examples. However, when considering delayed feedback, stability becomes crucial. The reason is that the

oracle's updates now depend on the arrival time of new loss observations. Hence, it is possible that the oracle is not updated for a considerably long time, but then, at some given time step, receives many examples at once, and will be fed with all of these examples. Thus, without being able to control the stability of the oracle, and due to the setting being adversarial, we might experience uncontrollable changes in the loss approximations and incur linear regret for unfavorable delay sequences. Hence, we impose the following natural stability assumption, stated below.

Assumption 5 (η**-stability).** *Let $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_T$ denote the function sequence outputted by the non-delayed oracle $\mathcal{O}_{sq}^{\mathcal{F}}$ on the observation sequence $\{(x_t, a_t, L_t)\}_{t=1}^{T}$. We assume that for all $t \in [T]$ and $x \in \mathcal{X}$, it holds that $\|\hat{f}_t(x,\cdot) - \hat{f}_{t-1}(x,\cdot)\|_\infty \leq \eta$ for $\eta > 0$. We denote the η-stable oracle for the function class $\mathcal{F}$ by $\mathcal{O}_{sq}^{\mathcal{F},\eta}$.*

We note that our online regression oracle is essentially an online optimization algorithm. This makes our stability assumption even more understandable because stability is an important property of many online optimization algorithms, and is in some cases essential in order to derive regret guarantees. Such algorithmic frameworks include, for instance, FTRL and OMD (see, e.g., [16]). See also Section 5 for further discussion on the topic.

4.1 Algorithm: Delay-Adapted Function Approximation for CMAB

In the following, we present algorithm DA-FA (Algorithm 2) for regret minimization in CMAB with delayed feedback under the function approximation framework described above. The algorithm is a delay-adapted version of algorithm OMG-CMDP! [26] applied to adversarial CMAB and its analysis for the delay-independent terms is similar. Algorithm 2 essentially uses the most up-to-date approximation of the loss until delayed observations arrive. When they arrive, the algorithm feeds them to the oracle one by one, ignores the midway approximations, and uses only the newest loss approximation. More specific details are given below.

In each round $t = 1, 2, \ldots, T$ the algorithm operates as follows. Let $\alpha(t) < t$ denote the number of observations that arrived at round t. Denote these observations by $\{(s_i^t, L(x_{s_i^t}, a_{s_i^t}))\}_{i=1}^{\alpha(t)}$, where $s_1^t \leq \ldots \leq s_{\alpha(t)}^t$ denote the time steps of the non-delayed related context and action associated with these delayed loss observation. It then holds that $s_i^t + d_{s_i^t} = t$ for all $i \in [\alpha(t)]$. Note that we assume that the delayed observations arrive in FIFO order, meaning, the delayed observation from round n always arrives before (or in parallel to) that of round $n+1$ for all $n \in [T]$. Then, for $i = 1, \ldots, \alpha(t)$, we feed the oracle with the example $(x_{s_i^t}, a_{s_i^t}, L(x_{s_i^t}, a_{s_i^t}))$ and observe the predicted function $\hat{f}_{t-d_{s_i^t}}$. Let $\tau^t = t - d_{s_{\alpha(t)}^t} = s_{\alpha(t)}^t$ denote the index of the last observed delayed loss. After processing all the data that arrived, the current context x_t is revealed and the algorithm uses the last predicted function $\hat{f}_{\tau^t}$ to solve the regularized convex optimization problem specified in Eq. (6), and plays an action sampled from the resulted distribution.

The following theorem states the performance of our algorithm.

Theorem 6. *For any $\delta \in (0,1)$ let $\gamma = \sqrt{\frac{T|\mathcal{A}|}{2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta}) + 16\log(4/\delta)}}$. Then, with probability at least $1-\delta$, the following regret bound holds,*

$$\mathcal{R}_T \leq \widetilde{O}\left(\sqrt{T|\mathcal{A}|\Big(\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta}) + \log(\delta^{-1})\Big)} + \eta D + d_{max}\right).$$

Computational efficiency. Note that the optimization problem given in Eq. (6) is convex and can be solved efficiently to an arbitrary precision, and thus Algorithm 2 is clearly efficient, assuming an efficient oracle implementation.

4.2 Analysis

In this subsection, we analyze algorithm DA-FA (Algorithm 2), proving Theorem 6.

