Journal of Machine Learning Research 23 (2022) 1-43 Submitted 4/20; Revised 1/22; Published 3/22

No Weighted-Regret Learning in Adversarial Bandits with Delays

Ilai Bistritz* BISTRITZ@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Zhengyuan Zhou ZZHOU@STERN.NYU.EDU
Stern School of Business
New York University
New York, NY 10003, USA

Xi Chen† XC13@STERN.NYU.EDU
Stern School of Business
New York University
New York, NY 10003, USA

Nicholas Bambos BAMBOS@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Jose Blanchet JOSE.BLANCHET@STANFORD.EDU
Department of Management Science & Engineering
Stanford University
Stanford, CA 94305, USA

Editor: Csaba Szepesvari

Abstract

Consider a scenario where a player chooses an action in each round t out of T rounds and observes the incurred cost after a delay of d_t rounds. The cost functions and the delay sequence are chosen by an adversary. We show that in a non-cooperative game, the expected weighted ergodic distribution of play converges to the set of coarse correlated equilibria if players use algorithms that have "no weighted-regret" in the above scenario, even if they have linear regret due to too large delays. For a two-player zero-sum game, we show that no weighted-regret is sufficient for the weighted ergodic average of play to converge to the set of Nash equilibria. We prove that the FKM algorithm with n dimensions achieves an expected regret of $O\left(nT^{\frac{3}{4}}+\sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ and the EXP3 algorithm with K arms achieves an expected regret of $O\left(\sqrt{\log K\left(KT+D\right)}\right)$ even when $D=\sum_{t=1}^{T}d_t$ and T are unknown. These bounds use a novel doubling trick that, under mild assumptions, provably retains the regret bound for when D and T are known. Using these bounds, we show that FKM and EXP3 have no weighted-regret even for $d_t=O\left(t\log t\right)$. Therefore, algorithms with no weighted-regret can be used to approximate a CCE of a finite or convex unknown game that can only be simulated with bandit feedback, even if the simulation involves significant delays.

Keywords: Online Learning, Adversarial Bandits, Non-cooperative Games, Delays

∗. This is an extended version of Bistritz et al. (2019). For details, see Subsection 1.1 on previous work. This research was supported by the Koret Foundation grant for Smart Cities and Digital Living.

†. Xi Chen would like to thank the support from NSF via IIS-1845444.

1. Introduction

Consider an agent that makes sequential decisions, and each decision incurs some cost. The agent's goal is to minimize this cost over time. The question of **what** the agent learns about the cost functions naturally influences the best performance the agent can guarantee. With full information, after acting at round t, the agent receives the cost function of round t as feedback. With bandit feedback, as we consider here, the agent only receives the cost of her decision. Another fundamental question is **when** the agent receives the feedback. In most practical learning environments, an agent does not get to learn the cost of her action immediately. For example, it takes a while to observe the effect of a decision on a treatment plan or before observing the market's response to an advertisement. With delayed feedback, decisions must be made before all the feedback from the past choices is received.

Practical environments are non-stationary since they typically consist of other learning agents, and the learning of one agent affects that of the others. Moreover, the costs are naturally correlated over time. Hence, guarantees for stochastic environments are not strong enough for multi-agent environments. Instead, we consider cost sequences that are chosen by an adversary that knows the agent's algorithm. Proving regret bounds against an adversary certifies the robustness of a learning algorithm, regardless of whether an actual malicious adversary exists or not. Following the same reasoning, proving regret bounds with adversarial delays certifies the robustness of an algorithm to non-stationary delays.

An algorithm is said to have "no-regret" (Bowling, 2005) if it has a sublinear regret in T. It is well known that when N agents in a non-cooperative game each use an algorithm that has no-regret against an adaptive adversary, the ergodic distribution of play converges to the set of coarse correlated equilibria (CCE) (Hannan, 1957; Hart, 2013). For a two-player zero-sum game, the ergodic average of play converges to the set of Nash equilibria (NE) (Cai and Daskalakis, 2011). The emergence of a CCE or a NE in a game between no-regret learners establishes their role as predictors for the outcome of the game. From a practical point of view, the convergence of the expected ergodic distribution to the set of CCE or of the ergodic average to the set of NE makes no-regret algorithms an appealing way to approximate a CCE or a NE when the reward functions are unknown so only simulating the game is possible (see Hellerstein et al. (2019)). When simulating an unknown game, bandit feedback is a more realistic assumption than full information (or gradient feedback). Approximating the equilibrium can help to predict the outcome of the interaction between deployed agents even if they use other algorithms than those used for the approximation. If the equilibrium is globally efficient, cooperative agents may agree to play it after using no-regret algorithms to distributedly approximate it first.

The convergence to the set of CCE is maintained if the algorithm still enjoys the no-regret property even with delayed feedback. However, for large enough delays (e.g. $d_t = O(t \log t)$), the regret of any algorithm becomes linear in the horizon T so the no-regret property no longer holds. Our first main contribution in this paper is to show that even with delays that cause a linear regret, the expected weighted ergodic distribution may still converge to the set of CCE, and the weighted ergodic average may still converge (in L^1) to the set of NE for a two-player zero-sum game.

Many practical multi-agent interactions (i.e., games) are complicated to model. Instead, we can simulate the game based on data or in an experiment. Since the agents' performance is only measured in hindsight, such a simulation typically involves delays. If our simulated agents each run an online learning algorithm independently (e.g., FKM or EXP3), we can approximate a CCE of the game (NE for a two-player zero-sum game) by computing the weighted ergodic distribution (average). Our results imply that by properly tuning the weights, this method can approximate an equilibrium even when the standard regret is linear.

Our game-theoretic results motivate analyzing the weighted-regret as opposed to the classical regret when delays are involved. Hence, we study the weighted-regret of some widely-applied algorithms, for both a discrete action set $1, ..., K$ (i.e., arms in multi-armed bandits) and a convex and

compact action set $\mathcal{K} \subset \mathbb{R}^n$. For bandit convex optimization with a convex compact action set, the most widely used adversarial bandits learning algorithm is FKM (Flaxman et al., 2005). With no delays, the expected regret of FKM is $O\left(nT^{\frac{3}{4}}\right)$ where n is the dimension of $\mathcal{K}$. For the discrete case, the most popular adversarial bandits learning algorithm is EXP3 (Auer et al., 1995, 2002; Bubeck et al., 2012; Neu et al., 2010). With no delays, the expected regret of EXP3 is $O\left(\sqrt{TK\log K}\right)$.

Our second main contribution is to bound the expected weighted-regret of FKM against an adaptive or oblivious adversary. As a special case, we show that with an arbitrary and possibly unbounded sequence of delays d_t, FKM achieves an expected regret of $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} + |\mathcal{M}|\right)$ against an oblivious adversary, where $\mathcal{M} = \{t \,|\, t + d_t > T,\, t \in [1,T]\}$ is the set of rounds that their feedback is not received before round T. Our third main contribution is to bound the expected weighted-regret of EXP3 against an adaptive or oblivious adversary. As a special case, we show that with an arbitrary and possibly unbounded sequence of delays d_t, EXP3 achieves an expected regret of $O\left(\sqrt{\left(KT + \sum_{t\notin\mathcal{M}} d_t\right)\log K} + |\mathcal{M}|\right)$ against an oblivious adversary. Our weighted-regret bounds reveal for which delay sequences FKM and EXP3 enjoy the no weighted-regret property.

Like the horizon T, the sum of delays $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$ might be unknown to the decision-maker, which may need them to tune the algorithm. While the standard doubling trick (Cesa-Bianchi et al., 1997) can deal with an unknown T, it does not help with an unknown D. Our fourth main contribution is a general novel two-dimensional doubling trick where epochs are indexed by a "delay index" as well as a "time index". The delay index doubles every time the number of missing samples so far doubles, and the time index doubles with the rounds as usual. We show that under mild conditions, this novel doubling trick can be applied to any online learning algorithm with delayed feedback, beyond the case of adversarial bandits. We apply this result to achieve an expected regret of $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ for FKM and of $O\left(\sqrt{(KT+D)\log K}\right)$ for EXP3.

1.1 Previous Work

In recent years, learning with delayed feedback has attracted considerable attention, ranging from multi-armed bandits (Mandel et al. (2015)) to Markov decision processes (Neu et al. (2010)) and even distributed optimization (Agarwal and Duchi (2011)).

Most literature on learning with delayed feedback deals with multi-armed bandits, i.e., with a discrete set of actions. Fixed delays were considered in Weinberger and Ordentlich (2002) and Zinkevich et al. (2009). Stochastic rewards and stochastic i.i.d. delays have been considered in Pike-Burke et al. (2018). Stochastic i.i.d. delays with random missing samples have been considered in Vernade et al. (2017). Bandits with adversarial rewards but still stochastic i.i.d. delays were considered in Joulani et al. (2013). Cesa-Bianchi et al. (2019) considered an interesting case of communicating agents that cooperate to solve a common adversarial bandit problem, where the messages between agents may arrive after a bounded delay with a known bound d. Recently, advancements were made for the case of stochastic delays, studying arm-dependent delays (Manegueu et al. (2020)), linear bandits (Vernade et al. (2020)) contextual bandits (Zhou et al. (2019)), and reward-dependent delays (Lancewicki et al. (2021)). In contrast, in our scenario, the adversary chooses the delay sequence.

In Quanrud and Khashabi (2015), the case of adversarial delays with full information feedback has been considered, where the feedback is the costs of all arms (or the gradient of the cost function). Our goal is to study bandit feedback instead, motivated by the multi-agent scenario.

In Cesa-Bianchi et al. (2018), a different adversarial bandits with delayed feedback scenario has been studied, where all the feedback that is received at the same round is summed up and cannot be distinguished, and delays are bounded by d. For both the multi-armed and convex cases, Cesa-Bianchi et al. (2018) designed a wrapper algorithm and proved a regret bound for their delayed feedback scenario as a function of the regret of the algorithm being wrapped for the no-delay scenario.

For EXP3, the resulting regret bound is $O\left(\sqrt{dTK\log K}\right)$. Compared to their scenario, we consider time-stamped feedback with delays that can be unbounded.

Multi-agent learning and convergence to NE under delays have been considered in Zhou et al. (2017); Héliou et al. (2020) for variationally stable games and monotone games (Rosen (1965)). We study general non-cooperative games and convergence to the set of CCE under delays by generalizing the framework of no-regret learning to no weighted-regret learning. Our analysis applies to both the multi-armed bandit and bandit convex optimization cases.

While their focus is on monotone games, Héliou et al. (2020) also prove a regret bound for an FKM-type algorithm for a single agent against an adversary. It is assumed in Héliou et al. (2020) that the delay sequence is of the form $d_t = O(t^\alpha)$ for $\alpha < 1$, so not even a small subset of the samples can be delayed by $O(t)$ rounds. Their algorithm puts the received samples in a queue and only uses one sample per round regardless of how many samples were received this round. This is different than our FKM version which uses all samples upon reception. The expected regret bound of this queuing version of FKM proved in Héliou et al. (2020) is $O(T^{\frac{3}{4}} + T^{\frac{2}{3}}T^{\frac{\alpha}{3}})$, regardless of the sum of delays, so it is much looser than our bound. In particular, if the sequence of delays is mostly zeros, but once in L rounds equals to t^α, the sum of delays is arbitrarily smaller, depending on L.

This paper extends a preliminary conference version (Bistritz et al., 2019) that only analyzed EXP3 under delays. In this journal version, we also analyze FKM for bandit convex optimization under delays. Additionally, we improve the doubling trick of Bistritz et al. (2019) and show that it can be applied to any online learning algorithm with delayed feedback. Last but not least, we generalize the game-theoretic results of Bistritz et al. (2019) to non-cooperative games and the CCE.

While preparing this journal version, we became aware that the concurrent work of Thune et al. (2019) published in the same conference as (Bistritz et al., 2019) provides a similar analysis for the single-agent EXP3 case with a constant step-size $\eta_t = \eta$. Taking a different approach to deal with unknown $D = \sum_t d_t$ and T, Thune et al. (2019) assume that the delays are available at action time. In this work, we instead provide a novel doubling trick that does not require this assumption and achieves the same $O\left(\sqrt{(TK+D)\log K}\right)$ that was achieved in Thune et al. (2019). This improves the doubling trick that was proposed in Bistritz et al. (2019) that achieved an expected regret of $O\left(\sqrt{(TK^2+D)\log K}\right)$. Replacing EXP3 with a novel follow-the-regulated-leader algorithm, Zimmert and Seldin (2020) improved the expected regret to the optimal $O\left(\sqrt{TK+D\log K}\right)$ even when D is unknown without using a doubling trick.

While this paper was under review, György and Joulani (2020) have achieved $O\left(\sqrt{(TK+D)\log K}\right)$ regret for EXP3 by adaptively tuning the step-size. As opposed to our EXP3 bound, their regret bound has been shown to also hold with high probability. Furthermore, assuming a bound on the maximal delay (or that the delay is available at action time) György and Joulani (2020) propose a data-adaptive version of EXP3 which yields a regret that depends on the cumulative cost.

The works Thune et al. (2019); Zimmert and Seldin (2020); György and Joulani (2020); Bistritz et al. (2019) all study multi-armed bandits, while this paper also studies convex bandit optimization, using the FKM algorithm under delayed feedback. Moreover, Thune et al. (2019); Zimmert and Seldin (2020); György and Joulani (2020) only studied the single-agent problem while we are mainly motivated by the multi-agent problem, proving convergence of the expected weighted ergodic distribution to the set of CCE under delayed feedback. Our emphasis on the multi-agent case leads to two technical differences even in our single-agent results. First, we prove regret bounds also against an adaptive adversary that can choose the cost function in response to the players' past actions. This distinction between an oblivious and adaptive adversary is necessary to show convergence to the set of CCE. Additionally, our single-agent results are formulated using the "weighted-regret", which weights the costs of different turns according to a given weight sequence. Even for the EXP3 analysis, this formulation leads to several subtleties that did not arise in Thune et al. (2019); Zimmert and Seldin (2020); György and Joulani (2020) (e.g., in Lemma 8 and Lemma 9). Last but not

least, this paper provides a novel doubling trick that can deal with an unknown sum of delays in addition to the unknown horizon. Our novel doubling trick can be applied to any online learning algorithm under delayed feedback, beyond the case of adversarial bandits. For example, the preliminary version of this doubling trick was employed in Lancewicki et al. (2020) that proposed novel learning algorithms for delayed feedback in adversarial Markov decision processes.

1.2 Outline

Section 2 formalizes the general problem of learning with delayed bandit feedback and highlights our main results. Section 3 discusses the outcome of the interaction between multiple learners that are each subjected to a possibly different delay sequence. We extend the well-known connection between no-regret learning and CCE to learning with delayed feedback. Surprisingly, even algorithms that have linear regret under delays can still lead to the set of CCE. Section 4 presents our general doubling trick that can be applied to online learning algorithms with delayed feedback, not necessarily in adversarial or bandit feedback environments. Section 5 and Section 6 consider the FKM algorithm for adversarial bandit convex optimization and the EXP3 algorithm for adversarial multi-armed bandits, respectively. Section 5 and Section 6 each starts by proving expected weighted-regret bounds under delayed bandit feedback for the algorithm in consideration, both against an oblivious and an adaptive adversary. Next, we apply the result on our doubling trick for both FKM and EXP3 to obtain expected regret bounds for the case where T and D are unknown. Then, we show that FKM and EXP3 have no weighted-regret even with respect to delay sequences for which they both have linear regret in T. This allows us to apply our game-theoretic results for both FKM and EXP3, showing that they can approximate a CCE or a NE (in a two-player zero-sum game) of a simulated game where only delayed bandit feedback is available. Finally, Section 7 concludes the paper. Long proofs are postponed to the appendix.

2. Problem Formulation

Consider a player that in each round t from 1 to T picks an action $\boldsymbol{a}_t \in \mathcal{K}$ from a set $\mathcal{K}$. The cost at round t from playing $\boldsymbol{a}_t$ is $l_t(\boldsymbol{a}_t) \in [0,1]$. We consider two types of adversaries:

1. **Oblivious Adversary**: chooses the cost functions $l_1, ..., l_T$ before the game starts.
2. **Adaptive Adversary**: chooses the cost function l_t after observing $\{a_1, \dots a_{t-1}\}$, for each t.

With full information and no delays, the player gets to know the function l_t immediately after playing $\boldsymbol{a}_t$. In the bandit delayed feedback scenario, the player only gets to know the value of $l_t(\boldsymbol{a}_t)$ at the beginning of round $t + d_t$. (i.e., after a delay of $d_t \geq 1$ rounds). The adversary (oblivious or adaptive) chooses the delay sequence $\{d_t\}$ before the game starts.

We assume that the cost feedback includes the timestamp of the action that incurred this cost. This is indeed the case in many applications, such as when robots take physical actions, a recommendation is made to a customer or a treatment is given to a patient. If the delays are bounded by d, Cesa-Bianchi et al. (2018) have shown that EXP3 can still be implemented even with no timestamps, with expected regret $O(\sqrt{dTK \log K})$ instead of $O\left(\sqrt{(d+K)\,T \log K}\right)$. For our unbounded delays case, it is not clear if FKM or EXP3 can be implemented without timestamps.

The set of costs (feedback samples) received **and** used at round t is denoted $\mathcal{S}_t$, so $s \in \mathcal{S}_t$ means that the cost of $\boldsymbol{a}_s$ from round s is received and used at round t. Since the game lasts for T rounds, all costs for which $t + d_t > T$ are never received. Of course, the value of d_t does not matter as long as $t + d_t > T$, and these are just samples that the adversary chose to prevent the player from receiving. We name these costs the missing samples and denote their set by $\mathcal{M}$.

While the rounds of the game are indexed by t, it will be useful to our analysis to index a finer time scale that counts the steps of the algorithm for every such t. We define s_-, s_+ as the steps a

moment before and after the algorithm uses the feedback from round s, respectively. These steps are taking place in round t if $s \in \mathcal{S}_t$.

The player wants to have a learning algorithm that uses past observations to make good decisions over time. The performance of the player's algorithm is measured using the regret. The expected regret is the total expected cost over a horizon of T rounds, compared to the total expected cost that results from playing the best fixed action in hindsight in all rounds:

Definition 1. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{K}} \sum_{t=1}^{T} l_t(\boldsymbol{a})$. The expected regret is defined as

$$\mathbb{E}\{R(T)\} \triangleq \sum_{t=1}^{T} \mathbb{E}\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\} \tag{1}$$

where $\mathbb{E}$ is the expectation over the (possibly) random actions $\boldsymbol{a}_1, ..., \boldsymbol{a}_T$ of the player.

We analyze two widely applied algorithms for the two central special cases of the scenario above:

1. **Bandit Convex Optimization** - $\mathcal{K} \subset \mathbb{R}^n$ is a compact and convex set and $l_t : \mathcal{K} \to [0,1]$ is convex and Lipschitz continuous with parameter L. With no delays, the FKM algorithm, also known as "gradient descent without the gradient" (Flaxman et al., 2005), achieves an expected regret of $O\left(nT^{\frac{3}{4}}\right)$ for this problem.

2. **Multi-Armed Bandit** - $\mathcal{K} = \{1, ..., K\}$, $l_t : \{1, ..., K\} \to [0,1]$. With no delays, the EXP3 algorithm (Auer et al., 2002) achieves an expected regret of $O\left(\sqrt{TK\log K}\right)$ for this problem.

2.1 Results and Contribution

Our main results for the single-agent case are summarized and compared to the literature in Table 1. They are based on the regret bounds proven in Theorem 5 for FKM and Theorem 7 for EXP3 for an unknown T and unknown $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$.

The regret bound $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ reveals a remarkable robustness of FKM to delayed feedback. For the sequence $d_t = t^{\frac{1}{4}}$, the expected regret maintains the same $O\left(nT^{\frac{3}{4}}\right)$ as in the no-delay case. Even for $d_t = t^{\frac{4}{5}}$, the expected regret is $O\left(nT^{\frac{14}{15}}\right)$, so FKM still has no-regret.

Similarly, the regret bound $O\left(\sqrt{(TK + D)\log K}\right)$ reveals a significant robustness of EXP3 to delayed feedback. This follows since the T term is factored by K while the delay term D is not. Consider bounded delays of the form $d_t = K$. Then, the order of magnitude of the regret as a function of T and K is $O\left(\sqrt{TK\log K}\right)$, exactly as that of EXP3 without delays. For comparison, consider the full information case where at each round the costs of all arms are received. Assume that the player uses the exponential weights algorithm, which is the equivalent of EXP3 for the full information case. For the same delay sequence $d_t = K$, exponential weights achieves a regret bound of $O\left(\sqrt{TK\log K}\right)$, $\sqrt{K}$ times worse than the $O\left(\sqrt{T\log K}\right)$ it achieves with no delays.

Both bandit feedback and delays are obstacles that hurt the performance of the learning of the agent, as reflected in the expected regret. Surprisingly, even when the adversary has control over both of these obstacles, the degradation in the regret is mild. Intuitively, with bandit feedback, the effect of delay is much weaker than with full information since less information is delayed. This is an encouraging finding since practical systems typically have both bandit feedback and delays.

As a benchmark, we provide a simple lower bound of $\Omega\left(\sqrt{D}\right)$ on the expected regret of any algorithm, even for a given $D = \sum_{t=1}^{T} d_t$, for multi-armed bandits or convex bandit optimization. With no delays, i.e., $D = T$, the bound coincides with the existing lower bounds that it invokes.

	Convex Optimization		K Arms		
	OGD (Gradient Feedback)	FKM (Bandit Feedback)	Exponential Weights (Full Information)	EXP3 (Bandit Feedback)	FTRL (Bandit Feedback)
No-delay	$O\left(\sqrt{T}\right)$ Zinkevich (2003)	$O\left(nT^{3/4}\right)$ Flaxman et al. (2005)	$O\left(\sqrt{T\log K}\right)$	$O\left(\sqrt{TK\log K}\right)$ Auer et al. (1995)	$O\left(\sqrt{TK}\right)$
Adversarial Delays	$O\left(\sqrt{D}\right)$ Quanrud and Khashabi (2015)	$O\left(nT^{\frac{3}{4}}+\sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ Theorem 5	$O\left(\sqrt{D\log K}\right)$ Quanrud and Khashabi (2015)	$O\left(\sqrt{(TK+D)\log K}\right)$ Theorem 7 and Thune et al. (2019); György and Joulani (2020)	$O\left(\sqrt{TK+D\log K}\right)$ Zimmert and Seldin (2020)

Table 1: Expected regret for adversarial bandits (assuming all feedback is received before T, for the ease of comparison of results). For shorthand, we use $D=\sum_{t=1}^{T}d_t$.

For multi-armed bandits a tight lower bound of $\Omega\left(\sqrt{TK}+\sqrt{D\log K}\right)$ was shown in Zimmert and Seldin (2020) based on the bound in Cesa-Bianchi et al. (2019).

For the bandit convex optimization problem, FKM with no delays does not meet the lower bound of $\Omega(\sqrt{T})$. However, with delays, the logarithmic gap between our FKM upper bound $O(T^{\frac{3}{4}}+T^{\frac{1}{3}}D^{\frac{1}{3}})$ and the lower bound $\Omega(\sqrt{D})$ shrinks. For $D=T^{\frac{5}{4}}$ FKM guarantees $O(T^{\frac{6}{8}})$ instead of $\Omega(T^{\frac{5}{8}})$ and for $D=T^{\frac{3}{2}}$ FKM guarantees $O(T^{\frac{10}{12}})$ instead of $\Omega(T^{\frac{9}{12}})$.

Proposition 1. *Consider multi-armed bandits or convex bandit optimization, as defined above. Then for any algorithm and for any integer $D\geq T$, there exists a sequence of delays $\{d_t\}$ such that $D-\sqrt{2D}\leq\sum_{t=1}^{T}d_t\leq D$ and the expected regret with an oblivious adversary is $\Omega\left(\sqrt{D}\right)$.*

Proof Divide the time horizon into T_0 and $T_1=T-T_0$. Set $d_t=1$ for $1\leq t\leq T_0$ so there are no delays in this period. Set $d_t=T-t+1$ for $T_0<t\leq T$ so that the feedback in this period is never received. Then $\sum_{t=1}^{T}d_t=T_0+\sum_{t=T_0+1}^{T}(T-t+1)=T_0+\frac{(T-T_0)(T-T_0+1)}{2}$ so we can choose $T_0=\Theta(T-\sqrt{D})$ such that $D-\sqrt{2D}\leq T_0+\frac{(T-T_0)(T-T_0+1)}{2}\leq D$. The lower bound for either multi-armed bandits (Lattimore and Szepesvári, 2020, Theorem 15.2) or convex bandit optimization (Hazan, 2019, Theorem 3.2) is $\Omega\left(\sqrt{T_0}\right)$ for some cost functions $l_1,\ldots,l_{T_0}$. The state of the algorithm at round T_0 is the initial condition for another adversarial bandit problem where no feedback is received for an horizon of T_1 rounds. Hence, for any initial condition, there exists a cost function g such that if $l_{T_0+1}=...=l_T=g$, then the T_1 last rounds incur a regret of $\Theta(T_1)=\Theta\left(\sqrt{D-T}\right)$. ■

Our results for non-cooperative games with convex cost functions under delays are summarized in Table 2. They are based on the sufficient conditions for no weighted-regret for FKM (Lemma 4) and EXP3 (Lemma 6). Surprisingly, the delays do not have to be bounded for the convergence to the set of CCE to hold (or to the set of NE for a two-player zero-sum game), and they can even increase as fast as $d_t=O(t\log t)$. Moreover, the feedback of the players does not need to be synchronized, and they may be subjected to different delay sequences. If $\frac{d_t}{t}\to 0$ as $t\to\infty$ the convergence to the set of CCE follows from the sublinear regret of FKM and EXP3. This is no longer the case for $d_t=\Theta(t)$ or $d_t=\Theta(t\log t)$, where the regret of FKM, EXP3, or any other algorithm is $\Theta(T)$, so the learning against the adversary fails. Our results show that against other agents the situation is more optimistic, as the weighted ergodic average can still converge to the set of CCE (see Proposition 2 and Proposition 3). To achieve that, agents need to use a time-varying step-size η_t, as can be seen in Table 2. In fact, one can go up to $d_t=\Theta(t\log t\log(\log t))$ and continue iteratively in this manner, as long as $\sum_{t=1}^{\infty}\frac{1}{d_t}=\infty$. For larger delays, it is not possible to converge to the set of CCE or NE using our approach.

	$d_t \leq t^{\frac{1}{4}}$	$d_t \leq t^{\frac{3}{4}}$	$d_t \leq t$	$d_t \leq t \log t$
Parameters for no weighted-regret: FKM	$\eta_t = \frac{1}{t^{\frac{5}{8}} \log(t+1)}$ $\delta = T^{-\frac{1}{8}}$	$\eta_t = \frac{1}{t^{\frac{7}{8}} \log(t+1)}$ $\delta = T^{-\frac{1}{24}}$	$\eta_t = \frac{1}{t \log(t+1)}$ $\delta = (\log\log T)^{-\frac{1}{3}}$	$\eta_t = \frac{1}{t \log(t+1) \log\log(t+1)}$ $\delta = (\log\log\log T)^{-\frac{1}{3}}$
Distance from CCE: FKM	$O\left(\frac{\log T}{T^{\frac{1}{8}}}\right)$	$O\left(\frac{\log T}{T^{\frac{1}{24}}}\right)$	$O\left(\frac{1}{(\log\log T)^{\frac{1}{3}}}\right)$	$O\left(\frac{1}{(\log\log\log T)^{\frac{1}{3}}}\right)$
FKM Regret	$O\left(nT^{\frac{3}{4}}\right)$	$O\left(nT^{\frac{11}{12}}\right)$	$O(T)$	$O(T)$
Parameters for no weighted-regret: EXP3	$\eta_t = \frac{1}{t^{\frac{5}{8}} \log(t+1)}$	$\eta_t = \frac{1}{t^{\frac{7}{8}} \log(t+1)}$	$\eta_t = \frac{1}{t \log(t+1)}$	$\eta_t = \frac{1}{t \log(t+1) \log\log(t+1)}$
Distance from CCE: EXP3	$O\left(\frac{\log T}{T^{\frac{3}{8}}}\right)$	$O\left(\frac{\log T}{T^{\frac{1}{8}}}\right)$	$O\left(\frac{1}{\log\log T}\right)$	$O\left(\frac{1}{\log\log\log T}\right)$
EXP3 Regret	$O\left(T^{\frac{5}{8}}\sqrt{\log K}\right)$	$O\left(T^{\frac{7}{8}}\sqrt{\log K}\right)$	$O(T)$	$O(T)$

Table 2: Conditions for no weighted-regret for different delay sequences, along with the corresponding single agent expected regret bounds. For shorthand, we use $D = \sum_{t=1}^{T} d_t$.

The implication of our results is for approximating a CCE (NE) of a (two-player zero-sum) game that we can only simulate "in the lab" based on data or an experiment. In such a scenario, we can only evaluate the agents' performance (i.e., cost) based on the effect of their actions, which means delayed bandit feedback for the agents. We show that algorithms with no weighted-regret can approximate a CCE or a NE even with large delays that yield linear (trivial) regret.

3. Non-cooperative Games with Delayed Bandit Feedback

One of the main reasons why adversarial regret bounds are needed is that practical environments consist of multiple interacting agents, leading to non-stationary reward processes. In this section, we study a non-cooperative game where each player only receives delayed bandit feedback, given some arbitrary sequence of delays that can be different for different players.

It is well known that without delays, players that use an online learning algorithm with sublinear regret (i.e., no-regret) against an adaptive adversary will converge to the set of CCE in the empirical distribution sense (Hannan, 1957; Hart, 2013), and to the set of NE for a two-player zero-sum game (Blackwell, 1956). With large enough delays, the regret becomes linear in T so there is no guarantee that the dynamics converge to the set of CCE or NE in any sense. Surprisingly, we show that a CCE (or a NE for a two-player zero-sum game) can still emerge even with linear regret that results from too large delays. Our weighted-regret bounds for FKM and EPX3 provide sufficient conditions under which CCE or a NE can be approximated this way for a convex or finite game, respectively. In this sense, the weighting in the weighted ergodic distribution "filters out" part of the delay noise in the approximation of the CCE/NE. In this section, we formulate our results for general continuous games and then explain how finite games can be viewed as a special case, using mixed actions.

Our key observation is that with delayed feedback, it is not the regret that matters for the game dynamics but rather what we call the weighted-regret. The weighted-regret weighs the costs in different rounds according to a given non-increasing sequence η_t so it coincides with the regret when $\eta_t = \eta, \forall t$. We define "no weighted-regret" to replace the traditional no-regret property:

Definition 2. Let $\{l_t : \mathcal{K} \to [0,1]\}_t$ be a sequence of cost functions, chosen by an adaptive adversary. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathcal{K}} \sum_{t=1}^{T} \eta_t l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. We say that an algorithm that produces the random sequence of (single-agent) actions $\{\boldsymbol{a}_t\}$ has no weighted-regret with respect to $\{d_t\}$ and the non-increasing weight

sequence $\{\eta_t\}$ if

$$\lim_{T\to\infty} \mathbb{E}\left\{ \frac{\sum_{t=1}^{T} \eta_t \left(l_t\left(\boldsymbol{a}_t\right) - l_t\left(\boldsymbol{a}^*\right)\right)}{\sum_{t=1}^{T} \eta_t} \right\} = 0 \tag{2}$$

where the expectation is with respect to the random $\{\boldsymbol{a}_t\}$ generated by the algorithm.

Having no weighted-regret is only non-trivial when $\sum_{t=1}^{\infty} \eta_t = \infty$. When $\sum_{t=1}^{\infty} \eta_t < \infty$, the feedback from the last $\frac{T}{2}$ rounds can be discarded without affecting (2). When $\sum_{t=1}^{\infty} \eta_t = \infty$, there is no round t after which we can discard all feedback and still maintain (2).

We define no weighted-regret with respect to an adaptive adversary since, in a non-cooperative game with cost functions $\{u_n\}$, the "adversarial" cost function of player n is $l_t(\boldsymbol{a}_n) = u_n(\boldsymbol{a}_n, \boldsymbol{a}_{-n,t})$, so it is determined by the actions of the other players. In turn, the actions of the other players depend on the past actions of player n. Hence, the cost function of player n depends on her past actions, as with an adaptive adversary. In general, the equilibrium does not consist of absolute strategies such as "min-max" in zero-sum games. Hence, regret guarantees against an adaptive adversary allow proving convergence to the set of CCE for general non-cooperative games.

When taking the limit $T \to \infty$, it is important to emphasize that we do not change the infinite sequence of delays $\{d_t\}$, but only reveal more elements in this sequence. In other words, we are looking at the same game but over a longer time horizon. Therefore, while $d_t = \frac{T}{2}$ makes sense for a constant T, it is misleading when taking $T \to \infty$ since it represents a delay that occurred at time t but changes with the limit, so the limit is no longer of the "same game".

3.1 Coarse Correlated Equilibrium for N-player Games

The CCE is a well-established equilibrium concept for learning in games (Hannan, 1957; Ashlagi et al., 2008; Hart, 2013). Our convergence argument utilizes the notion of an ε-CCE:

Definition 3. Consider a non-cooperative game where the action set of all players is some compact set $\mathcal{A}$ and the reward function of each player $u_n : \mathcal{A}^N \to [0,1]$ is continuous. Recall that $\boldsymbol{a}_{-n}$ is the action profile of all players except player n. Let $\mathcal{P}\left(\mathcal{A}^N\right)$ be the set of all Borel probability measures over $\mathcal{A}^N$, equipped with the weak-* topology (see Simmons (1963)). The set of all ε-CCE points is the set of distributions over $\mathcal{A}^N$ such that:

$$\mathcal{C}_\varepsilon = \left\{ \rho \in \mathcal{P}\left(\mathcal{A}^N\right) \,\middle|\, \mathbb{E}^{\boldsymbol{a}^*\sim\rho}\left\{u_n\left(a_n^*, \boldsymbol{a}_{-n}^*\right)\right\} \geq \max_{a_n \in \mathcal{A}} \mathbb{E}^{\boldsymbol{a}^*\sim\rho}\left\{u_n\left(a_n, \boldsymbol{a}_{-n}^*\right)\right\} - \varepsilon\,, \forall n \right\} \tag{3}$$

and the set of CCE points is $\mathcal{C}_0$ with $\varepsilon = 0$.

The $\varepsilon-$CCE is a distribution over the action profiles such that no player can improve her expected reward by more than ε by playing any pure action if other players keep playing according to this distribution. The CCE can be interpreted as a coordinator that uses a random signal to instruct the players what to play such that they all want to follow this instruction given that the others do. This equilibrium is called "correlated" since the actions of the players are statistically dependent, potentially through the coordinator's signal they all observe. The history of the game, and even as little as the bandit feedback each player received in the past, can implement such a coordinator.

The CCE should not be confused with the more refined correlated equilibrium (CE). The CCE coincides with the CE if only constant departure functions are considered in the definition of the CE (Hart, 2013, Page 11), instead of all measurable mappings. Hence, by definition, every CE is a CCE. We focus on the CCE since it relates to the regret (Definition 1) directly, while the CE is related to the internal regret which is less common for online learning algorithms (Blum and Mansour (2007)).