Our main technical challenge is reflected in the regret analysis. As in all previous literature regarding delayed feedback, the main challenge is to derive a bound where the sum of delays D is separated

Algorithm 2 Delay-Adapted Function Approximation for CMAB (DA-FA)

inputs:
- Function class $\mathcal{F}$ for loss approximation
- Learning rate parameter γ.
- η-stable OLSR oracle $\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta}$.

for round $t = 1, \ldots, T$ **do**
  observe $\alpha(t) < t$ losses $\{(s_i^t, L(x_{s_i^t}, a_{s_i^t}))\}_{i=1}^{\alpha(t)}$ where $\forall i \in [\alpha(t)],\ s_i^t + d_{s_i^t} = t$ and $s_1^t \leq \ldots \leq s_{\alpha(t)}^t$.
  for $i = 1, 2, \ldots, \alpha(t)$ **do**
    update $\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta}$ with the example $((x_{s_i^t}, a_{s_i^t}), L(x_{s_i^t}, a_{s_i^t}))$.
    observe the oracle's output $\hat{f}_{t-d_{s_i^t}} \leftarrow \mathcal{O}_{\text{sq}}^{\mathcal{F},\eta}$.
  let $\tau^t := t - d_{s_{\alpha(t)}^t} = s_{\alpha(t)}^t$ denote index of the last observed loss.
  use $\hat{f}_{\tau^t}$ as the current loss approximation.
  observe context $x_t \in \mathcal{X}$.
  solve

$$p_t = \arg\min_{p \in \Delta_A} \sum_{a \in \mathcal{A}} p(a) \cdot \hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma} \sum_{a \in \mathcal{A}} \log(p(a)). \tag{6}$$

  play the action a_t sampled from $p_t(\cdot)$

from the number of actions. Usually, this separation is obtained by a delicate choice of loss estimators. In our case, the loss estimator choice is done by the oracle, hence not transparent to the algorithm. Our way to create the desired separation is by the decomposition of the regret described in Eq. (7). Let $\{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_T\}$ denote the functions predicted by the OLSR oracle on the non-delayed observation sequence $\{(x_1, a_1, L(x_1, a_1)), (x_2, a_2, L(x_2, a_2)), \ldots, (x_T, a_T, L(x_T, a_T))\}$. That is, for all $i \in [T-1]$, $\hat{f}_{i+1} = \mathcal{O}_{\text{sq}}^{\mathcal{F},\eta}(\cdot; (x_1, a_1, L(x_1, a_1)), \ldots, (x_i, a_i, L(x_i, a_i)))$. In the following analysis, for convenience, we denote the optimal (randomized) policy by $p_\star(\cdot|x)$ for all $x \in \mathcal{X}$. Then, the regret is given by $\mathcal{R}_T = \sum_{t=1}^T (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \ell(x_t, \cdot)$ and decomposed as follows.

$$\begin{aligned}
\mathcal{R}_T = &\underbrace{\sum_{t=1}^{d_{\max}} (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \ell(x_t, \cdot)}_{(a)} + \underbrace{\sum_{t=d_{\max}+1}^{T} (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot)}_{(b)} \\
&+ \underbrace{\sum_{t=d_{\max}+1}^{T} p_t(\cdot) \cdot \left(\ell(x_t, \cdot) - \hat{f}_t(x_t, \cdot)\right)}_{(c)} + \underbrace{\sum_{t=d_{\max}+1}^{T} p_\star(\cdot \mid x_t) \cdot \left(\hat{f}_t(x_t, \cdot) - \ell(x_t, \cdot)\right)}_{(d)} \\
&+ \underbrace{\sum_{t=d_{\max}+1}^{T} (p_t - p_\star(\cdot \mid x_t)) \cdot \left(\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot)\right)}_{(e)}.
\end{aligned} \tag{7}$$

In the above decomposition, term (a) is bounded trivially by $d_{\max}$. Term (b) is the regret with respect to the approximated delayed loss. Term (c) is the approximation error with respect to the policy induced by $p_t(\cdot)$, when considering the non-delayed approximated loss. This term will be bounded by the oracle regret. Term (d) is the approximation error with respect to the optimal $p_\star(\cdot|\cdot)$ when considering the non-delayed approximated loss. Lastly, term (e) is the regret caused by the delay drift in approximation. This term will be shown to be bounded by ηD, independently of the number of actions.

We bound each term individually in the following lemmas, and then combine the results to conclude Theorem 6. We start by term (a) that is bounded trivially by $d_{\max}$, since the losses are in $[0, 1]$.

We then proceed to bound term (b), whose bound follows from first-order optimality conditions for convex optimization.

Lemma 7 (Term (b) bound). *It holds true that*

$$\sum_{t=d_{\max}+1}^{T}(p_t(\cdot)-p_\star(\cdot|x_t))\cdot\hat{f}_{\tau^t}(x_t,\cdot)\leq\frac{T|\mathcal{A}|}{\gamma}-\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star(a|x_t)}{\gamma p_t(a)}.$$

Term (c) is bounded using AM-GM inequality, and applying the non-delayed oracle concentration bound stated in Lemma 4. We obtain the following.

Lemma 8 (Term (c) bound). *With probability at least* $1-\delta/2$*,*

$$\sum_{t=d_{\max}+1}^{T}p_t(\cdot)\cdot\Big(\ell(x_t,\cdot)-\hat{f}_t(x_t,\cdot)\Big)\leq\frac{T|\mathcal{A}|}{\gamma}+\gamma\big(2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta})+16\log(4/\delta)\big).$$

Term (d) is bounded using AM-GM inequality to change the measure from $p_\star(\cdot|x_t)$ to $p_t(\cdot)$, to then apply the non-delayed oracle concentration bound. We obtained the following.

Lemma 9 (Term (d) bound). *With probability at least* $1-\delta/2$ *it holds that*

$$\sum_{t=d_{\max}+1}^{T}p_\star(\cdot|x_t)\cdot\Big(\hat{f}_t(x_t,\cdot)-\ell(x_t,\cdot)\Big)\leq\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star(a|x_t)}{\gamma p_t(a)}+\gamma\big(2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta})+16\log(4/\delta)\big).$$

The proofs of Lemmas Lemmas 7 to 9 are inspired by those of [26], and included for completeness in Appendix A.2.

Lastly, we bound term (e). We apply Hölder's inequality and η-stability (Assumption 5) to obtain the following result. The full proof of the lemma can also be found in Appendix A.2.

Lemma 10 (Term (e) bound). *Under Assumption 5 the following holds true.*

$$\sum_{t=d_{\max}+1}^{T}(p_t(\cdot)-p_\star(\cdot\mid x_t))\cdot\Big(\hat{f}_t(x_t,\cdot)-\hat{f}_{\tau^t}(x_t,\cdot)\Big)\leq 2\eta D.$$

We now have what we need to prove Theorem 6.

Proof of Theorem 6. Putting the results of Lemmas 7 to 10 all together, with probability at least $1-\delta$ the regret of Algorithm 2 is bounded as follows.

$$\mathcal{R}_T\leq d_{\max}+2\frac{T|\mathcal{A}|}{\gamma}+2\gamma\big(2\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})+16\log(4/\delta)\big)+2\eta D.$$

Choosing $\gamma=\sqrt{\frac{T|\mathcal{A}|}{2\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})+16\log(4/\delta)}}$ yields the desired bound. ■

5 Conclusions and Discussion

In this paper we present regret minimization algorithms for adversarial CMAB with delayed feedback, where both the contexts and delays are chosen by a (possibly adaptive) adversary. We consider the problem under the two mainstream frameworks for CMAB learning: (1) policy class learning and (2) online function approximation.

For (1) we present the algorithm EXP4-DALE (Algorithm 1) and prove that it obtains a regret bound of $O\Big(\sqrt{KT\log|\Pi|}+\sqrt{D\log|\Pi|}\Big)$ which is known to be optimal, up to logarithmic factors, in terms of the sum of delays D. We remark that while our approach is not designed to handle delay sequences with overly excessive delays, we strongly believe that it is possible to employ skipping techniques similar to those of [38] in order to obtain more refined bounds in terms of the delay sequence. Additionally, while Algorithm 1 requires knowledge of T and D in order to tune the

learning rate, using a doubling approach similar to the one suggested by [21], the same bound can be obtained without knowledge of T or D.

For (2) we present the algorithm DA-FA (Algorithm 2) and analyze its regret under a natural stability assumption related to the online regression oracle in use, which affects our bound. By the applicability of this model, studying minimal and natural assumptions regarding the oracle in use that enables achieving sublinear regret guarantees is a truly interesting question we leave for future research. We will remark, however, that for online oracles such as the least-squares regression oracles employed throughout the adversarial CMAB literature, assuming that they satisfy certain stability properties seems fairly natural. The reason is that a plethora of online learning algorithms (with an online oracle being one such algorithm) are based on either FTRL or OMD updates, which inherently make use of a step size η which governs the algorithm's stability. While we make no assumption on the specific algorithmic structure of the given online oracle for our setting, a practical implementation of such an oracle would most likely involve FTRL / OMD style updates which would also induce stability.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396 and grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 993/17 and 2549/19), Tel Aviv University Center for AI and Data Science (TAD), the Yandex Initiative for Machine Learning at Tel Aviv University, the Len Blavatnik and the Blavatnik Family Foundation.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

[3] A. Agarwal, M. Dudík, S. Kale, J. Langford, and R. Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

[4] A. Agarwal, M. Dudik, S. Kale, J. Langford, and R. Schapire. Contextual bandit learning with predictable rewards. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 19–26, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[5] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1638–1646, 2014.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[7] I. Bistritz, Z. Zhou, X. Chen, N. Bambos, and J. Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. *Advances in neural information processing systems*, 32, 2019.

[8] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

[9] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[10] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

[11] D. Foster and A. Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[12] D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.

[13] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[14] M. A. Gael, C. Vernade, A. Carpentier, and M. Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.

[15] A. Gyorgy and P. Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pages 3988–3997. PMLR, 2021.