In the game of Definition 3 a CCE always exists since Theorem (Hart, 2013, Theorem 3) shows that a correlated equilibrium exists. Hence, $\mathcal{C}_0$ is non-empty.

The entity that converges to the set of CCE $\mathcal{C}_0$ in our non-cooperative game scenario is the expected weighted ergodic distribution of the actions $\boldsymbol{a}_t$. For the special case of $\eta_t = \eta$ for all t for some $\eta > 0$, the weighted ergodic distribution of $\boldsymbol{a}_t$ is simply its ergodic (i.e., empirical) distribution.

Definition 4. For a weight sequence $\{\eta_t\}$ and horizon T, the weighted ergodic distribution of a sequence of actions $\{\boldsymbol{a}_t\}$ is defined as:

$$\rho_T \triangleq \frac{\sum_{t=1}^{T} \eta_t \delta_{\boldsymbol{a}_t}}{\sum_{t=1}^{T} \eta_t} \tag{4}$$

where $\delta_{\boldsymbol{a}_t}$ is Dirac's measure, so $\delta_{\boldsymbol{a}_t}(A) = 1$ if $\boldsymbol{a}_t \in A$ and $\delta_{\boldsymbol{a}_t}(A) = 0$ otherwise. Note that $\mathbb{E}\{\delta_{\boldsymbol{a}_t}\} = \mathbb{E}\{p_{\boldsymbol{a}_t}\}$ where $p_{\boldsymbol{a}_t}$ is the distribution of $\boldsymbol{a}_t$ given the information the algorithm has at round t. Hence, we can use $\frac{\sum_{t=1}^{T} \eta_t p_{\boldsymbol{a}_t}}{\sum_{t=1}^{T} \eta_t}$ instead to estimate a CCE, which exploits more information.

The following theorem establishes the convergence of $\mathbb{E}\{\rho_T\}$ to the set of CCE $\mathcal{C}_0$.

Theorem 1. *Let N players play a non-cooperative game where the action set of all players $\mathcal{A}$ is compact and the reward function of each player $u_n : \mathcal{A}^N \to [0,1]$ is continuous. Let $\{\eta_t\}$ be the non-increasing weight sequence. If each player n runs an algorithm that has no weighted-regret with respect to its delay sequence $\{d_t^n\}_t$ then $\mathbb{E}\{\rho_T\}$ converges to the set of CCE $\mathcal{C}_0$ as $T \to \infty$.*

Proof See Appendix. ■

The expectation $\mathbb{E}\{\rho_T\}$ is with respect to the random actions. By definition, the set of ε-CCE is convex, so the average of multiple ε-CCE is also an ε-CCE. Hence, the implication of Theorem 1 is that to approximate a CCE using T samples, one can run $\sqrt{T}$ independent simulations of the game (possibly not identically distributed) and then average the resulting $\left\{\rho_{\sqrt{T}}^{(i)}\right\}_{i=1}^{\sqrt{T}}$. From the strong law of large numbers, this estimation converges as $T \to \infty$ with probability 1 to $\mathcal{C}_0$ since $\rho_{\sqrt{T}}^{(i)}$ is bounded and $\left\{\rho_{\sqrt{T}}^{(i)}\right\}_{i=1}^{\sqrt{T}}$ are independent.

Remark 1 (Finite Games)**.** *One useful special case of Theorem 1 is that of finite games (i.e., multi-armed bandits). To see that, choose the action set to be the K-dimensional simplex, i.e., $\mathcal{A} = \Delta^K$. Let $U_n : \{1,\dots,K\}^N \to [0,1]$ be the reward function of player n in the finite game. Then we can define the reward function of the continuous game $u_n : \mathcal{A}^N \to [0,1]$ for every $\boldsymbol{x} \in \Delta^K$ as follows:*

$$u_n(\boldsymbol{x}) \triangleq \mathbb{E}^{\boldsymbol{a}\sim\boldsymbol{x}}\{U_n(\boldsymbol{a})\} \tag{5}$$

where the expectation averages $\boldsymbol{a} \in \{1,\dots,K\}^N$ according to the distribution that $\boldsymbol{x}$ defines on $1,\dots,K$. This is indeed a special case of our formulation above since the simplex Δ^K is compact, and $u_n(\boldsymbol{x})$ is linear and therefore continuous. Moreover, we have

$$\mathbb{E}^{\boldsymbol{x}^*\sim\rho}\{u_n(\boldsymbol{x}^*)\} = \mathbb{E}^{\boldsymbol{x}^*\sim\rho}\left\{\mathbb{E}^{\boldsymbol{a}^*\sim\boldsymbol{x}^*}\{U_n(\boldsymbol{a}^*)\}\right\} = \mathbb{E}^{\boldsymbol{a}^*\sim\phi(\rho,\boldsymbol{x}^*)}\{U_n(\boldsymbol{a}^*)\} \tag{6}$$

where $\phi(\rho,\boldsymbol{x})$ is the distribution over $\{1,\dots,K\}$ that is induced by randomizing $\boldsymbol{x}$ according to ρ and then randomizing $\boldsymbol{a}$ according to $\boldsymbol{x}$ (i.e., the compound distribution with $\boldsymbol{x}$ as the random parameter). Since the maximum of u_n is attained at the corners of the simplex, we also have

$$\max_{\boldsymbol{x}_n\in\mathcal{A}} \mathbb{E}^{\boldsymbol{x}^*\sim\rho}\left\{u_n\left(\boldsymbol{x}_n, \boldsymbol{x}_{-n}^*\right)\right\} = \max_{a_n\in\{1,\dots,K\}} \mathbb{E}^{\boldsymbol{a}^*\sim\phi(\rho,\boldsymbol{x}^*)}\left\{U_n\left(a_n, \boldsymbol{a}_{-n}^*\right)\right\}. \tag{7}$$

Due to (6) and (7), Definition 3 coincides with the definition of a CCE for a finite game, which results from Definition 3 with $\mathcal{A} = \{1,\dots,K\}$ and any arbitrary reward functions $\{u_n : \{1,\dots,K\}^N \to [0,1]\}_n$. Then $\mathcal{P}\left(\mathcal{A}^N\right)$ takes the form of the K-dimensional simplex.

It is important to notice that FKM cannot be applied in a finite game with bandit feedback. While FKM can certainly be applied to linear cost functions, it would require the player to obtain $u_n(\boldsymbol{x})$ as feedback, which is the expected cost incurred to a player that only picks one discrete arm each turn at random, according to a distribution $\boldsymbol{x}_n$. For this reason, we also analyze the weighted-regret of

the EXP3 algorithm, which can work with discrete arm choices and bandit feedback. Remarkably, requiring less feedback, EXP3 also achieves better expected regret for this linear case. If $l_t^{(i)}$ is the cost of playing arm i at round t, then the expected cost of playing a random arm a_t according to the distribution $\boldsymbol{x}_t \in \Delta^K$ is

$$\mathbb{E}\left\{l_t^{(a_t)}\right\} = \mathbb{E}\left\{\mathbb{E}\left\{\sum_{i=1}^{K} x_t^{(i)} l_t^{(i)} \mid \boldsymbol{x}_t\right\}\right\} = \mathbb{E}\left\{\sum_{i=1}^{K} x_t^{(i)} l_t^{(i)}\right\}. \tag{8}$$

Hence, EXP3 also leads to faster convergence to the set of CCE in finite games.

The general result of Theorem 1 implies stronger results for special classes of games where the set of CCE has an interesting structure. For example, in strictly monotone games the unique CCE places probability one on the unique pure NE (Ui (2008)). Another example is a polymatrix game, which is a finite action set game where each player plays a separate two-player zero-sum game against each of her neighbors on a given graph. For polymatrix games, for which a two-player zero-sum game is a special case, it was shown in Cai and Daskalakis (2011) that the marginal distributions of the CCE are a NE. However, this result holds only for multi-armed bandits since it assumes discrete action sets. Our next section establishes that the weighted ergodic average of two no weighted-regret algorithms in a two-player zero-sum game converges to the set of NE for multi-armed bandits **and** bandit convex optimization.

3.2 Nash Equilibrium for Two-Player Convex-Concave Zero-Sum Games

In this subsection, we consider two-player zero-sum games where the action set $\mathcal{A} \subset \mathbb{R}^n$ of both players is convex and compact. The cost function $u : \mathcal{A} \times \mathcal{A} \to [0, 1]$ is assumed to be continuous, convex in the first argument and concave in the second. When the row player plays $\boldsymbol{y}$ and the column player plays $\boldsymbol{z}$, the first pays a cost of $u(\boldsymbol{y}, \boldsymbol{z})$ and the second gains a reward of $u(\boldsymbol{y}, \boldsymbol{z})$.

For two-player zero-sum games, we show that algorithms with no weighted-regret lead to the set of Nash equilibria (NE). A NE is an action profile such that no player wants to switch an action given that the other players keep their actions. For our convergence argument, we define the set of all approximate (pure) NE of a two-player zero-sum game:

Definition 5. Define a two-player zero-sum game where the action set of both players is $\mathcal{A} \subset \mathbb{R}^n$ and the cost function is $u : \mathcal{A} \times \mathcal{A} \to [0, 1]$. The set of all ε-NE points of this game is defined as:

$$\mathcal{N}_\varepsilon = \left\{(\boldsymbol{y}^*, \boldsymbol{z}^*) \in \mathcal{A} \times \mathcal{A} \mid u(\boldsymbol{y^*}, \boldsymbol{z^*}) \leq \min_{\boldsymbol{y} \in \mathcal{A}} u(\boldsymbol{y}, \boldsymbol{z^*}) + \varepsilon, u(\boldsymbol{y^*}, \boldsymbol{z^*}) \geq \max_{\boldsymbol{z} \in \mathcal{A}} u(\boldsymbol{y^*}, \boldsymbol{z}) - \varepsilon\right\} \tag{9}$$

and the set of NE points is $\mathcal{N}_0$ with $\varepsilon = 0$.

The NE is a more exclusive solution concept than the CCE, so our result is stronger for this special case of a two-player zero-sum game. This holds since if a NE exists, it is always a CCE, which follows immediately from Definition 3 and Definition 5 by substituting the distribution that gives the NE action profile with probability 1 as the distribution ρ. For a two-player zero-sum game with a convex and continuous cost function and compact action sets, a pure NE always exists (Nikaidô and Isoda, 1955; Debreu, 1952) so $\mathcal{N}_0$ is non-empty.

It was shown in Bailey and Piliouras (2018) that for the no-delay case, the last iterate $(\boldsymbol{y}_t, \boldsymbol{z}_t)$ does not converge in general to a NE and even moves away from it. Instead, it is the ergodic average action that converges to the set of NE $\mathcal{N}_0$. With delayed feedback, the entity that converges to $\mathcal{N}_0$ in our two-player zero-sum game scenario is the weighted ergodic average of the actions $\{\boldsymbol{a}_t\}_t$. For the special case of $\eta_t = \eta$ for all t for some $\eta > 0$, the weighted ergodic average of $\boldsymbol{a}_t$ is just its ergodic average.

Definition 6. For a weight sequence $\{\eta_t\}$ and horizon T, the weighted ergodic average of a sequence of actions $\{\boldsymbol{a}_t\}$ is defined as

$$\bar{\boldsymbol{a}}_T \triangleq \frac{\sum_{t=1}^{T} \eta_t \boldsymbol{a}_t}{\sum_{t=1}^{T} \eta_t}. \tag{10}$$

Then, the following theorem establishes the convergence to the set of NE $\mathcal{N}_0$. Following Remark 1, we can apply Theorem 2 to a finite game with cost table $U : \{1,\dots,K\} \times \{1,\dots,K\} \to [0,1]$ such that $\boldsymbol{y}, \boldsymbol{z}$ are mixed actions (distributions over $\{1,\dots,K\}$), and $u(\boldsymbol{y},\boldsymbol{z}) \triangleq \sum_{i=1}^{K}\sum_{j=1}^{K} y^{(i)} z^{(j)} U(i,j)$.

Since the players' payoffs ("value of the game") are the same in all NE of a two-player zero-sum game (Cesa-Bianchi and Lugosi, 2006, Page 182), algorithms with no weighted-regret can be used to approximate this outcome, regardless of the NE that the weighted ergodic average approximates.

Theorem 2. *Let two players play a zero-sum game with a convex and compact action set $\mathcal{A} \subset \mathbb{R}^n$ and a cost function $u(\boldsymbol{y},\boldsymbol{z}) : \mathcal{A} \times \mathcal{A} \to [0,1]$. Assume that $u(\boldsymbol{y},\boldsymbol{z})$ is convex in $\boldsymbol{y}$ and concave in $\boldsymbol{z}$ and is continuous. Let $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$ be the actions of the row and column players at round t, and let $\bar{\boldsymbol{y}}_T$ and $\bar{\boldsymbol{z}}_T$ be their weighted ergodic averages. Let $\{\eta_t\}$ be the non-increasing weight sequence. Let $\{d_t^r\}$ and $\{d_t^c\}$ be the delay sequence of the row player and the column player. If both players use a no weighted-regret algorithm with respect to $\{d_t^r\}, \{d_t^c\}$ then, as $T \to \infty$:*

1. *$(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T)$ converges in L^1 to the set of Nash equilibria $\mathcal{N}_0$ of the two-player zero-sum game.*
2. *$U(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T)$ converges in L^1 to the value of the game $\min\limits_{\boldsymbol{y}}\max\limits_{\boldsymbol{z}} U(\boldsymbol{y},\boldsymbol{z}) = \max\limits_{\boldsymbol{z}}\min\limits_{\boldsymbol{y}} U(\boldsymbol{y},\boldsymbol{z})$.*

Proof See Appendix. ■

4. Doubling Trick for Online Learning with Delays

Online learning under delayed feedback introduces another key parameter, which is the sum of delays $D = \sum_{t=1}^{T} d_t$. The sum of delays appears in the expected regret bound of many online algorithms and is required to tune their step-size and other parameters. If D or a tight upper bound for it is unknown, then an adaptive algorithm is needed.

With no delays, the standard doubling trick (see Cesa-Bianchi et al. (1997)) can be used if T is unknown. However, the same doubling trick does not work with delayed feedback. We now present a novel doubling trick for the delayed feedback case, where T and D are unknown.

Our enhanced doubling trick is two-dimensional, as each epoch is indexed by a delay index w as well as a time index h. Compared to that, our previous doubling trick in Bistritz et al. (2019) only tracked a delay index w. As a result, the new doubling trick leads to tighter regret bounds. For example, it yields $O\left(\sqrt{(TK+D)\log K}\right)$ instead of $O\left(\sqrt{(TK^2+D)\log K}\right)$ for EXP3. More importantly, the doubling trick in Bistritz et al. (2019) required to analyze the regret of the algorithm that employed it. In comparison, the new doubling trick is plug and play since it provably retains the regret bound for when D and T are known for a wide class of online learning algorithms.

Our doubling trick assumes that a regret bound of the form $O\left(k_1 D^d + k_2 T^c + k_3 T^a D^b\right)$ is available if T, D are known, for some constants $k_1, k_2, k_3 \geq 0$ and $0 \leq a, b, c, d \leq 1$. We divide the time horizon into super-epochs, indexed by ν, each consists of epochs as explained below. A super-epoch is a set $\mathcal{E}_\nu$ of consecutive rounds that all use the same algorithm parameters $\mathcal{P}_\nu$ (e.g., a step-size η_ν). Let ν_t be the index of the super-epoch that contains round t. Let m_t be the number of missing feedback samples at round t. A missing feedback sample at round t is a sample from $\tau \leq t$ such that $\nu_t = \nu_\tau$ (i.e., belongs to the same super-epoch) that was not received before or at round t.

An epoch is the set of consecutive rounds where the sum of delays is within a given interval and the time index is within another given interval. To enable that, we employ a delay index counter w

and a time index counter h. We increase w every time when $\sum_{\tau=1}^{t} m_\tau$, that tracks D, doubles. We increase h every time when the number of rounds t doubles. We then define the (h,w) epoch as

$$\mathcal{T}_{h,w} = \left\{ t \,|2^{w-1} \leq \sum_{\tau=1}^{t} m_\tau < 2^w,\ 2^{h-1} \leq t < 2^h \right\}. \tag{11}$$

The (h,w) epochs are then partitioned into super-epochs as follows (also illustrated in Fig. 1):

- For each h, the super-epoch $\mathcal{E} = \mathcal{S}_h$ is the set of (h,w) such that $2k_2 2^{ch} \geq k_1 2^{dw} + k_3 2^{ah} 2^{bw}$.
- For each w, the super-epoch $\mathcal{E} = \mathcal{S}_w$ is the set of (h,w) such that $2k_1 2^{dw} \geq k_2 2^{ch} + k_3 2^{ah} 2^{bw}$.
- All other epochs $(h,w) \notin \mathcal{S}_h \cup \mathcal{S}_w$ are each a separate super-epoch $\mathcal{E} = \{(h,w)\}$.

During the $(h,w) \in \mathcal{E}_\nu$ epoch, the algorithm equipped with our doubling trick uses the parameters $\mathcal{P}_\nu$ (e.g., $\mathcal{P}_\nu = \{\eta_\nu, \delta_\nu\}$ for FKM). Different epochs (h_1,w_1) and (h_2,w_2) in the same super-epoch use the same set of parameters. As shown in Fig. 1, this can only be the case if $h_1 = h_2$ or $w_1 = w_2$. Feedback samples from previous super-epochs are discarded once received, and are no longer counted in $\sum_{\tau=1}^{t} m_\tau$ after their super-epoch has ended. The resulting algorithm is detailed in Algorithm 1.