[16] E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[17] B. Howson, C. Pike-Burke, and S. Filippi. Delayed feedback in generalised linear bandits revisited. In *International Conference on Artificial Intelligence and Statistics*, pages 6095–6119. PMLR, 2023.

[18] S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems*, 33:4872–4883, 2020.

[19] T. Jin, T. Lancewicki, H. Luo, Y. Mansour, and A. Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35: 33469–33481, 2022.

[20] T. Lancewicki, S. Segal, T. Koren, and Y. Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978. PMLR, 2021.

[21] T. Lancewicki, A. Rosenberg, and Y. Mansour. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7281–7289, 2022.

[22] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

[23] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[24] O. Levy and Y. Mansour. Optimism in face of a context: Regret guarantees for stochastic contextual mdp. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8510–8517, 2023.

[25] O. Levy, A. Cassel, A. Cohen, and Y. Mansour. Eluder-based regret for stochastic contextual mdps. *arXiv preprint arXiv:2211.14932*, 2022.

[26] O. Levy, A. Cohen, A. B. Cassel, and Y. Mansour. Efficient rate optimal regret for adversarial contextual mdps using online function approximation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19287–19314. PMLR, 2023.

[27] B. Li, T. Chen, and G. B. Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 993–1002. PMLR, 2019.

[28] S. Masoudian, J. Zimmert, and Y. Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. *Advances in Neural Information Processing Systems*, 35:11752–11762, 2022.

[29] S. Masoudian, J. Zimmert, and Y. Seldin. An improved best-of-both-worlds algorithm for bandits with delayed feedback. *arXiv preprint arXiv:2308.10675*, 2023.

[30] D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

[31] A. Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

[32] T. S. Thune, N. Cesa-Bianchi, and Y. Seldin. Nonstochastic multiarmed bandits with unrestricted delays. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] D. van der Hoeven, L. Zierahn, T. Lancewicki, A. Rosenberg, and N. Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and mdps. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1285–1321. PMLR, 2023.

[34] C. Vernade, O. Cappé, and V. Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

[35] Y. Xu and A. Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

[36] Z. Zhou, R. Xu, and J. Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.

[37] Y. Zhu, D. J. Foster, J. Langford, and P. Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pages 27428–27453. PMLR, 2022.

[38] J. Zimmert and Y. Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3285–3294. PMLR, 2020.

A Proofs

A.1 Proofs for Section 3

Throughout this section, we use the notation $\mathbb{E}_t[\cdot]$ to denote an expectation conditioned on the entire history up to round t.

Theorem 11. *Algorithm 1 attains the following expected regret bound:*

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{\log N}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t+d_t,i}\hat{c}_{t,i}^2\right] + 2\mathbb{E}\left[\sum_{t=1}^{T}\|p_{t+d_t} - p_t\|_1\right].$$

Proof. The regret may be decomposed as follows:

$$\begin{aligned}\mathcal{R}_T &= \sum_{t=1}^{T} c_t \cdot (p_t - p^\star) \\ &= \underbrace{\sum_{t=1}^{T} p_t \cdot (c_t - \hat{c}_t)}_{Bias_1} + \underbrace{\sum_{t=1}^{T} p^\star \cdot (\hat{c}_t - c_t)}_{Bias_2} + \underbrace{\sum_{t=1}^{T}(p_t - p_{t+d_t}) \cdot \hat{c}_t}_{Drift} + \underbrace{\sum_{t=1}^{T}(p_{t+d_t} - p^\star) \cdot \hat{c}_t}_{OMD}, \end{aligned} \tag{8}$$

where $c_{t,i} = L(x_t, \pi_i(x_t))$ for $i \in [N]$. The OMD term can be bounded by referring to Lemma 9 of [32] which asserts that

$$\sum_{t=1}^{T}(p_{t+d_t} - p^\star) \cdot \hat{c}_t \leq \frac{\log N}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{|\Pi|} p_{t+d_t,i}\hat{c}_{t,i}^2. \tag{9}$$

while noting that this lemma does not require a specific form of loss estimators, only that they are nonnegative, as is the case for our delay-adapted estimators defined in Eq. (2). We also note that the $Bias_2$ term is non-positive in expectation, since the delay-adapted estimators satisfy $\mathbb{E}_t[\hat{c}_{t,i}] \leq c_{t,i}$ for $i \in [N]$. Thus, to conclude the proof we are left with bounding the $Drift$ and $Bias_1$ terms, whose bounds are given in Lemma 12 and Lemma 13 that follow. ■