Fig. 1 illustrates the partition of epochs into super epochs used for our doubling trick. We can see that the (h,w) epoch space is split into three regions with three different types of super-epochs. In the upper one (in blue) D^d dominates the regret, in the middle one (in pink) $T^a D^b$ dominates the regret and in the lower one (in orange) T^c dominates the regret. Each blue, pink, or orange box is a super-epoch on which we apply our regret bound separately. The grey arrows represent the actual path that the epoch indices $\{(h,w)\}$ went through.

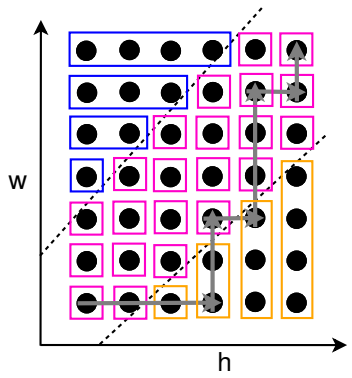

Figure 1: 2D Doubling Trick. Each box is a different super-epoch, and each dot is a different epoch.

The next Lemma proves that our doubling trick tracks D, similar in spirit to how the standard doubling trick tracks T up to a factor of 2. It also bounds the largest delay and time indices possible.

Lemma 1. *Let H, W be the last time and delay indices, respectively. Let $\mathcal{S}_w, \mathcal{S}_h$ be super-epochs, as defined below* (11)*. Let $\mathcal{T}_w$ be the set of all rounds in $\mathcal{S}_w$ and $\tau_w \triangleq \max_{t \in \mathcal{T}_w} t$. Let W_h be the maximal w such that $(h,w) \in \mathcal{S}_h$. Let $\mathcal{T}_h$ be the set of all rounds in $\mathcal{S}_h$ and $\tau_h \triangleq \max_{t \in \mathcal{T}_h} t$. Algorithm 1 maintains:*

1. *For every w and h, $\sum_{t \in \mathcal{T}_w} \min\{d_t, \tau_w - t + 1\} \leq 2^{w-1}$ and $\sum_{t \in \mathcal{T}_h} \min\{d_t, \tau_h - t + 1\} \leq 2^{W_h}$.*
2. *$W \leq \log_2 \sum_{t=1}^{T} \min\{d_t, T - t + 1\} + 1$ and $H \leq \log_2(T+2) - 1$.*

Algorithm 1 2D Doubling Trick Wrapper for Unknown T and D

Initialization: Set $h=1, w=1, \nu=1$. Choose an algorithm $\mathrm{Alg}\left(\mathcal{P}_\nu\right)$, where $\mathcal{P}_\nu$ is the set of the algorithm parameters. Initialize $\boldsymbol{p}_1, \mathcal{P}_1$. Let $R\left(T, D, \mathcal{P}^*\left(T, D\right)\right) = k_1 D^d + k_2 T^c + k_3 T^a D^b$ be an expected regret bound for when D, T are known, that applies to $\mathrm{Alg}\left(\mathcal{P}^*\left(T, D\right)\right)$. Divide the (h, w) space into super-epochs as detailed below (11).
For $t = 1, ..., T$ **do**

1. Pick an action $\boldsymbol{a}_t \in \mathcal{K}$ at random according to the distribution $\boldsymbol{p}_t$.
2. Let $\tilde{\mathcal{C}}_t$ be the set of delayed costs $l_s\left(\boldsymbol{a}_s\right)$ received at round t such that $\nu_s = \nu_t$ (i.e., originated in the current super-epoch). Calculate the number of missing samples at round t by computing the difference between the number of rounds since the beginning of the super-epoch and the number of samples from this super-epoch received so far:
$$m_t = \left(t - \min_{\tau \in \{t' \,|\, \nu_t = \nu_{t'}\}} \tau + 1\right) - \sum_{\tau \in \{t' \,|\, \nu_t = \nu_{t'}\}} \left|\tilde{\mathcal{C}}_\tau\right|. \tag{12}$$
3. Update the delay index: if $\sum_{\tau=1}^{t} m_\tau \geq 2^w$, then update $w \leftarrow w + 1$.
4. Update the time index: if $t \geq 2^h$, then update $h \leftarrow h + 1$.
5. Start a new super-epoch: if the new h, w indices are outside the current super-epoch, i.e., $(h, w) \notin \mathcal{E}_\nu$, then start a new super-epoch with parameters $\mathcal{P}_{\nu+1} \leftarrow \mathcal{P}^*\left(2^{h_{\nu+1}}, 2^{w_{\nu+1}}\right)$, where h_ν, w_ν are the maximal h, w indices in super-epoch $\mathcal{E}_\nu$:
   (a) Initialize the algorithm with these parameters, i.e., $\mathrm{Alg}\left(\mathcal{P}_{\nu+1}\right)$.
   (b) Increase the super-epoch index $\nu \leftarrow \nu + 1$.
6. Using only the samples in $\tilde{\mathcal{C}}_t$, update the distribution $\boldsymbol{p}_t$ according to $\mathrm{Alg}\left(\mathcal{P}_\nu\right)$.

End

Proof Let $\mathcal{M}_{\mathcal{S}_w}$ be the set of feedback samples for costs in $\mathcal{S}_w$ that are not received within $\mathcal{S}_w$. Every round $t \in \mathcal{T}_w$ such that $t \notin \mathcal{M}_{\mathcal{S}_w}$ contributes exactly d_t to $\sum_{t \in \mathcal{T}_w} m_t$, since the t-th feedback is missing for d_t rounds in $\mathcal{T}_w$. Every round $t \in \mathcal{M}_{\mathcal{S}_w}$ contributes $\tau_w - t + 1 \leq d_t$ to $\sum_{t \in \mathcal{T}_w} m_t$ before it stops being counted. Therefore

$$\sum_{t \in \mathcal{T}_w} \min\left\{d_t, \tau_w - t + 1\right\} \leq \sum_{t \in \mathcal{T}_w} m_t \underset{(a)}{\leq} 2^{w-1} \tag{13}$$

where (a) follows since if $\sum_{t \in \mathcal{T}_w} m_t > 2^{w-1}$ then $\sum_{\tau=1}^{t} m_\tau \geq 2^{w-1} + 2^{w-1} = 2^w$ and the delay index w should have already increased to $w+1$. Applying the same argument on $\mathcal{T}_h$, we obtain that

$$\sum_{t \in \mathcal{T}_h} \min\left\{d_t, \tau_h - t + 1\right\} \leq \sum_{t \in \mathcal{T}_h} m_t \underset{(a)}{\leq} 2^{W_h} \tag{14}$$

where (a) follows since if $\sum_{t \in \mathcal{T}_h} m_t > 2^{W_h}$ then $\sum_{\tau=1}^{t} m_\tau \geq 2^{W_h}$ so w must have increased to $W_h + 1$.

For the second part of the lemma, $2^{H+1} - 2 = \sum_{h=1}^{H} 2^h \leq T$ so $H \leq \log_2\left(T+2\right) - 1$, and

$$\sum_{t=1}^{T} \min\left\{d_t, T - t + 1\right\} \underset{(a)}{\geq} \sum_{t=1}^{T} m_t \geq 2^{W-1} \tag{15}$$

where (a) uses that every sample is counted in $\sum_{t=1}^{T} m_t$ for at most d_t or $T-t+1$ rounds. ■

Now we can prove our main result of this section. The first assumption is merely the regret bound that one can obtain if T and D are known. The second assumption says that holding the parameters $\mathcal{P}$ fixed, the regret bound is non-decreasing with T and D. As we see in Section 5 and Section 6, these basic assumptions hold for FKM and EXP3.

Theorem 3. *Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Let T be the time horizon and define $D=\sum_{t=1}^{T}\min\{d_t, T-t+1\}$. Let $R(T,D,\mathcal{P}(T,D))$ be an upper bound on the expected regret (Definition 1) of an online learning algorithm that uses the parameters $\mathcal{P}(T,D)$. Assume:*

1. *There exists a sequence $\mathcal{P}^*(T,D)$ and constants $k_1,k_2,k_3\geq 0$ and $0\leq a,b,c,d\leq 1$ such that for all T,D:*
$$R(T,D,\mathcal{P}^*(T,D)) \leq k_1 D^d + k_2 T^c + k_3 T^a D^b. \tag{16}$$
2. *For a fixed $\mathcal{P}^*$, $R(T,D,\mathcal{P}^*)$ is non-decreasing with T and D.*

Then if Algorithm 1 is used to track $\mathcal{P}^*(T,D)$, it achieves a total expected regret of
$$R(T,D) = O\left(k_1 D^d + k_2 T^c + k_3 T^a D^b\right). \tag{17}$$

Proof We apply the regret bound for each of the super-epoch types $\mathcal{S}_h, \mathcal{S}_w$ and $\{(h,w)\}$ as follows (defined below (11) and illustrated in Fig. 1):

Type 1 ($\mathcal{S}_h$): Let W_h be the largest w such that $(h,w)\in\mathcal{S}_h$. Let T_h be the length of $\mathcal{S}_h$ and let $D_h \triangleq \sum_{t\in\mathcal{T}_h}\min\{d_t,\tau_h-t+1\}\leq 2^{W_h}$, using Lemma 1. By applying the regret bound on $\mathcal{S}_h$:

$$R_{\mathcal{S}_h} \underset{(a)}{\leq} R\left(T_h, D_h, \mathcal{P}^*\left(2^h,2^{W_h}\right)\right) \underset{(b)}{\leq} R\left(2^h,2^{W_h},\mathcal{P}^*\left(2^h,2^{W_h}\right)\right) \leq k_1 2^{W_h d}+k_2 2^{hc}+k_3 2^{ha}2^{W_h b} \leq 3k_2 2^{hc} \tag{18}$$

where (a) follows since Algorithm 1 uses $\mathcal{P}^*\left(2^h,2^{W_h}\right)$, and (b) from condition 2 of the Theorem.

Type 2 ($\mathcal{S}_w$): Let H_w be the largest h such that $(h,w)\in\mathcal{S}_w$. Let T_w be the length of $\mathcal{S}_w$ and let $D_w \triangleq \sum_{t\in\mathcal{T}_w}\min\{d_t,\tau_w-t+1\}\leq 2^{w-1}$, using Lemma 1. By applying the regret bound on $\mathcal{S}_w$:

$$R_{\mathcal{S}_w} \leq R\left(T_w, D_w, \mathcal{P}^*\left(2^{H_w},2^w\right)\right) \leq R\left(2^{H_w},2^w,\mathcal{P}^*\left(2^{H_w},2^w\right)\right) \leq k_1 2^{wd}+k_2 2^{H_w c}+k_3 2^{H_w a}2^{wb} \leq 3k_1 2^{wd}. \tag{19}$$

Type 3 ($\{(h,w)\}$): For this case, we must have $2k_3 2^{ha}2^{wb}\geq k_1 2^{wd}+k_2 2^{ch}$, so

$$R_{h,w} \leq R\left(T_{h,w}, D_{h,w}, \mathcal{P}^*\left(2^h,2^w\right)\right) \leq R\left(2^h,2^w,\mathcal{P}^*\left(2^h,2^w\right)\right) \leq k_1 2^{wd}+k_2 2^{hc}+k_3 2^{ha}2^{wb} \leq 3k_3 2^{ha}2^{wb}. \tag{20}$$

Then the total regret is bounded by

$$\begin{aligned}
&\mathbb{E}\{R(T)\} = \sum_{h=1}^{H} R_{\mathcal{S}_h} + \sum_{w=1}^{W} R_{\mathcal{S}_w} + \sum_{(h,w)\notin\mathcal{S}_h\cup\mathcal{S}_w} R_{h,w} \underset{(a)}{\leq} \\
&3k_1\sum_{w=1}^{W}2^{dw}+3k_2\sum_{h=1}^{H}2^{ch}+3k_3\sum_{h=1}^{H}2^{ah}\sum_{w=1}^{W}2^{bw} \leq 6k_1\frac{2^{dW}-1}{2^d-1}+6k_2\frac{2^{cH}-1}{2^c-1}+12k_3\frac{2^{aH}-1}{2^a-1}\frac{2^{bW}-1}{2^b-1} \underset{(b)}{\leq} \\
&6k_1\frac{(2D)^d-1}{2^d-1}+6k_2\frac{(T+2)^c-1}{2^c-1}+12k_3\frac{(T+2)^a-1}{2^a-1}\frac{(2D)^b-1}{2^b-1} = O\left(k_1D^d+k_2T^c+k_3T^aD^b\right)
\end{aligned} \tag{21}$$

where (a) uses the bounds in (18),(19),(20) even for epochs that do not occur (that trivially have zero regret), or that end after the last round. Inequality (b) uses part 2 of Lemma 1 to upper bound H and W. We define the expressions before the last equality of (21) according to their continuous extensions at $a,b,c,d=0$, which yield logarithmic terms in T or D. Note that $\log D\leq 2\log T$. ■

5. The FKM Algorithm for Adversarial Bandit Convex Optimization with Delayed Feedback

In bandit convex optimization, the action $\boldsymbol{a}_t$ is chosen from a convex and compact set $\mathcal{K} \subset \mathbb{R}^n$ with diameter $|\mathcal{K}| \triangleq \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}} \|\boldsymbol{x}-\boldsymbol{y}\|$. Without the loss of generality, we assume that $\mathcal{K}$ contains the unit ball centered at the zero vector. The cost functions $l_t(\boldsymbol{a}_t) \in [0,1]$ are convex for all t and Lipschitz continuous with parameter L. The player has no access to the gradient of l_t, and only receives the value of $l_t(\boldsymbol{a}_t)$ at the beginning of round $t+d_t$. In the FKM algorithm, the player uses the cost value to estimate the gradient. The idea is to use a perturbation $\boldsymbol{u}_t$, drawn uniformly at random on the n-dimensional unit sphere $\mathbb{S}_1$. Then, instead of playing the unperturbed action $\boldsymbol{x}_t$, the player plays $\boldsymbol{a}_t = \boldsymbol{x}_t + \delta \boldsymbol{u}_t$ for a sampling radius $\delta > 0$. Let $\mathcal{K}_\delta = \left\{\boldsymbol{x} \,|\, \frac{1}{1-\delta}\boldsymbol{x} \in \mathcal{K}\right\}$. To ensure that $\boldsymbol{x}_t + \delta\boldsymbol{u}_t \in \mathcal{K}$, we maintain $\boldsymbol{x}_t \in \mathcal{K}_\delta$ by projecting into $\mathcal{K}_\delta$ after each gradient step. Since the cost functions are Lipschitz continuous, this projection creates a bias that decreases with δ.

Define the following filtration

$$\mathcal{F}_t = \sigma\left(\{\boldsymbol{x}_\tau \,|\, \tau \le t\}\right) \tag{22}$$

which is generated from all the past unperturbed actions. With a slight abuse of notation, we use $\mathcal{F}_{s_-}$ to denote the filtration induced from all $\boldsymbol{x}_\tau$ for $\tau \le t$ and all $\boldsymbol{x}_{q_-}$ for rounds $q \in \mathcal{S}_t$ that the algorithm used their feedback before using the feedback from round s, but including $\boldsymbol{x}_{s_-}$.

The purpose of the action perturbation is to allow for an estimator for the gradient at $\boldsymbol{a}_t$ with a bias that decreases with δ. This bias adds up to the bias that results from projecting into $\mathcal{K}_\delta$. On the other hand, the variance of the estimator increases with δ, introducing a bias-variance trade-off.

Lemma 2 (Flaxman et al. (2005, Lemma 2.1)). *Let $\delta > 0$ and define $\hat{l}(\boldsymbol{x}) \triangleq \mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{l(\boldsymbol{x}+\delta\boldsymbol{u})\}$ where $\mathbb{S}_1$ is the unit sphere. Let $\boldsymbol{g} = \frac{n}{\delta} l(\boldsymbol{x}+\delta\boldsymbol{u})\,\boldsymbol{u}$. Then $\mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{\boldsymbol{g}\} = \nabla\hat{l}(\boldsymbol{x})$.*

The next Lemma is the main result of this section, used to prove both Theorem 4 and Lemma 4.