Proof of Theorem 1. First, we show that

$$\mathbb{E}\left[\sum_t\sum_i p_{t+d_t,i}\hat{c}_{t,i}^2\right] \leq KT.$$

Indeed, using the definition of the delay-adapted loss estimators $\hat{c}_t$, it holds that

$$\begin{aligned}\mathbb{E}\left[\sum_t\sum_i p_{t+d_t,i}\hat{c}_{t,i}^2\right] &= \mathbb{E}\left[\sum_t\sum_i p_{t+d_t,i}\left(\frac{L(x_t,a_t)\mathbb{I}[\pi_i(x_t)=a_t]}{\max\left\{Q_{t,a_t}, \tilde{Q}_{t,a_t}^{t+d_t}\right\}}\right)^2\right] \\ &\leq \mathbb{E}\left[\sum_t \frac{1}{\tilde{Q}_{t,a_t}^{t+d_t}}\sum_i \frac{p_{t+d_t,i}\mathbb{I}[\pi_i(x_t)=a_t]}{Q_{t,a_t}}\right] \\ &= \mathbb{E}\left[\sum_t \frac{1}{Q_{t,a_t}}\right] = \mathbb{E}\left[\sum_t\sum_a \frac{Q_{t,a}}{Q_{t,a}}\right] = KT.\end{aligned}$$

Thus, using Theorem 11 together with Lemma 14 gives the bound claimed in Theorem 1. ■

Lemma 12 (Bounding the Drift term). *The $Drift$ term given in Eq. (8) is bounded in expectation as follows:*

$$\mathbb{E}\left[\sum_{t=1}^{T}(p_t - p_{t+d_t}) \cdot \hat{c}_t\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\|p_t - p_{t+d_t}\|_1\right].$$

Proof. First, we note that the delay-adapted loss estimators $\hat{c}_t$ are upper-bounded by the standard, conditionally unbiased importance-weighted estimators $\tilde{c}_t$ defined in Eq. (1). Therefore, we can bound the $Drift$ term as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}(p_t - p_{t+d_t})\cdot\hat{c}_t\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}|p_{t,i} - p_{t+d_t,i}|\hat{c}_{t,i}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}|p_{t,i} - p_{t+d_t,i}|\tilde{c}_{t,i}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}|p_{t,i} - p_{t+d_t,i}|\cdot c_{t,i}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\|p_t - p_{t+d_t}\|_1\right],
\end{aligned}$$

where the last step follows from Hölder's inequality and the fact that $\|c_t\|_\infty \leq 1$. ■

Lemma 13. *The $Bias_1$ term given in Eq. (8) is bounded in expectation as follows:*

$$\mathbb{E}\left[\sum_t p_t\cdot(c_t - \hat{c}_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\|p_t - p_{t+d_t}\|_1\right].$$

Proof. We note losses and loss estimators can be indexed by actions rather than policies and use the the notation $c_{t,a} = L(x_t, a)$ and $\hat{c}_{t,a} = \frac{c_{t,a}\mathbb{I}[a_t=a]}{M_{t,a}}$ where $M_{t,a} = \max\left\{Q_{t,a}, \tilde{Q}_{t,a}^{t+d_t}\right\}$. Therefore, using the fact that $\mathbb{E}_t[\hat{c}_{t,a}] = c_{t,a}\frac{Q_{t,a}}{M_{t,a}}$, the $Bias_1$ term can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} p_t\cdot(c_t - \hat{c}_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} Q_{t,a}(L(x_t,a) - \hat{c}_{t,a})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} Q_{t,a}L(x_t,a)\left(1 - \frac{Q_{t,a}}{M_{t,a}}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\frac{Q_{t,a}}{M_{t,a}}(M_{t,a} - Q_{t,a})\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\max\left\{Q_{t,a}, \tilde{Q}_{t,a}^{t+d_t}\right\} - Q_{t,a}\right)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\left|\tilde{Q}_{t,a}^{t+d_t} - Q_{t,a}\right|\right].
\end{aligned}$$

Now, by the definition of $Q_{t,a}, \tilde{Q}_{t,a}^{t+d_t}$ and the triangle inequality, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\left|\tilde{Q}_{t,a}^{t+d_t} - Q_{t,a}\right|\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\sum_{i:\pi_i(x_t)=a}|p_{t+d_t,i} - p_{t,i}|\right] = \mathbb{E}\left[\sum_{t=1}^{T}\|p_t - p_{t+d_t}\|_1\right],$$

concluding the proof. ■

Lemma 14 (Distribution drift). *The following holds for the iterates $\{p_t\}_{t=1}^{T}$ of Algorithm 1:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\|p_{t+d_t} - p_t\|_1\right] \leq \eta(D+T).$$