Lemma 3 (Weighted-Regret Bound for FKM). *Let $\{\eta_t\}$ be a non-increasing step-size sequence. Let δ be the sampling radius. For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{K}} \sum_{t=1}^T \eta_t l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Define the set $\mathcal{M} = \{t \,|\, t+d_t > T,\, t \in [1,T]\}$ of all samples that are not received before round T. Then using FKM (Algorithm 2) guarantees:*

1. *For an oblivious adversary:*

$$\sum_{t=1}^T \eta_t \left(\mathbb{E}\left\{l_t(\boldsymbol{a}_t)\right\} - l_t(\boldsymbol{a}^*)\right) \le \sum_{t\in\mathcal{M}} \eta_t + \frac{|\mathcal{K}|^2}{2} + (3+|\mathcal{K}|)\,L\delta \sum_{t\notin\mathcal{M}} \eta_t + \frac{1}{2}\frac{n^2}{\delta^2}\sum_{t\notin\mathcal{M}} \eta_t^2 + 2L\frac{n}{\delta}\sum_{t\notin\mathcal{M}} \eta_t^2 d_t. \tag{23}$$

2. *For an adaptive adversary:*

$$\sum_{t=1}^T \eta_t \mathbb{E}\left\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right\} \le$$
$$\sum_{t\in\mathcal{M}} \eta_t + \frac{|\mathcal{K}|^2}{2} + (3+|\mathcal{K}|)\,L\delta \sum_{t\notin\mathcal{M}} \eta_t + |\mathcal{K}| \sqrt{2\sum_{t\notin\mathcal{M}} \eta_t^2 \left(\frac{n^2}{\delta^2} + L^2\right)} + \frac{1}{2}\frac{n^2}{\delta^2}\sum_{t\notin\mathcal{M}} \eta_t^2 (1+4d_t). \tag{24}$$

Proof See Appendix. ■

Algorithm 2 FKM with delays

Initialization: Let $\{\eta_t\}$ be a positive non-increasing sequence. Let $\delta < 1$. Set $\boldsymbol{x}_1 = 0$.
For $t = 1, ..., T$ **do**

1. Draw $\boldsymbol{u}_t \in \mathbb{S}_1$ uniformly at random, where $\mathbb{S}_1$ is the n-dimensional unit sphere.
2. Play $\boldsymbol{a}_t = \boldsymbol{x}_t + \delta \boldsymbol{u}_t$.
3. Obtain a set of delayed costs $l_s(\boldsymbol{a}_s)$ for all $s \in \mathcal{S}_t$ and compute $\boldsymbol{g}_s = \frac{n}{\delta} l_s(\boldsymbol{a}_s)\boldsymbol{u}_s$ for each.
4. Let $s_{\min} = \min_{s \in \mathcal{S}_t} s$ and $s_{\max} = \max_{s \in \mathcal{S}_t} s$. Set $\boldsymbol{x}_{s_{\min}^-} = \boldsymbol{x}_t$. For every $s \in \mathcal{S}_t$, update

$$\boldsymbol{x}_{s_+} = \prod_{\mathcal{K}_\delta} \left(\boldsymbol{x}_{s_-} - \eta_s \boldsymbol{g}_s\right) \tag{25}$$

where $\mathcal{K}_\delta = \left\{\boldsymbol{x} \,|\, \frac{1}{1-\delta}\boldsymbol{x} \in \mathcal{K}\right\}$, and then set $\boldsymbol{x}_{t+1} = \boldsymbol{x}_{s_{\max}^+}$.

End

The following theorem establishes the expected regret bound for FKM with delays. It is proved by optimizing over a constant step-size η and sampling radius δ in Lemma 3.

Theorem 4. *Let $\eta > 0$ and $0 < \delta < 1$. For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathcal{K}} \sum_{t=1}^{T} l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Define the set $\mathcal{M} = \{t \,|\, t + d_t > T,\, t \in [1,T]\}$ of all samples that are not received before round T. Then the expected regret of FKM (Algorithm 2) against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \sum_{t=1}^{T} \mathbb{E}\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\} \leq$$
$$|\mathcal{M}| + \left((3 + |\mathcal{K}|)\delta L + \frac{1}{2}\eta\frac{n^2}{\delta^2}\right)(T - |\mathcal{M}|) + \frac{|\mathcal{K}|^2}{2\eta} + 2Ln\frac{\eta}{\delta}\sum_{t \notin \mathcal{M}} d_t. \tag{26}$$

Furthermore, for

$$\eta = |\mathcal{K}| \min\left\{\frac{1}{n}T^{-\frac{3}{4}}, \frac{1}{\sqrt{n}}T^{-\frac{1}{3}}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{-\frac{1}{3}}\right\} \text{ and } \delta = \max\left\{T^{-\frac{1}{4}}, T^{-\frac{2}{3}}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{\frac{1}{3}}\right\} \tag{27}$$

we obtain

$$\mathbb{E}\{R(T)\} = O\left(nT^{\frac{3}{4}} + \sqrt{n}\left(\sum_{t \notin \mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}} + |\mathcal{M}|\right). \tag{28}$$

Proof First note that if $|\mathcal{M}| = \Theta(T)$ then $\mathbb{E}\{R(T)\} = O(T)$ and otherwise $T - |\mathcal{M}| = \Theta(T)$. To obtain (26), substitute $\eta_t = \eta$ in Lemma 3 for an oblivious adversary and divide both sides by η. We have $\Theta(\delta T) = \Theta\left(\frac{\eta}{\delta^2}T\right) = \Theta\left(\frac{1}{\eta}\right)$ for $\eta = \frac{|\mathcal{K}|}{n}T^{-\frac{3}{4}}$ and $\delta = T^{-\frac{1}{4}}$, hence this choice of parameters minimizes the order of magnitude of the T dependence of the following expression:

$$\min_{\delta,\eta}\left((3 + |\mathcal{K}|)\delta L(T - |\mathcal{M}|) + \frac{1}{2}\eta\frac{n^2}{\delta^2}(T - |\mathcal{M}|) + \frac{|\mathcal{K}|^2}{2\eta}\right) = O\left(nT^{\frac{3}{4}}\right). \tag{29}$$

Since $\Theta(\delta T) = \Theta\left(\frac{\eta}{\delta}\sum_{t\notin\mathcal{M}} d_t\right) = \Theta\left(\frac{1}{\eta}\right)$ for $\eta = \frac{1}{\sqrt{n}}\left(\frac{1}{T\sum_{t\notin\mathcal{M}} d_t}\right)^{\frac{1}{3}}$ and $\delta = T^{-\frac{2}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}}$, then this choice of parameters minimizes the order of magnitude of the T dependence of the following expression:

$$\min_{\delta,\eta}\left((3+|\mathcal{K}|)\,\delta L\,(T-|\mathcal{M}|) + \frac{|\mathcal{K}|^2}{2\eta} + 2L\eta\frac{n}{\delta}\sum_{t\notin\mathcal{M}} d_t\right) = O\left(\sqrt{n}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}}\right). \tag{30}$$

Therefore (26) cannot have a better T dependence than in (28) for any η, δ. For the choice in (27) we have

$$n^2\frac{\eta}{\delta^2}T = \frac{|\mathcal{K}|\min\left\{nT^{\frac{1}{4}}, n^{\frac{3}{2}}T^{\frac{2}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{-\frac{1}{3}}\right\}}{\max\left\{T^{-\frac{1}{2}}, T^{-\frac{4}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{2}{3}}\right\}} \underset{(a)}{\leq} O\left(nT^{\frac{3}{4}}\right) \tag{31}$$

where (a) follows since $\frac{\min\{a,b\}}{\max\{c,d\}} \leq \frac{a}{c}$. For the choice in (27) we also have

$$\frac{\eta}{\delta}n\sum_{t\notin\mathcal{M}} d_t = \frac{|\mathcal{K}|\min\left\{T^{-\frac{3}{4}}\sum_{t\notin\mathcal{M}} d_t, \sqrt{n}T^{-\frac{1}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{2}{3}}\right\}}{\max\left\{T^{-\frac{1}{4}}, T^{-\frac{2}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}}\right\}} \underset{(a)}{\leq} O\left(\sqrt{n}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}}\right) \tag{32}$$

where (a) follows since $\frac{\min\{a,b\}}{\max\{c,d\}} \leq \frac{b}{d}$. Therefore, for the choice in (27)

$$\begin{aligned}\sum_{t=1}^{T}\left(\mathbb{E}\left\{l_t\left(\boldsymbol{a}_t\right)\right\} - l_t\left(\boldsymbol{a}^*\right)\right) &\leq |\mathcal{M}| + \left((3+|\mathcal{K}|)L + \frac{|\mathcal{K}|}{2}\right)\max\left\{nT^{\frac{3}{4}}, \sqrt{n}T^{\frac{1}{3}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}}\right\} \\ &\quad + O\left(nT^{\frac{3}{4}}\right) + O\left(\sqrt{n}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}}\right) = O\left(nT^{\frac{3}{4}} + \sqrt{n}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} T^{\frac{1}{3}} + |\mathcal{M}|\right). \end{aligned} \tag{33}$$

■

The next Corollary shows that the bound of Theorem 4 is slightly tighter than $O\left(nT^{\frac{3}{4}} + \sqrt{n}D^{\frac{1}{3}}T^{\frac{1}{3}}\right)$, where one sums $D = \sum_{t=1}^{T}\min\{d_t, T-t+1\}$ instead of $\sum_{t\notin\mathcal{M}} d_t$ in our regret bound (28), depending on the pattern of the missing samples.

Corollary 1. *Choose the fixed η and δ according to (27). For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{K}}\sum_{t=1}^{T} l_t(\boldsymbol{a})$. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Let $D = \sum_{t=1}^{T}\min\{d_t, T-t+1\}$. Then the expected regret of FKM (Algorithm 2) against an oblivious adversary satisfies*

$$\mathbb{E}\left\{R(T)\right\} = \sum_{t=1}^{T}\left(\mathbb{E}\left\{l_t\left(\boldsymbol{a}_t\right)\right\} - l_t\left(\boldsymbol{a}^*\right)\right) = O\left(nT^{\frac{3}{4}} + \sqrt{n}D^{\frac{1}{3}}T^{\frac{1}{3}}\right). \tag{34}$$

Proof The $m \triangleq |\mathcal{M}|$ missing samples contribute at least $\frac{m(m+1)}{2} \geq \frac{m^2}{2}$ to $D = \sum_{t=1}^{T}\min\{d_t, T-t+1\}$. This follows since the best case is when the feedback of round T is delayed by one and arrives after T, the feedback of round $T-1$ now has to be delayed by at least 2 to arrive after T and so on, m times. Therefore

$$T^{\frac{1}{3}}D^{\frac{1}{3}} \geq \left(T\sum_{t\notin\mathcal{M}} d_t + T\frac{m^2}{2}\right)^{\frac{1}{3}} \underset{(a)}{\geq} \frac{1}{2}\left(2T\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} + \frac{1}{2}\left(Tm^2\right)^{\frac{1}{3}} \geq \frac{T^{\frac{1}{3}}}{2^{\frac{2}{3}}}\left(\sum_{t\notin\mathcal{M}} d_t\right)^{\frac{1}{3}} + \frac{m}{2} \tag{35}$$

where (a) follows from the concavity of $f(x) = x^{\frac{1}{3}}$. We conclude that the regret in (34) is greater than that in (28).

■

5.1 Agnostic FKM

If the horizon T and sum of delays D are unknown, then we can apply Algorithm 1 to wrap FKM. The next Theorem is an immediate application of Theorem 3 on the FKM regret bound for this agnostic case. The resulting bound retains the same order of magnitude as the bound of Corollary 1 even though D and T are unknown. The only difference with the bound of Theorem 4 arises because the doubling trick discards samples that cross super-epochs. Hence, the bound below uses $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$ instead of $\sum_{t \notin \mathcal{M}} d_t$ and $|\mathcal{M}|$.

Theorem 5. *For every t, let $l_t : \mathcal{K} \to [0,1]$ be a convex cost function that is Lipschitz continuous with parameter L. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Let $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$. If the player uses Algorithm 1 to wrap FKM (Algorithm 2) such that in the (h,w) epoch $\eta_{h,w} = |\mathcal{K}| \min\left\{\frac{1}{n}2^{-\frac{3h}{4}}, \frac{1}{\sqrt{n}}2^{-\frac{h+w}{3}}\right\}$ and $\delta_{h,w} = \delta_0 \max\left\{2^{-\frac{h}{4}}, 2^{\frac{w-2h}{3}}\right\}$ for $\delta_0 < 1$, then the expected regret of FKM (Algorithm 2) against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = O\left(nT^{\frac{3}{4}} + \sqrt{n}D^{\frac{1}{3}}T^{\frac{1}{3}}\right). \tag{36}$$

Proof Consider the regret bound in (26) after bounding $|\mathcal{M}| \leq \sqrt{2D}$ (as in Corollary 1), $T - |\mathcal{M}| \leq T$ and $\sum_{t \notin \mathcal{M}} d_t \leq D$. For the choice in (27), this bound takes the form $O(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}} + D^{\frac{1}{2}})$ (where $T^{\frac{1}{3}}D^{\frac{1}{3}} \geq D^{\frac{1}{2}}$) so it matches Assumption 1 in Theorem 3 with $a = b = \frac{1}{3}, c = \frac{3}{4}$ and $d = \frac{1}{2}$. This bound also satisfies Assumption 2 since it is increasing with T and D for any η, δ. ■

5.2 No Weighted-Regret Property for FKM

In this subsection, we provide conditions for FKM to have no weighted-regret with respect to the delay sequence and its step-size sequence as the weight sequence of Section 3. As discussed in Section 3, $\sum_{t=1}^{\infty} \eta_t = \infty$ is necessary for the no weighted-regret property to be non-trivial. All other conditions of Lemma 4 are as tight as the bound of Lemma 3.

Lemma 4. *FKM with a non-increasing and positive step-size sequence $\{\eta_t\}$ and sampling radius δ_T has no weighted-regret with respect to the sequence of delays $\{d_t\}$ and $\{\eta_t\}$ as the weight sequence, if the following three conditions hold:*

1. $\sum_{t=1}^{\infty} \eta_t = \infty$.
2. $\lim_{t\to\infty} \eta_t d_t < \infty$ *and* $\sum_{t=1}^{\infty} \eta_t^2 d_t < \infty$.
3. $\lim_{T\to\infty} \delta_T = 0$ *and* $\lim_{T\to\infty} \delta_T^2 \sum_{t=1}^{T} \eta_t = \infty$.

Proof Let $\mathcal{M} = \{t \,|\, t + d_t > T, t \in [1,T]\}$. Define $t^*(T) = \min_{t \in \mathcal{M}} t$, and note that $t^*(T) \to \infty$ as $T \to \infty$ since $t + d_t \geq t$, and $f(t) = t$ is increasing. Since η_t is non-increasing then

$$\sum_{t \in \mathcal{M}} \eta_t \leq |\mathcal{M}| \, \eta_{t^*(T)} \leq (T - t^*(T) + 1) \, \eta_{t^*(T)} \leq d_{t^*(T)} \eta_{t^*(T)}. \tag{37}$$

Let $A = (3 + |\mathcal{K}|) L$. Therefore

$$\lim_{T\to\infty} \mathbb{E}\left\{\frac{\sum_{t=1}^{T} \eta_t \left(l_t\left(\boldsymbol{a}_t\right) - l_t\left(\boldsymbol{a}^*\right)\right)}{\sum_{t=1}^{T} \eta_t}\right\}$$
$$\underset{(a)}{\leq} \lim_{T\to\infty} \frac{d_{t^*(T)}\eta_{t^*(T)} + \frac{|\mathcal{K}|^2}{2} + A\delta_T \sum_{t=1}^{T} \eta_t + \frac{2n|\mathcal{K}|}{\delta_T}\sqrt{\sum_{t=1}^{T} \eta_t^2} + \frac{1}{2}\frac{n^2}{\delta_T^2}\sum_{t=1}^{T} \eta_t^2 \left(1 + 4d_t\right)}{\sum_{t=1}^{T} \eta_t} \underset{(b)}{=} 0 \quad (38)$$

where (a) is Lemma 3 and (37), and (b) follows since $\lim_{t\to\infty} d_t\eta_t < \infty$, $\sum_{t=1}^{\infty} \eta_t^2 d_t < \infty$, $\sum_{t=1}^{\infty} \eta_t = \infty$, $\lim_{T\to\infty} \delta_T = 0$ and $\lim_{T\to\infty} \delta_T^2 \sum_{t=1}^{T} \eta_t = \infty$. ■

Note that one can choose $\eta_t = \frac{1}{t \log t \log\log t \log\log\log t}$ and $\delta_T = O\left(\left(\log\log\log\left(\log T\right)\right)^{-\frac{1}{3}}\right)$ to guarantee no weighted-regret for all sequences such that $d_t = O\left(t \log t \log\log t\right)$. This boundary can only be slightly improved by adding $\log\left(\log\left(\ldots \log\left(T\right)\right)\right)$ iteratively in this manner as long as $\sum_{t=1}^{T} \frac{1}{d_t} = \infty$. Outside this boundary, we cannot guarantee that FKM has no weighted-regret even if d_t is known. Hence, no knowledge of the individual terms of the sequence d_t is required to tune η_t and δ_T such that FKM has no weighted-regret for all sequences inside this boundary, which can be arbitrarily extended to include all the sequences for which Lemma 4 holds. However, if a tighter bound on the rate of growth of d_t is available then one can improve the convergence rate to the set of CCE by picking a more slowly decaying η_t than $\eta_t = \frac{1}{t \log t \log\log t \log\log\log t}$. This still would not require knowledge of the individual terms in d_t. In general, for a given T, FKM gives an ε-CCE with $\varepsilon = O\left(\max\left\{\frac{1}{\delta_T^2 \sum_{t=1}^{T} \eta_t}, \delta_T\right\}\right)$, as given in (38).

The following Proposition shows that FKM can have no weighted-regret even when it has linear regret in T. As a result, FKM can be used to approximate a CCE in a non-cooperative game with convex cost functions or a NE in a two-player convex-concave zero-sum game despite not having no-regret guarantees.

Proposition 2. *There exist a delay sequence $\{d_t\}$ and Lipschitz continuous convex functions $\{l_1, ..., l_T\}$ on $[0, 1]$ such that for a large enough T*

$$\mathbb{E}\left\{R\left(T\right)\right\} = \sum_{t=1}^{T} \mathbb{E}\left\{l_t\left(\boldsymbol{a}_t\right) - l_t\left(\boldsymbol{a}^*\right)\right\} \geq \frac{T}{4\left(|\mathcal{K}| + 1\right)^2} \quad (39)$$

but still the step-sizes $\{\eta_t\}$ and sampling radius δ_T for Algorithm 2 (FKM) can be chosen such that it has no weighted-regret with respect to $\{d_t\}$ and $\{\eta_t\}$ as the weight sequence.