Proof of Lemma 14. Define

$$F_t(p) = p\cdot\sum_{s:s+d_s<t}\hat{c}_s + \frac{1}{\eta}R(p),$$

where $R(p) = \sum_{i=1}^{N} p_i \log p_i$, so that $p_t = \arg\min_{p \in \Delta_\Pi} F_t(p)$. Note that $R(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_1$, and therefore $F_t(\cdot)$ are $1/\eta$-strongly convex. Thus, using first-order optimality conditions for p_t and p_{t+1}, we have:

$$F_t(p_{t+1}) \geq F_t(p_t) + \nabla F_t(p_t) \cdot (p_{t+1} - p_t) + \frac{1}{2\eta}\|p_{t+1} - p_t\|_1^2 \geq F_t(p_t) + \frac{1}{2\eta}\|p_{t+1} - p_t\|_1^2,$$

$$F_{t+1}(p_t) \geq F_{t+1}(p_{t+1}) + \nabla F_{t+1}(p_{t+1}) \cdot (p_t - p_{t+1}) + \frac{1}{2\eta}\|p_{t+1} - p_t\|_1^2 \geq F_{t+1}(p_{t+1}) + \frac{1}{2\eta}\|p_{t+1} - p_t\|_1^2.$$

Summing the two inequalities, we obtain

$$\begin{aligned}
\frac{1}{\eta}\|p_{t+1} - p_t\|_1^2 &\leq F_{t+1}(p_t) - F_t(p_t) + F_t(p_{t+1}) - F_{t+1}(p_{t+1}) \\
&= \left(\sum_{s:s+d_s=t} \hat{c}_s\right) \cdot (p_t - p_{t+1}) \\
&\leq \sum_i \left(\sum_{s:s+d_s=t} \hat{c}_{s,i}\right) |p_{t,i} - p_{t+1,i}| \\
&\leq \sum_i \left(\sum_{s:s+d_s=t} \tilde{c}_{s,i}\right) |p_{t,i} - p_{t+1,i}|,
\end{aligned}$$

where $\tilde{c}_{s,i}$ are the standard (unbiased) importance-weighted loss estimators. Taking expectations while using $\mathbb{E}[(\cdot)^2] \geq (\mathbb{E}[\cdot])^2$ and Hölder's inequality, we obtain

$$\begin{aligned}
\frac{1}{\eta}(\mathbb{E}\|p_{t+1} - p_t\|_1)^2 &\leq \frac{1}{\eta}\mathbb{E}\left[\|p_{t+1} - p_t\|_1^2\right] \\
&\leq \mathbb{E}\left[\sum_i \left(\sum_{s:s+d_s=t} c_{s,i}\right) \cdot |p_{t+1,i} - p_{t,i}|\right] \\
&\leq m_t \mathbb{E}\|p_{t+1} - p_t\|_1,
\end{aligned}$$

where $m_t = |\{s : s + d_s = t\}|$ is the number of observations that arrive on round t. Dividing through by the right-hand side of the inequality above, we obtain

$$\mathbb{E}\|p_{t+1} - p_t\|_1 \leq \eta m_t,$$

and using the triangle inequality we have

$$\mathbb{E}\|p_{t+d_t} - p_t\|_1 \leq \sum_{s=1}^{d_t} \mathbb{E}\|p_{t+s} - p_{t+s-1}\|_1 \leq \eta \sum_{s=1}^{d_t} m_{t+s-1} = \eta M_{t,d_t},$$

where M_{t,d_t} is the number of observations that arrive between rounds t and $d + d_t - 1$. Using Lemma C.7 in [19], we conclude the proof via

$$\mathbb{E}\left(\sum_{t=1}^{T} \|p_{t+d_t} - p_t\|_1\right) \leq \eta \sum_{t=1}^{T} M_{t,d_t} \leq \eta(D + T).$$

■

A.2 Proofs for Section 4.2

In this subsection, we provide the proofs of the lemmas required to derive regret guarantees for algorithm DA-FA (Algorithm 2), proving Theorem 6.

Recall the following regret decomposition,

$$\begin{aligned}\mathcal{R}_T = &\sum_{t=1}^{d_{\max}} (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \ell(x_t, \cdot) + \sum_{t=d_{\max}+1}^{T} (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot) \\ &+ \sum_{t=d_{\max}+1}^{T} p_t(\cdot) \cdot \Big(\ell(x_t, \cdot) - \hat{f}_t(x_t, \cdot)\Big) + \sum_{t=d_{\max}+1}^{T} p_\star(\cdot \mid x_t) \cdot \Big(\hat{f}_t(x_t, \cdot) - \ell(x_t, \cdot)\Big) \\ &+ \sum_{t=d_{\max}+1}^{T} (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \Big(\hat{f}_t(x_t, \cdot) - \hat{f}_{\tau^t}(x_t, \cdot)\Big).\end{aligned}$$

We bound each term individually in the following lemmas and claims, and then we combine all the bounds to derive Theorem 6.