Proof Let $d_t = t$ and choose $\delta_T = 0.1\left(\log\log(T+1)\right)^{-\frac{1}{3}}, \eta_t = \frac{1}{t\log(t+1)}$ for all $t \geq 1$, for which $\sum_{t=1}^{T} \eta_t = O\left(\log\log T\right), \sum_{t=1}^{\infty} d_t\eta_t^2 < \infty$ and also $\delta_T \to 0$ and $\delta_T^2 \sum_{t=1}^{T} \eta_t \to \infty$ as $T \to \infty$. Hence, by Lemma 4 we obtain that FKM has no weighted-regret with respect to d_t and η_t. However, the feedback for the last $\frac{T}{2}$ rounds is never received. Therefore, the unperturbed action $\boldsymbol{x}_t$ is $\boldsymbol{x}_{\frac{T}{2}}$ for all $t \geq \frac{T}{2}$. Consider the sequence of costs $l_t\left(\boldsymbol{a}\right) = \boldsymbol{0}$ for all $t \leq \frac{T}{2}$ and $l_t\left(\boldsymbol{a}\right) = \frac{\left\|\boldsymbol{a} - \frac{1}{\sqrt{n}}\mathbf{1}\right\|^2}{\left(|\mathcal{K}|+1\right)^2}$ for all $t > \frac{T}{2}$ where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones. Starting from $\boldsymbol{x}_1 = \boldsymbol{0}$ and computing $\boldsymbol{g}_t = \boldsymbol{0}$ for all $t \leq \frac{T}{2}$ we have $\boldsymbol{x}_{\frac{T}{2}} = \boldsymbol{0}$. Then, from the Lipschitz continuity of l_t we obtain for all $t > \frac{T}{2}$, for large enough T,

$$\mathbb{E}\left\{l_t\left(\boldsymbol{a}_t\right)\right\} = \mathbb{E}\left\{l_t\left(\boldsymbol{x}_{\frac{T}{2}} + \delta_T \boldsymbol{u}_t\right)\right\} \geq \mathbb{E}\left\{l_t\left(\boldsymbol{0}\right)\right\} - \delta_T L = \frac{1}{\left(|\mathcal{K}| + 1\right)^2} - \delta_T L \geq \frac{1}{2\left(|\mathcal{K}| + 1\right)^2} \quad (40)$$

which means that this sequence yields an expected regret of at least $\frac{T}{4(|\mathcal{K}|+1)^2}$. ■

6. The EXP3 Algorithm for Adversarial Multi-Armed Bandits with Delayed Feedback

Consider a player that at each round t picks one out of K arms. Let a_t be the arm the player chooses at round t. The cost at round t of playing arm i is $l_t^{(i)} \in [0,1]$, and let $\boldsymbol{l}_t = \left(l_t^{(1)}, ..., l_t^{(K)}\right)$ be the cost vector. At round t, the EXP3 algorithm, detailed in Algorithm 3, chooses an arm at random using a distribution that depends on the history of the game. The variant when $\gamma_t \neq 0$, as we use against an adaptive adversary, is known as EXP3-IX (see Neu (2015)). We denote the vector of probabilities of the player for choosing arms at round t by $\boldsymbol{p}_t \in \Delta^K$, where Δ^K denotes the K-simplex. This is also known as the mixed action of the player. We also define the following filtration

$$\mathcal{F}_t = \sigma\left(\{a_s \,|\, s + d_s \leq t\} \cup \{\boldsymbol{l}_s \,|\, s \leq t\}\right) \tag{41}$$

which is generated from all the actions for which the feedback was received up to round t and all cost functions up to round t. Note that the mixed action $\boldsymbol{p}_t$ is a $\mathcal{F}_t$-measurable random variable since $\mathcal{F}_t$ includes everything that could affect the algorithm up to round t. With a slight abuse of notation, we use $\mathcal{F}_{s_-}$ to denote the filtration induced from all actions for which the feedback has been received and used up to step s_- and all the cost functions up to round t.

The next Lemma is the main result of this section, used to prove both Theorem 6 and Lemma 6.

Lemma 5 (Weighted-Regret Bound for EXP3)**.** *Let $\{\eta_t\}$ be a non-increasing step-size sequence such that $\eta_t \leq \frac{1}{2}e^{-2}$ for all t. Let $\left\{l_t^{(i)}\right\}$ be a cost sequence such that $l_t^{(i)} \in [0,1]$ for every t, i. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Define the set $\mathcal{M}^*$ of all samples that are not received before round T or that were delayed by $d_t \geq \frac{1}{e^2\eta_t} - 1$. Then using EXP3 (Algorithm 3) guarantees:*

1. *With an oblivious adversary and $\gamma_t = 0$ for all t:*

$$\mathbb{E}\left\{\sum_{t=1}^{T} \eta_t l_t^{(a_t)} - \min_i \sum_{t=1}^{T} \eta_t l_t^{(i)}\right\} \leq \log K + \frac{e^2}{2} K \sum_{t=1}^{T} \eta_t^2 + 4 \sum_{t \notin \mathcal{M}^*} \eta_t^2 d_t + \sum_{t \in \mathcal{M}^*} \eta_t. \tag{42}$$

2. *With an adaptive adversary and $\gamma_t = \eta_t$ for all t:*

$$\mathbb{E}\left\{\sum_{t=1}^{T} \eta_t l_t^{(a_t)} - \min_i \sum_{t=1}^{T} \eta_t l_t^{(i)}\right\} \leq 2 + 2\log K + (1 + \frac{e^2}{2}) K \sum_{t=1}^{T} \eta_t^2 + 4e^2 K \sum_{t \notin \mathcal{M}^*} \eta_t^2 d_t + \sum_{t \in \mathcal{M}^*} \eta_t. \tag{43}$$

Proof See Appendix. ■

The following theorem establishes the expected regret bound for EXP3 with delays. It is proved by optimizing over a constant step-size η in Lemma 5.

Theorem 6. *Let $\left\{l_t^{(i)}\right\}$ be a cost sequence such that $l_t^{(i)} \in [0,1]$ for every t, i. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t + d_t$. Define the set $\mathcal{M} = \{t \,|\, t + d_t > T,\, t \in [1,T]\}$ of all samples that are not received before round T. Let us choose the fixed step-size $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\log K}{KT + \sum_{t \notin \mathcal{M}} d_t}}$. Then the expected regret of EXP3 (Algorithm 3) against an oblivious adversary satisfies*

$$\mathbb{E}\left\{R\left(T\right)\right\} = \mathbb{E}\left\{\sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{p}_t \rangle - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} = O\left(\sqrt{\log K \left(KT + \sum_{t \notin \mathcal{M}} d_t\right)} + |\mathcal{M}|\right). \tag{46}$$

Algorithm 3 EXP3 with delays

Initialization: Let $\{\eta_t\}$ and $\{\gamma_t\}$ be non-negative non-increasing sequences such that $\eta_1 \leq \frac{e^{-2}}{2}$ and set $\tilde{L}_1^{(i)} = 0$ and $p_1^{(i)} = \frac{1}{K}$ for $i = 1, ..., K$.
For $t = 1, ..., T$ **do**

1. Choose an arm a_t at random according to the distribution $\boldsymbol{p}_t$.
2. Collect in $\mathcal{S}_t$ all the rounds s for which $l_s^{(a_s)}$ arrived at round t after a delay of $d_s \leq \frac{1}{e^2 \eta_s} - 1$.
3. Set $\tilde{L}_t^{(i)} = \tilde{L}_{t-1}^{(i)}$ for all i. Update the weights of arm a_s for all $s \in \mathcal{S}_t$ using
$$\tilde{L}_t^{(a_s)} = \tilde{L}_t^{(a_s)} + \eta_s \frac{l_s^{(a_s)}}{p_s^{(a_s)} + \gamma_s}. \tag{44}$$
4. Update the mixed action for $i = 1, \ldots, K$ using
$$p_{t+1}^{(i)} = \frac{e^{-\tilde{L}_t^{(i)}}}{\sum_{j=1}^{K} e^{-\tilde{L}_t^{(j)}}}. \tag{45}$$

End

Proof We choose $\eta_t = \eta$ in (42) of Lemma 5 and define $\mathcal{D} = \left\{ t \,|\, d_t \geq \frac{1}{e^2 \eta} - 1 \text{ and } t + d_t \leq T \right\}$ so $\mathcal{M}^* = \mathcal{M} \cup \mathcal{D}$ in the statement of Lemma 5. Now divide both sides by η:

$$\mathbb{E}\left\{ \sum_{t=1}^{T} l_t^{(a_t)} - \min_i \sum_{t=1}^{T} l_t^{(i)} \right\} \leq \frac{\log K}{\eta} + \frac{e^2}{2} \eta K T + 4\eta \sum_{t \notin \mathcal{M}^*} d_t + |\mathcal{M}| + |\mathcal{D}|. \tag{47}$$

Then, choosing $\eta = \frac{e^{-2}}{2} \sqrt{\frac{\log K}{KT + \sum_{t \notin \mathcal{M}} d_t}}$ yields (46). Note that for this choice $|\mathcal{D}| \leq \frac{\sum_{t \notin \mathcal{M}} d_t}{e^{-2} \eta^{-1} - 1} \leq \sqrt{\left(\sum_{t \notin \mathcal{M}} d_t\right) \log K}$, since $\sum_{t \notin \mathcal{M}} d_t \geq \left(\frac{1}{e^2 \eta} - 1 \right) |\mathcal{D}|$ (discarded samples in $\mathcal{D}$ are not missing). ■

Similar to the bandit convex optimization case, the bound of Theorem 6 is tighter than $O\left(\sqrt{\log K \left(KT + D\right)}\right)$ for $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$, as the next Corollary shows.

Corollary 2. *Let* $\eta = \frac{e^{-2}}{2} \sqrt{\frac{\log K}{KT + \sum_{t \notin \mathcal{M}} d_t}}$. *Let* $\left\{ l_t^{(i)} \right\}$ *be a cost sequence such that* $l_t^{(i)} \in [0, 1]$ *for every* t, i. *Let* $\{d_t\}$ *be a delay sequence such that the cost from round* t *is received at round* $t + d_t$. *Let* $D = \sum_{t=1}^{T} \min\{d_t, T - t + 1\}$. *Then the expected regret of EXP3 (Algorithm 3) against an oblivious adversary satisfies*

$$\mathbb{E}\{R(T)\} = \mathbb{E}\left\{ \sum_{t=1}^{T} \langle \boldsymbol{l}_t, \boldsymbol{p}_t \rangle - \min_i \sum_{t=1}^{T} l_t^{(i)} \right\} = O\left(\sqrt{\log K \left(KT + D\right)}\right). \tag{48}$$

Proof The $m = |\mathcal{M}|$ missing samples contribute at least $\frac{m(m+1)}{2}$ to D (as in Corollary 1), so

$$\sqrt{\log K\left(KT+D\right)} \geq \sqrt{\log K\left(KT+\sum_{t\notin\mathcal{M}} d_t + \frac{m\left(m+1\right)}{2}\right)}$$
$$\underset{(a)}{\geq} \frac{1}{2}\sqrt{2\log K\left(KT+\sum_{t\notin\mathcal{M}} d_t\right)} + \frac{1}{2}\sqrt{\log Km\left(m+1\right)} \geq O\left(\sqrt{\log K\left(KT+\sum_{t\notin\mathcal{M}} d_t\right)}\right) + \frac{|\mathcal{M}|}{4} \tag{49}$$

where (a) follows from the concavity of $f\left(x\right) = \sqrt{x}$. ■

6.1 Agnostic EXP3

The step-size $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\log K}{KT+\sum_{t\notin\mathcal{M}} d_t}}$ used in Algorithm 3 requires knowing the horizon T and the sum of delays D. When these parameters are unknown, we can apply Algorithm 1 to wrap EXP3. The next Theorem is an immediate application Theorem 3 on the EXP3 regret bound for this agnostic case. The resulting bound retains the same order of magnitude as the bound of Corollary 2 even though D and T are unknown. The only difference with the bound of Theorem 6 arises because the doubling trick discards samples that cross super-epochs. Hence, the bound below uses $D = \sum_{t=1}^{T} \min\left\{d_t, T-t+1\right\}$ instead of $\sum_{t\notin\mathcal{M}} d_t$ and $|\mathcal{M}|$.

Theorem 7. *Let $\left\{l_t^{(i)}\right\}$ be a cost sequence such that $l_t^{(i)} \in [0,1]$ for every t, i. Let $\{d_t\}$ be a delay sequence such that the cost from round t is received at round $t+d_t$. Let $D = \sum_{t=1}^{T} \min\left\{d_t, T-t+1\right\}$. If the player uses Algorithm 1 to wrap EXP3 (Algorithm 3) with step size $\eta_{h,w} = \frac{e^{-2}}{2}\sqrt{\frac{\log K}{K2^h+2^w}}$ for epoch (h,w), then the regret against an oblivious adversary satisfies*

$$\mathbb{E}\left\{R\left(T\right)\right\} = \mathbb{E}\left\{\sum_{t=1}^{T}\left\langle \boldsymbol{l}_t, \boldsymbol{p}_t\right\rangle - \min_i \sum_{t=1}^{T} l_t^{(i)}\right\} = O\left(\sqrt{\log K\left(KT+D\right)}\right). \tag{50}$$

Proof Consider the regret bound in (47) after bounding $|\mathcal{M}| \leq \sqrt{2D}$, (as in the proof of Corollary 2), $T - |\mathcal{M}| \leq T$, $\sum_{t\notin\mathcal{M}^*} d_t \leq D$ and $|\mathcal{D}| \leq \sqrt{D\log K}$ (as in the proof of Theorem 6). For $\eta = \frac{e^{-2}}{2}\sqrt{\frac{\log K}{KT+D}}$, this bound yields that of Corollary 2 which is of the form $\sqrt{TK\log K} + \sqrt{D\log K}$, so it matches Assumption 1 in Theorem 3 with $a = b = 0$ and $c = d = \frac{1}{2}$. This bound also satisfies Assumption 2 since it is increasing with T and D for any η. ■

6.2 No Weighted-Regret Property for EXP3

In this subsection, we provide conditions for EXP3 to have no weighted-regret with respect to the delay sequence and its step-size sequence as the weight sequence of Section 3. As discussed in Section 3, $\sum_{t=1}^{\infty} \eta_t = \infty$, is necessary for the no weighted-regret property to be non-trivial. All other conditions of Lemma 6 are as tight as the bound of Lemma 5.

Lemma 6. *EXP3 with a non-increasing and positive step-size sequence $\{\eta_t\}$ such that $\eta_t \leq \frac{1}{2}e^{-2}$ for all t has no weighted-regret with respect to the sequence of delays $\{d_t\}$ and $\{\eta_t\}$ as the weight sequence, if the following two conditions hold:*

1. $\sum_{t=1}^{\infty}\eta_t=\infty$.
2. $\lim_{t\to\infty}\eta_t d_t<\infty$ *and* $\sum_{t=1}^{\infty}\eta_t^2 d_t<\infty$.

Proof Define the set of missing samples $\mathcal{M}=\{t\,|\,t+d_t>T\}$ and the set of discarded samples $\mathcal{D}_T=\left\{t\,|d_t\eta_t>e^{-2}-\eta_t\right\}$. Define $t^*(T)=\min_{t\in\mathcal{M}} t$, and note that $t^*(T)\to\infty$ as $T\to\infty$ since $t+d_t\geq t$, and $f(t)=t$ is increasing. Since η_t is non-increasing then

$$\sum_{t\in\mathcal{M}}\eta_t\leq|\mathcal{M}|\,\eta_{t^*(T)}\leq(T-t^*(T)+1)\,\eta_{t^*(T)}\leq d_{t^*(T)}\eta_{t^*(T)}. \tag{51}$$

Given $\sum_{t=1}^{\infty}\eta_t^2 d_t<\infty$ and $\sum_{t=1}^{\infty}\eta_t=\infty$ we must have $\lim_{t\to\infty}\eta_t d_t=0$ if this limit exists, so $\lim_{T\to\infty}|\mathcal{D}_T|<\infty$. Therefore for the optimal arm i^*

$$\lim_{T\to\infty}\frac{\mathbb{E}\left\{\sum_{t=1}^{T}\eta_t\left(l_t^{(a_t)}-l_t^{(i^*)}\right)\right\}}{\sum_{t=1}^{T}\eta_t}\underset{(a)}{\leq}\lim_{T\to\infty}\frac{\eta_1|\mathcal{D}_T|+d_{t^*(T)}\eta_{t^*(T)}+4\log K+5K\sum_{t=1}^{T}\eta_t^2(1+6d_t)}{\sum_{t=1}^{T}\eta_t}\underset{(b)}{=}0 \tag{52}$$

where (a) is Lemma 5 and (51), and (b) uses $d_t\eta_t\to 0$ as $t\to\infty$, $\sum_{t=1}^{\infty}\eta_t=\infty$ and $\sum_{t=1}^{\infty}d_t\eta_t^2<\infty$. ■

Note that one can choose $\eta_t=\frac{1}{t\log t\log\log t\log\log\log t}$ to guarantee no weighted-regret for all sequences such that $d_t=(t\log t\log\log t)$. This boundary can only be slightly improved by adding $\log(\log(\ldots\log(T)))$ iteratively in this manner as long as $\sum_{t=1}^{\infty}\frac{1}{d_t}=\infty$. Outside this boundary, we cannot guarantee that EXP3 has no weighted-regret even if d_t is known. Hence, no knowledge of the individual terms of the sequence d_t is required to tune η_t such that EXP3 has no weighted-regret for all sequences inside this boundary, which can be arbitrarily extended to include all the sequences for which Lemma 6 holds. However, if a tighter bound on the rate of growth of d_t is available then one can improve the convergence rate to the set of CCE by picking a more slowly decaying η_t than $\eta_t=\frac{1}{t\log t\log\log t\log\log\log t}$. This still would not require knowledge of the individual terms in d_t. In general, for a given T, EXP3 gives an ε-CCE with $\varepsilon=O\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right)$, as given in (52).