Claim 15. *It holds that*

$$\sum_{t=1}^{d_{\max}} (p_t(\cdot) - p_\star(\cdot \mid x_t)) \cdot \ell(x_t, \cdot) \le d_{\max}.$$

Proof. Followed by the fact that for ℓ is bounded in $[0, 1]$. ■

Lemma 16 (Restatement of Lemma 7). *It holds that*

$$\sum_{t=d_{\max}+1}^{T} (p_t(\cdot) - p_\star(\cdot|x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot) \le \frac{T|\mathcal{A}|}{\gamma} - \sum_{t=d_{\max}+1}^{T} \sum_{a \in \mathcal{A}} \frac{p_\star(a|x_t)}{\gamma p_t(a)}.$$

Proof. For $t \in \{d_{\max}+1, d_{\max}+2, \ldots, T\}$, let $R_t(p)$ denote the objective of the convex minimization problem in Eq. (6), i.e,

$$R_t(p) = \sum_{a \in \mathcal{A}} p(a) \cdot \hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma} \sum_{a \in \mathcal{A}} \log(p(a)).$$

Hence,

$$(\nabla R_t(p))_a = \hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma p(a)}.$$

Since $p_\star(\cdot|x_t)$ is a feasible solution and p_t is the optimal solution, by first-order optimality conditions we have

$$\sum_{a \in \mathcal{A}} p_\star(a|x_t) \left(\hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma p_t(a)}\right) - \sum_{a \in \mathcal{A}} p_t(a) \left(\hat{f}_{\tau^t}(x_t, a) - \frac{1}{\gamma p_t(a)}\right) \ge 0,$$

Thus,

$$\sum_{a \in \mathcal{A}} (p_\star(a|x_t) - p_t(a)) \hat{f}_{\tau^t}(x_t, a) \ge \sum_{a \in \mathcal{A}} \frac{p_\star(a|x_t)}{\gamma p_t(a)} - \frac{|\mathcal{A}|}{\gamma}.$$

Which implies that

$$\sum_{a \in \mathcal{A}} (p_t(a) - p_\star(a|x_t)) \hat{f}_{\tau^t}(x_t, a) \le \frac{|\mathcal{A}|}{\gamma} - \sum_{a \in \mathcal{A}} \frac{p_\star(a|x_t)}{\gamma p_t(a)}.$$

We conclude that

$$\sum_{t=d_{\max}+1}^{T} (p_t - p_\star(\cdot|x_t)) \cdot \hat{f}_{\tau^t}(x_t, \cdot) \le \frac{T|\mathcal{A}|}{\gamma} - \sum_{t=d_{\max}+1}^{T} \sum_{a \in \mathcal{A}} \frac{p_\star(a|x_t)}{\gamma p_t(a)}.$$

■

Lemma 17 (Restatement of Lemma 8). *With probability at least $1-\delta/2$ it holds that*

$$\sum_{t=d_{\max}+1}^{T} p_t(\cdot)\cdot\Big(\ell(x_t,\cdot)-\hat{f}_t(x_t,\cdot)\Big)\le\frac{T|\mathcal{A}|}{\gamma}+\gamma\big(2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta})+16\log(4/\delta)\big).$$

Proof. For this term, we apply the oracle regret bound for the non-delayed function approximation. By Lemma 4 the following holds with probability at least $1-\delta/2$.

$$\begin{aligned}
&\sum_{t=d_{\max}+1}^{T} p_t(\cdot)\cdot\Big(\ell(x_t,\cdot)-\hat{f}_t(x_t,\cdot)\Big)\le\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\Big(\ell(x_t,a)-\hat{f}_t(x_t,a)\Big)=\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\sqrt{\frac{\gamma}{\gamma}}p_t(a)\Big(\ell(x_t,a)-\hat{f}_t(x_t,a)\Big)\le\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_t(a)}{\gamma}+\gamma\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\Big(\ell(x_t,a)-\hat{f}_t(x_t,a)\Big)^2= && \text{(AM-GM)}\\
&\frac{(T-d_{\max})|\mathcal{A}|}{\gamma}+\gamma\sum_{t=d_{\max}+1}^{T}\mathbb{E}_{a_t\sim p_t}\left[\Big(\hat{f}_t(x_t,a_t)-\ell(x_t,a_t)\Big)^2\right]\le\\
&\frac{(T-d_{\max})|\mathcal{A}|}{\gamma}+\gamma\sum_{t=1}^{T}\mathbb{E}_{a_t\sim p_t}\left[\Big(\hat{f}_t(x_t,a_t)-\ell(x_t,a_t)\Big)^2\right]\le\\
&\frac{T|\mathcal{A}|}{\gamma}+\gamma\big(2\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})+16\log(4/\delta)\big). && \text{(W.p. } 1-\delta/2)
\end{aligned}$$

■

Lemma 18 (Restatement of Lemma 9). *With probability at least $1-\delta/2$ it holds that*

$$\begin{aligned}\sum_{t=d_{\max}+1}^{T} p_\star(\cdot|x_t)\cdot\Big(\hat{f}_t(x_t,\cdot)-\ell(x_t,\cdot)\Big)\le&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star(a|x_t)}{\gamma p_t(a)}\\&+\gamma\big(2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta})+16\log(4/\delta)\big).\end{aligned}$$