The following Proposition shows that EXP3 can have no weighted-regret even when it has linear regret in T. As a result, EXP3 can be used to approximate a CCE in a discrete non-cooperative game or a NE in a discrete two-player zero-sum game despite not having no-regret guarantees.

Proposition 3. *There exist a delay sequence* $\{d_t\}$ *and a cost sequence* $\left\{l_t^{(1)},\ldots,l_t^{(K)}\right\}_t$ *with* $0\leq l_t^{(i)}\leq 1$ *for all* t *and* i*, such that*

$$\mathbb{E}\{R(T)\}=\mathbb{E}\left\{\sum_{t=1}^{T}\langle\boldsymbol{l}_t,\boldsymbol{p}_t\rangle-\min_i\sum_{t=1}^{T}l_t^{(i)}\right\}\geq\left(1-\frac{1}{K}\right)\frac{T}{2} \tag{53}$$

but still the step-sizes $\{\eta_t\}$ *for Algorithm 3 (EXP3) can be chosen such that it has no weighted-regret with respect to* $\{d_t\}$ *and* $\{\eta_t\}$ *as the weight sequence.*

Proof Let $d_t=t$ and $\eta_t=\frac{1}{t\log(t+1)}$ for all t, for which $d_t\eta_t^2=\frac{1}{t\log^2(t+1)}$ so $\sum_{t=1}^{\infty}\eta_t=\infty$, $\sum_{t=1}^{\infty}d_t\eta_t^2<\infty$ and $\lim_{t\to\infty}\eta_t d_t=0$. Hence, by Lemma 6 we obtain that EXP3 has no weighted-regret with respect to d_t and η_t. However, the feedback for the last $\frac{T}{2}$ rounds is never received. Therefore, the mixed action $\boldsymbol{p}_t$ does not change for all $t\geq\frac{T}{2}$. Then the cost sequence such that $l_t^{(i)}=0$ for all i and all $t\leq\frac{T}{2}$ and $l_t^{(1)}=0$, $l_t^{(j)}=1$ for all $j>1$ and all $t>\frac{T}{2}$ yields an expected regret of exactly $\left(1-\frac{1}{K}\right)\frac{T}{2}$. ■

7. Conclusions

We studied the weighted-regret of online learning with adversarial bandit feedback and an arbitrary delay sequence $\{d_t\}$. Our results have implications both for the single-agent and multi-agent cases.

For the single-agent case, our weighted-regret bounds yield standard regret bounds as a special case. We showed an expected regret bound of $O\left(nT^{\frac{3}{4}} + \sqrt{n}T^{\frac{1}{3}}D^{\frac{1}{3}}\right)$ for FKM and $O\left(\sqrt{\log K\,(KT+D)}\right)$ for EXP3, where $D = \sum_{t=1}^{T} \min\{d_t, T-t+1\}$. These bounds hold even if D, T are unknown thanks to a novel doubling trick. Our doubling trick can be applied to any online learning algorithm with delays in a plug-and-play manner. Under mild conditions, the novel doubling trick provably retains the order of magnitude dependence on D, T (and other parameters) of the regret bound for when D, T are known.

Our single-agent results in this paper focus on FKM and EXP3 since they are the most widely used algorithms for bandit convex optimization and multi-armed bandits, respectively. Therefore it is crucial to understand how they perform under delays, which are prevalent in practical systems. However, this leaves open the question of what are the best algorithms for delayed bandit feedback.

For multi-armed bandits the lower bound is $O\left(\sqrt{KT + D\log K}\right)$, which is achieved by the algorithm of Zimmert and Seldin (2020). EXP3, which has lower computational complexity, achieves this lower bound up to the $\log K$ that factors KT, which is negligible if the average delay is larger than $O\left(\frac{K}{\log K}\right)$.

For bandit convex optimization, much less is understood. A breakthrough was made in Bubeck et al. (2017), that introduced a bandit convex optimization algorithm that achieves an expected regret of $O\left(n^{9.5}\sqrt{T}\log^{7.5} T\right)$. Recently, it was shown in Lattimore (2020) that an algorithm exists that achieves an expected regret of $O(n^{2.5}\sqrt{T}\log T)$, which improves the bound from Bubeck and Eldan (2016). However, the result in Lattimore (2020) is non-constructive so an algorithm that achieves the improved bound is still unknown. Compared to FKM, the algorithm in Bubeck et al. (2017) suffers from a few drawbacks. First, the n dependence is a high-degree polynomial, which is much worse than the linear dependence of FKM. Second, the algorithm proposed in Bubeck et al. (2017) has a T dependent complexity per round. Since this new algorithm may need a very large T to have lower regret than FKM, a T dependant complexity is a serious practical concern. Finally, it is an open question how robust the algorithm in Bubeck et al. (2017) is to delays. The main difficulty seems to be that their algorithm requires increasing the step-size multiplicatively by a factor larger than one once a certain condition holds. An increasing step-size sequence conflicts with the T, D-dependent tuning that optimizes the expected regret with delays or the decreasing step-size sequence that guarantees no weighted-regret for unbounded delay sequences. In contrast, FKM gives a weighted-regret bound as a function of η_t that can be easily tuned.

For the multi-agent case, we proved that if the algorithms have no weighted-regret, then the expected weighted ergodic distribution of play converges to the set of coarse correlated equilibria (CCE) for a general non-cooperative game. For a two-player zero-sum game, the weighted ergodic average of play converges in L^1 to the set of Nash equilibria. Then, we showed that FKM and EXP3 have no weighted-regret with their step-size sequence as the weight sequence even under significant unbounded delay sequences (e.g., $d_t = O(t\log t)$) for which their regret is $\Theta(T)$. Hence, by simulating a game and endowing the players with FKM or EXP3, we can use the weighted ergodic distribution or average to approximate an equilibrium. By tuning the weights according to the conditions we provide, this approximation method can still converge even if the algorithms have linear regret. Since delays are prevalent when simulating model-free multi-agent interactions (i.e., games), this extends the set of tools that can approximate equilibria in practice. Approximating equilibria of model-free games in a simulated environment can help to predict their outcomes in practice, or design distributed algorithms in case the equilibria have good global performance.

Our results highlight the role of no weighted-regret in online learning under delayed feedback. This motivates to further study the analogy between no weighted-regret under delays to no-regret

in the standard no-delay case. In particular, it might be possible to prove results on the internal weighted-regret, based on the techniques introduced in Blum and Mansour (2007) for multi-armed bandits. An internal weighted-regret bound will allow approximating the correlated equilibrium in delayed feedback environments, generalizing our result on external weighted-regret and the CCE.

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, 2011.

Itai Ashlagi, Dov Monderer, and Moshe Tennenholtz. On the value of correlation. *Journal of Artificial Intelligence Research*, 33:575–613, 2008.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 2018.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2019.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

Michael Bowling. Convergence and no-regret in multiagent learning. In *Advances in neural information processing systems*, 2005.

Sébastien Bubeck and Ronen Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589, 2016.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85. ACM, 2017.

Yang Cai and Constantinos Daskalakis. On minmax theorems for multiplayer games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 217–234, 2011.

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

Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pages 750–773, 2018.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Gerard Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

András György and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. *arXiv preprint arXiv:2010.06022*, 2020.

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

Sergiu Hart. *Simple adaptive strategies: from regret-matching to uncoupled dynamics*, volume 4. World Scientific, 2013.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Amélie Héliou, Panayotis Mertikopoulos, and Zhengyuan Zhou. Gradient-free online learning in games with delayed rewards. *arXiv preprint arXiv:2006.10911*, 2020.

Lisa Hellerstein, Thomas Lidbetter, and Daniel Pirutinsky. Solving zero-sum games using best-response oracles with applications to search games. *Operations Research*, 2019.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, 2013.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. *arXiv preprint arXiv:2012.14843*, 2020.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. *arXiv preprint arXiv:2106.02436*, 2021.

Tor Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3):311–334, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Anne Gael Manegueu, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, 2020.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, 2010.

Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(Suppl. 1):807–815, 1955.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, 2018.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in neural information processing systems*, 2015.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

George F Simmons. *Introduction to topology and modern analysis*, volume 44. Tokyo, 1963.

Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1):187–208, 2007.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, 2019.

Takashi Ui. Correlated equilibrium and concave games. *International Journal of Game Theory*, 37 (1):1–13, 2008.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, 2020.

Marcelo J Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter W Glynn, and Claire Tomlin. Countering feedback delays in multi-agent learning. In *Advances in Neural Information Processing Systems*, 2017.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, 2019.

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

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003.

Martin Zinkevich, John Langford, and Alex J Smola. Slow learners are fast. In *Advances in neural information processing systems*, 2009.

8. Proof of Theorem 1

We start by showing that $\mathbb{E}\{\rho_T\} = \mathbb{E}\left\{\frac{\sum_{t=1}^{T}\eta_t \delta_{\boldsymbol{a}_t}}{\sum_{t=1}^{T}\eta_t}\right\}$ converges to an ε-CCE of the game as $T \to \infty$, for every $\varepsilon > 0$. For each n, define the cost function $l_{n,t}(\boldsymbol{a}_n) = 1 - u_n(\boldsymbol{a}_n, \boldsymbol{a}_{-n,t})$. Let $\varepsilon > 0$. Since each player n has no weighted-regret with respect to d_t^n and η_t, then there exists a $T_0 > 0$ such that for all $T > T_0$, we have for every n and every action $\boldsymbol{a}_n \in \mathcal{A}$:

$$\begin{aligned}\mathbb{E}\left\{\mathbb{E}^{\boldsymbol{a}^*\sim\rho_T}\left\{u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n}^*\right) - u_n\left(\boldsymbol{a}^*\right)\right\}\right\} &\underset{(a)}{=} \mathbb{E}\left\{\frac{\sum_{t=1}^{T}\eta_t\left(u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n,t}\right) - u_n\left(\boldsymbol{a}_t\right)\right)}{\sum_{t=1}^{T}\eta_t}\right\} \\ &= \mathbb{E}\left\{\frac{\sum_{t=1}^{T}\eta_t\left(l_{n,t}\left(\boldsymbol{a}_{n,t}\right) - l_{n,t}\left(\boldsymbol{a}_n\right)\right)}{\sum_{t=1}^{T}\eta_t}\right\} \underset{(b)}{\leq} \varepsilon \end{aligned} \tag{54}$$

where (a) uses the definition of ρ_T and (b) follows since player n has no weighted-regret. Now pick $\boldsymbol{a}_n = \underset{\boldsymbol{a}_n' \in \mathcal{A}}{\arg\max}\mathbb{E}^{\boldsymbol{a}^*\sim\rho_T}\left\{u_n\left(\boldsymbol{a}_n', \boldsymbol{a}_{-n}^*\right)\right\}$ in (54). Since $\mathbb{E}^{\boldsymbol{a}^*\sim\rho_T}\left\{u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n}^*\right) - u_n\left(\boldsymbol{a}^*\right)\right\}$ is linear in ρ_T, then

$$\mathbb{E}\left\{\mathbb{E}^{\boldsymbol{a}^*\sim\rho_T}\left\{u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n}^*\right) - u_n\left(\boldsymbol{a}^*\right)\right\}\right\} = \mathbb{E}^{\boldsymbol{a}^*\sim\mathbb{E}\{\rho_T\}}\left\{u_n\left(\boldsymbol{a}_n, \boldsymbol{a}_{-n}^*\right) - u_n\left(\boldsymbol{a}^*\right)\right\} \tag{55}$$

so we conclude that by definition $\mathbb{E}\{\rho_T\}$ is an ε-CCE.

Let $\Delta > 0$. From Lemma 7, we know that there exists an $\varepsilon_\Delta > 0$ such that for all $\varepsilon \leq \varepsilon_\Delta$ we have $\min_{\rho^* \in \mathcal{C}_0} \|\rho_\varepsilon - \rho^*\| \leq \Delta$ for all $\rho_\varepsilon \in \mathcal{C}_\varepsilon$. From (54) we know that there exists a large enough T_1 such that for all $T > T_1$ we have $\mathbb{E}\{\rho_T\} \in \mathcal{C}_{\varepsilon_\Delta}$, which implies that $\min_{\rho^* \in \mathcal{C}_0} \|\mathbb{E}\{\rho_T\} - \rho^*\| \leq \Delta$. Therefore, $\mathbb{E}\{\rho_T\}$ converges to the set of CCE $\mathcal{C}_0$ as $T \to \infty$.

9. Proof of Theorem 2

Recall that the weighted ergodic average of $\boldsymbol{a}_t$ is $\bar{\boldsymbol{a}}_T \triangleq \frac{\sum_{t=1}^{T}\eta_t \boldsymbol{a}_t}{\sum_{t=1}^{T}\eta_t}$. Let $\varepsilon > 0$. Define the ergodic average of the value of the game by

$$\overline{u}_T = \frac{\sum_{t=1}^{T}\eta_t u\left(\boldsymbol{y}_t, \boldsymbol{z}_t\right)}{\sum_{t=1}^{T}\eta_t}. \tag{56}$$

Define the row cost function $l_{r,t}(\boldsymbol{y}) = u(\boldsymbol{y}, \boldsymbol{z}_t)$. Since the row player has no weighted-regret with respect to d_t^r and η_t then there exists a $T_1 > 0$ such that for all $T > T_1$ and every $\boldsymbol{y} \in \mathcal{A}$ (even in hindsight):

$$\begin{aligned}\mathbb{E}\left\{\overline{u}_T - u\left(\boldsymbol{y}, \bar{\boldsymbol{z}}_T\right)\right\} &\underset{(a)}{\leq} \mathbb{E}\left\{\frac{\sum_{t=1}^{T}\eta_t\left(u\left(\boldsymbol{y}_t, \boldsymbol{z}_t\right) - u\left(\boldsymbol{y}, \boldsymbol{z}_t\right)\right)}{\sum_{t=1}^{T}\eta_t}\right\} \\ &= \mathbb{E}\left\{\frac{\sum_{t=1}^{T}\eta_t\left(l_{r,t}\left(\boldsymbol{y}_t\right) - l_{r,t}\left(\boldsymbol{y}\right)\right)}{\sum_{t=1}^{T}\eta_t}\right\} \underset{(b)}{\leq} \frac{\varepsilon}{2} \end{aligned} \tag{57}$$

where (a) uses the concavity of $u(\boldsymbol{y}, \bar{\boldsymbol{z}}_T)$ in $\bar{\boldsymbol{z}}_T$ and (b) uses the no weighted-regret of the algorithm.

Define the column cost function $l_{c,t}(\boldsymbol{z}) = 1 - u(\boldsymbol{y}_t, \boldsymbol{z})$. Since the column player has no weighted-regret with respect to d_t^c and η_t, then there exists a $T_2 > 0$ such that for all $T > T_2$ and for every

$\boldsymbol{z} \in \mathcal{A}$ (even in hindsight):

$$\begin{aligned}\mathbb{E}\left\{u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}\right) - \overline{u}_T\right\} &\underset{(a)}{\leq} \mathbb{E}\left\{\frac{\sum_{t=1}^T \eta_t \left(u\left(\boldsymbol{y}_t, \boldsymbol{z}\right) - u\left(\boldsymbol{y}_t, \boldsymbol{z}_t\right)\right)}{\sum_{t=1}^T \eta_t}\right\} \\ &= \mathbb{E}\left\{\frac{\sum_{t=1}^T \eta_t \left(l_{c,t}\left(\boldsymbol{z}_t\right) - l_{c,t}\left(\boldsymbol{z}\right)\right)}{\sum_{t=1}^T \eta_t}\right\} \underset{(b)}{\leq} \frac{\varepsilon}{2} \end{aligned} \tag{58}$$

where (a) uses the convexity of $u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}\right)$ in $\bar{\boldsymbol{y}}_T$ and (b) uses the no weighted-regret of the algorithm.

Now define the best-response to $\bar{\boldsymbol{z}}_T$ as $\boldsymbol{y}_T^b = \arg\min_{\boldsymbol{y}'} u\left(\boldsymbol{y}', \bar{\boldsymbol{z}}_T\right)$ and the best-response to $\bar{\boldsymbol{y}}_T$ as $\boldsymbol{z}_T^b = \arg\max_{\boldsymbol{z}'} u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}'\right)$. By choosing $\boldsymbol{y} = \boldsymbol{y}_T^b$, $\boldsymbol{z} = \bar{\boldsymbol{z}}_T$ in (57) and (58) and adding them together we conclude that for all $T > \max\left\{T_1, T_2\right\}$

$$\mathbb{E}\left\{\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \min_{\boldsymbol{y}'} u\left(\boldsymbol{y}', \bar{\boldsymbol{z}}_T\right)\right|\right\} \underset{(a)}{=} \mathbb{E}\left\{\overline{u}_T - u\left(\boldsymbol{y}_T^b, \bar{\boldsymbol{z}}_T\right)\right\} + \mathbb{E}\left\{u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \overline{u}_T\right\} \leq \varepsilon \tag{59}$$

where (a) follows since $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \geq u\left(\boldsymbol{y}_T^b, \bar{\boldsymbol{z}}_T\right)$. By choosing instead $\boldsymbol{y} = \bar{\boldsymbol{y}}_T$, $\boldsymbol{z} = \boldsymbol{z}_T^b$ in (57) and (58) and adding them together we conclude that for all $T > \max\left\{T_1, T_2\right\}$

$$\mathbb{E}\left\{\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \max_{\boldsymbol{z}'} u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}'\right)\right|\right\} \underset{(a)}{=} \mathbb{E}\left\{\overline{u}_T - u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right)\right\} + \mathbb{E}\left\{u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}_T^b\right) - \overline{u}_T\right\} \leq \varepsilon \tag{60}$$

where (a) follows since $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \leq u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}_T^b\right)$.