Proof. For this term, we would like to use a change-of-measure technique using AM-GM to be able to apply the oracel's regret bound for the non-delayed function approximation. Again, by Lemma 4 the following holds with probability at least $1-\delta/2$.

$$\begin{aligned}
&\sum_{t=d_{\max}+1}^{T} p_\star(\cdot|x_t)\cdot\Big(\hat{f}_t(x_t,\cdot)-\ell(x_t,\cdot)\Big)\le\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}} p_\star(a|x_t)\cdot\Big(\hat{f}_t(x_t,a)-\ell(x_t,a)\Big)=\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}} p_\star(a|x_t)\sqrt{\frac{\gamma p_t(a)}{\gamma p_t(a)}}\cdot\Big(\hat{f}_t(x_t,a)-\ell(x_t,a)\Big)\le\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star^2(a|x_t)}{\gamma p_t(a)}+\gamma\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}} p_t(a)\Big(\hat{f}_t(x_t,a)-\ell(x_t,a)\Big)^2\le && \text{(AM-GM)}\\
&\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star(a|x_t)}{\gamma p_t(a)}+\gamma\sum_{t=d_{\max}+1}^{T}\mathbb{E}_{a_t\sim p_t}\left[\Big(\hat{f}_t(x_t,a_t)-\ell(x_t,a_t)\Big)^2\right]\le
\end{aligned}$$

$$\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star(a|x_t)}{\gamma p_t(a)}+\gamma\sum_{t=1}^{T}\mathbb{E}_{a_t\sim p_t}\left[\left(\hat{f}_t(x_t,a_t)-\ell(x_t,a_t)\right)^2\right]\leq$$
$$\sum_{t=d_{\max}+1}^{T}\sum_{a\in\mathcal{A}}\frac{p_\star(a|x_t)}{\gamma p_t(a)}+\gamma\big(2\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})+16\log(4/\delta)\big). \qquad \text{(W.p. at least } 1-\delta/2)$$

■

We now move to proving our final lemma.

Lemma 19 (Restatement of Lemma 10). *Under Assumption 5 it holds true that*

$$\sum_{t=d_{\max}+1}^{T}(p_t-p_\star(\cdot\mid x_t))\cdot\left(\hat{f}_t(x_t,\cdot)-\hat{f}_{\tau^t}(x_t,\cdot)\right)\leq 2\eta D.$$

Proof. We denote $\hat{f}_0:=\hat{f}_1$. Then we use Assumption 5 and Hölder's inequality to obtain

$$\sum_{t=d_{\max}+1}^{T}(p_t-p_\star(\cdot\mid x_t))\cdot\left(\hat{f}_t(x_t,\cdot)-\hat{f}_{\tau^t}(x_t,\cdot)\right)\leq$$
$$\sum_{t=d_{\max}+1}^{T}\|p_t-p_\star(\cdot\mid x_t)\|_1\cdot\|\hat{f}_t(x_t,\cdot)-\hat{f}_{\tau^t}(x_t,\cdot)\|_\infty\leq$$
$$2\sum_{t=d_{\max}+1}^{T}\sum_{i=1}^{d_{s^t_{\alpha(t)}}}\|\hat{f}_{t-i}(x_t,\cdot)-\hat{f}_{t-(i-1)}(x_t,\cdot)\|_\infty\leq \qquad (\tau^t=t-d_{s^t_{\alpha(t)}})$$
$$2\sum_{t=d_{\max}+1}^{T}d_{s^t_{\alpha(t)}}\eta\leq$$
$$2\eta D.$$

■

Finally, we came to prove Theorem 6.

Theorem 20 (Restatement of Theorem 6). *For any $\delta\in(0,1)$ let $\gamma=\sqrt{\frac{T|\mathcal{A}|}{2\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta})+16\log(4/\delta)}}$. Then, with probability at least $1-\delta$, the following regret bound holds.*

$$\mathcal{R}_T\leq\widetilde{O}\left(\sqrt{T|\mathcal{A}|\left(\mathcal{R}_T(\mathcal{O}_{sq}^{\mathcal{F},\eta})+\log(\delta^{-1})\right)}+\eta D+d_{\max}\right).$$

Proof of Theorem 6. Putting the results of Claim 15 and Lemmas 7 to 10 all together, with probability at $1-\delta$ the regret is bounded as follows.

$$\mathcal{R}_T\leq d_{\max}+2\frac{T|\mathcal{A}|}{\gamma}+2\gamma\big(2\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})+16\log(4/\delta)\big)+2\eta D.$$

Choosing $\gamma=\sqrt{\frac{T|\mathcal{A}|}{2\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F},\eta})+16\log(4/\delta)}}$ yields the desired bound. ■