Let $\Delta > 0$. From Lemma 7, we know that there exists an $\varepsilon_\Delta > 0$ such that for all $\varepsilon \leq \varepsilon_\Delta$ we have $\min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\| \leq \Delta$ for all $\boldsymbol{x}_\varepsilon \in \mathcal{N}_\varepsilon$. Let $\delta > 0$ and let $\varepsilon = \frac{\varepsilon_\Delta \delta}{2} > 0$. Then from (59),(60) we know that there exists a large enough T_3 such that for all $T > T_3$, using Markov inequality:

$$\mathbb{P}\left(\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \min_{\boldsymbol{y}'} u\left(\boldsymbol{y}', \bar{\boldsymbol{z}}_T\right)\right| \geq \varepsilon_\Delta\right) \leq \frac{\mathbb{E}\left\{\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \min_{\boldsymbol{y}'} u\left(\boldsymbol{y}', \bar{\boldsymbol{z}}_T\right)\right|\right\}}{\varepsilon_\Delta} = \frac{\delta}{2} \tag{61}$$

and

$$\mathbb{P}\left(\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \max_{\boldsymbol{z}'} u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}'\right)\right| \geq \varepsilon_\Delta\right) \leq \frac{\mathbb{E}\left\{\left|u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) - \max_{\boldsymbol{z}'} u\left(\bar{\boldsymbol{y}}_T, \boldsymbol{z}'\right)\right|\right\}}{\varepsilon_\Delta} = \frac{\delta}{2}. \tag{62}$$

Hence by the union bound over (61) and (62):

$$\mathbb{P}\left(\min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) - \boldsymbol{x}^*\| \leq \Delta\right) \geq \mathbb{P}\left((\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) \in \mathcal{N}_{\varepsilon_\Delta}\right) \geq 1 - \delta \tag{63}$$

so $(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T)$ converges in probability to the set of NE. Since $(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) - \boldsymbol{x}^*$ is bounded, it implies that $\mathbb{E}\left\{\min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T) - \boldsymbol{x}^*\|\right\} \to 0$ as $T \to \infty$. Since u is continuous, $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \to v$ in probability as $T \to \infty$ where v is the value of the game. Since u is bounded, $u\left(\bar{\boldsymbol{y}}_T, \bar{\boldsymbol{z}}_T\right) \to v$ in L^1 as $T \to \infty$.

10. The Set of Approximate Equilibria Approaches the Set of Equilibria

The following lemma shows that for a given game, the sets of ε-NE and ε-CCE converge to the sets of NE and CCE, respectively, when $\varepsilon \to 0$. It allows us to convert convergence to $\mathcal{N}_\varepsilon$ and $\mathcal{C}_\varepsilon$ for each $\varepsilon > 0$ to convergence to $\mathcal{N}_0$ and $\mathcal{C}_0$.

Lemma 7. *Let* $d_{\mathcal{N}}(\varepsilon) = \max_{\boldsymbol{x}_\varepsilon \in \mathcal{N}_\varepsilon} \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\|$ *and* $d_{\mathcal{C}}(\varepsilon) = \max_{\rho_\varepsilon \in \mathcal{C}_\varepsilon} \min_{\rho^* \in \mathcal{C}_0} \|\rho_\varepsilon - \rho^*\|$. *Then* $d_{\mathcal{N}}(\varepsilon) \to 0$ *and* $d_{\mathcal{C}}(\varepsilon) \to 0$ *as* $\varepsilon \to 0$.

Proof Let $A_\varepsilon(\boldsymbol{y}) = \left\{\boldsymbol{x} \in \mathcal{A}^N \,|\, u_i(\boldsymbol{x}) \geq u_i(\boldsymbol{y}_i, \boldsymbol{x}_{-i}) - \varepsilon, \forall i\right\}$. This is a compact set since $A_\varepsilon(\boldsymbol{y}) = \bigcap_i f_{i,\boldsymbol{y}}^{-1}([-\varepsilon, u_{i,\max}])$ for the continuous $f_{i,\boldsymbol{y}}(\boldsymbol{x}) = u_i(\boldsymbol{x}) - u_i(\boldsymbol{y}_i, \boldsymbol{x}_{-i})$ and $u_{i,\max} = \max_{\boldsymbol{x}\in\mathcal{A}^N} f_{i,\boldsymbol{y}}(\boldsymbol{x})$. Now note that $\mathcal{N}_\varepsilon = \bigcap_{\boldsymbol{y}\in\mathcal{A}^N} A_\varepsilon(\boldsymbol{y})$, since $\mathcal{N}_\varepsilon$ only includes action profiles $\boldsymbol{x}$ where no deviation $\boldsymbol{y}$ gives any player more than ε gain. Hence $\mathcal{N}_\varepsilon$ is compact for all $\varepsilon \geq 0$. Since $\mathcal{N}_{\frac{1}{n}}$ and $\mathcal{N}_0$ are compact, we can define the sequence $\{\tilde{\boldsymbol{x}}_n\}$ such that

$$\tilde{\boldsymbol{x}}_n \in \arg\max_{\boldsymbol{x}_\varepsilon \in \mathcal{N}_{\frac{1}{n}}} \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\|. \tag{64}$$

Since $\mathcal{A}^N$ is compact then the infinite sequence $\tilde{\boldsymbol{x}}_n$ has a subsequence $\tilde{\boldsymbol{x}}_{n_k}$ that converges to a point $\tilde{\boldsymbol{x}} \in \mathcal{A}^N$ (i.e., Bolzano–Weierstrass Theorem, see Simmons (1963)). Since $\mathcal{N}_{\varepsilon_1} \subseteq \mathcal{N}_{\varepsilon_2}$ if $\varepsilon_2 \geq \varepsilon_1$, we must have $\tilde{\boldsymbol{x}} \in \bigcap_{n=1}^\infty \mathcal{N}_{\frac{1}{n}}$ so $\max_{\boldsymbol{y}_i} u_i(\boldsymbol{y}_i, \tilde{\boldsymbol{x}}_{-i}) - u_i(\tilde{\boldsymbol{x}}) \leq \frac{1}{n}$ for all n, implying that $\max_{\boldsymbol{y}_i} u_i(\boldsymbol{y}_i, \tilde{\boldsymbol{x}}_{-i}) \leq u_i(\tilde{\boldsymbol{x}})$ and $\tilde{\boldsymbol{x}} \in \mathcal{N}_0$. The fact that $\mathcal{N}_{\varepsilon_1} \subseteq \mathcal{N}_{\varepsilon_2}$ if $\varepsilon_2 \geq \varepsilon_1$ also implies that $d_{\mathcal{N}}(\varepsilon)$ is non-increasing. Additionally, $d_{\mathcal{N}}(\varepsilon)$ is bounded since $\mathcal{A}^N$ is bounded. Hence, $\lim_{n\to\infty} d_{\mathcal{N}}\left(\frac{1}{n}\right)$ exists. However, since $\tilde{\boldsymbol{x}}_{n_k} \in \mathcal{N}_{\frac{1}{n_k}}$ and $\tilde{\boldsymbol{x}} \in \mathcal{N}_0$ then

$$d_{\mathcal{N}}\left(\frac{1}{n_k}\right) = \max_{\boldsymbol{x}_\varepsilon \in \mathcal{N}_{\frac{1}{n_k}}} \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|\boldsymbol{x}_\varepsilon - \boldsymbol{x}^*\| = \min_{\boldsymbol{x}^* \in \mathcal{N}_0} \|\tilde{\boldsymbol{x}}_{n_k} - \boldsymbol{x}^*\| \leq \|\tilde{\boldsymbol{x}}_{n_k} - \tilde{\boldsymbol{x}}\| \tag{65}$$

so $\lim_{k\to\infty} d_{\mathcal{N}}\left(\frac{1}{n_k}\right) = 0$ since $\tilde{\boldsymbol{x}}_{n_k} \to \tilde{\boldsymbol{x}}$. Hence, we must have $\lim_{n\to\infty} d_{\mathcal{N}}\left(\frac{1}{n}\right) = 0$ and $\lim_{\varepsilon\to 0} d_{\mathcal{N}}(\varepsilon) = 0$.

Let $\mathcal{P}\left(\mathcal{A}^N\right)$ be the set of all Borel probability measures over $\mathcal{A}^N$, equipped with the weak-* topology (see Simmons (1963)). As discussed in Section 3, a CCE is a CE when all of the departure functions are constant and therefore continuous. Hence, by simply replacing the zero constant of the half-space with $-\varepsilon$ in the last line of the proof of Stoltz and Lugosi (2007, Theorem 9, Page 194), we conclude from their Theorem 9 that $\mathcal{C}_\varepsilon$ is compact for all $\varepsilon \geq 0$. Define $\tilde{\rho}_n \in \arg\max_{\rho_\varepsilon \in \mathcal{C}_{\frac{1}{n}}} \min_{\rho^* \in \mathcal{C}_0} \|\rho_\varepsilon - \rho^*\|$. Then, Prokhorov's Theorem, given as Proposition 8 in (Stoltz and Lugosi, 2007, Page 194), states that there exists a subsequence $\tilde{\rho}_{n_k}$ that converges to a point $\tilde{\rho} \in \mathcal{P}\left(\mathcal{A}^{\mathcal{N}}\right)$. Since $\mathcal{C}_{\varepsilon_1} \subseteq \mathcal{C}_{\varepsilon_2}$ if $\varepsilon_2 \geq \varepsilon_1$ then $\tilde{\rho} \in \bigcap_{n=1}^\infty \mathcal{C}_{\frac{1}{n}}$ and therefore $\mathbb{E}^{\boldsymbol{x}^*\sim\tilde{\rho}}\left\{\max_{\boldsymbol{y}_i} u_i\left(\boldsymbol{y}_i, \boldsymbol{x}^*_{-i}\right) - u_i\left(\boldsymbol{x}^*\right)\right\} \leq 0$ so $\tilde{\rho} \in \mathcal{C}_0$. Following the same argument as in (65) we conclude that $\lim_{\varepsilon\to 0} d_{\mathcal{C}}(\varepsilon) = 0$. ■

11. Upper Bound on the Effect of the Delays on the Regret

Lemma 8. *Let $\{\eta_t\}$ be a non-increasing step-size sequence. Let $\{d_t\}$ be a delay such that the cost from round t is received at round $t + d_t$. Let $\mathcal{S}_t$ be the set of feedback samples received and used at round t. Define the set $\mathcal{M}^*$ of all samples that are not received and used before round T. Then*

$$\sum_{t=1}^{T} \sum_{s\in\mathcal{S}_t} \eta_s \left(\sum_{r=s}^{t-1} \sum_{q\in\mathcal{S}_r} \eta_q + \sum_{q\in\mathcal{S}_t, q<s} \eta_q \right) \leq 2 \sum_{t\notin\mathcal{M}^*} \eta_t^2 d_t. \tag{66}$$

Proof Up to the weights η_q, the quantity inside the parentheses in (66):

$$Q_{s,t} \triangleq \sum_{r=s}^{t-1} \sum_{q\in\mathcal{S}_r} \eta_q + \sum_{q\in\mathcal{S}_t, q<s} \eta_q \tag{67}$$

counts of the number of feedback samples received and used between round s and round t such that $s \in \mathcal{S}_t$, before the feedback from round s was used. To prove (66), we want to upper bound $\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s Q_{s,t}$ for all possible delay sequences $\{d_t\}$. A sample can be received and used between round s and round t if it belongs to one of two types:

1. The first type is a feedback sample from $q \geq s$ that is received and used before the feedback from round s is used. There are a maximum of d_s feedback samples of this type. We denote the contribution of these samples to $Q_{s,t}$ by $Q_{s,t}^1$. Each feedback sample can contribute $\eta_q \leq \eta_s$ with $q \geq s$ (since η_t is non-increasing) to $Q_{s,t}^1$ for $s \in \mathcal{S}_t$. We over count them by giving each $Q_{s,t}^1$ term all of its d_s possible samples of this type. Summing over all t:

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s Q_{s,t}^1 \leq \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s^2 d_s = \sum_{t \notin \mathcal{M}^*} \eta_s^2 d_s. \tag{68}$$

2. The second type is a feedback sample from $q < s$ that is received and used before s is used. We denote the contribution of these samples to $Q_{s,t}$ by $Q_{s,t}^2$. The samples from round q can contribute to a maximum of d_q different $Q_{s,t}^2$ terms, all with $s \geq q$. This follows simply because the feedback sample of q is not received before $q + d_q$. Let Γ_q be the set of rounds s such that the samples from round q contribute to $Q_{s,t}^2$. Then

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s Q_{s,t}^2 \underset{(a)}{=} \sum_{q \notin \mathcal{M}^*} \sum_{s \in \Gamma_q} \eta_s \eta_q \underset{(b)}{\leq} \sum_{q \notin \mathcal{M}^*} \eta_q^2 \left|\Gamma_q\right| \leq \sum_{q \notin \mathcal{M}^*} \eta_q^2 d_q \tag{69}$$

   where (a) follows since only rounds q that their feedback is received and used sometime before round T are counted in $Q_{s,t}^2$ for some s, t. Inequality (b) uses $\eta_s \leq \eta_q$ since η_t is non-increasing and $s \geq q$ for all $s \in \Gamma_q$.

Summing the contribution to (67) from these two possible types of samples, we have $Q_{s,t} = Q_{s,t}^1 + Q_{s,t}^2$ so by summing (68) and (69) we obtain (66). ■

12. Proof of Lemma 3: Weighted-Regret of FKM with Delays

Recall that $\mathcal{S}_t$ is the set of feedback samples received and used at round t, and that $\mathcal{M} = \{t \,|\, t + d_t > T\}$ is the set of samples that are not received before round T. Let $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathcal{K}} \sum_{t=1}^{T} \eta_t l_t(\boldsymbol{a})$ and note that $\boldsymbol{a}^*$ is random for an adaptive adversary. We have

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \mathbb{E}\left\{l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right\} &= \mathbb{E}\left\{\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(l_s(\boldsymbol{a}_s) - l_s(\boldsymbol{a}^*)\right)\right\} + \mathbb{E}\left\{\sum_{t \in \mathcal{M}} \eta_t \left(l_t(\boldsymbol{a}_t) - l_t(\boldsymbol{a}^*)\right)\right\} \\
&\underset{(a)}{\leq} \mathbb{E}\left\{\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(l_s(\boldsymbol{a}_s) - l_s(\boldsymbol{a}^*)\right)\right\} + \sum_{t \in \mathcal{M}} \eta_t \\
&\underset{(b)}{\leq} \mathbb{E}\left\{\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \left(l_s(\boldsymbol{x}_s) - l_s(\boldsymbol{a}^*)\right)\right\} + \sum_{t \in \mathcal{M}} \eta_t + \delta L \sum_{t \notin \mathcal{M}} \eta_t
\end{aligned} \tag{70}$$

where (a) uses $l_t(\boldsymbol{a}) \in [0,1]$ and (b) uses $\left|l_s(\boldsymbol{x}_s) - l_s(\boldsymbol{a}_s)\right| \leq L \left\|\boldsymbol{x}_s - \boldsymbol{a}_s\right\| \leq \delta L$.

Define s_-, s_+ as the step a moment before and a moment after the algorithm uses the feedback from round s, which updates the action from $\boldsymbol{a}_{s_-}$ to $\boldsymbol{a}_{s_+}$. Both s_- and s_+ are algorithm update

steps that take place in round t of the game if $s \in \mathcal{S}_t$. Define the projection $\boldsymbol{a}_\delta^* = \prod_{\mathcal{K}_\delta}(\boldsymbol{a}^*)$, where $\mathcal{K}_\delta = \left\{\boldsymbol{x} \,|\, \frac{1}{1-\delta}\boldsymbol{x} \in \mathcal{K}\right\}$ as defined in Algorithm 2. Recall that $\hat{l}(\boldsymbol{x}) = \mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{l(\boldsymbol{x}+\delta\boldsymbol{u})\}$. We bound the first term in (70) as follows

$$\begin{aligned}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(l_s(\boldsymbol{x}_s)-l_s(\boldsymbol{a}^*)\right) &\underset{(a)}{\leq} L|\mathcal{K}|\,\delta\sum_{t\notin\mathcal{M}}\eta_t+\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(l_s(\boldsymbol{x}_s)-l_s(\boldsymbol{a}_\delta^*)\right)\\ &\underset{(b)}{\leq}(2+|\mathcal{K}|)\,L\delta\sum_{t\notin\mathcal{M}}\eta_t+\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\hat{l}_s(\boldsymbol{x}_s)-\hat{l}_s(\boldsymbol{a}_\delta^*)\right)\end{aligned} \tag{71}$$

where (a) follows from the Lipschitz continuity of l_s since $\left\|\boldsymbol{a}^*-\prod_{\mathcal{K}_\delta}(\boldsymbol{a}^*)\right\| \leq \|\boldsymbol{a}^*-(1-\delta)\,\boldsymbol{a}^*\| \leq \delta|\mathcal{K}|$. Inequality (b) uses the Lipschitz continuity again, this time on $l_s(\boldsymbol{x}_s)$ and $l_s(\boldsymbol{a}_\delta^*)$:

$$\hat{l}_s(\boldsymbol{x})-l_s(\boldsymbol{x})=\mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{l_s(\boldsymbol{x}+\delta\boldsymbol{u})-l_s(\boldsymbol{x})\}\leq\delta L\mathbb{E}^{\boldsymbol{u}\in\mathbb{S}_1}\{\|\boldsymbol{u}\|\}=\delta L. \tag{72}$$

Now recall that $\boldsymbol{g}_t=\frac{n}{\delta}l_t(\boldsymbol{x}_t+\delta\boldsymbol{u}_t)\,\boldsymbol{u}_t$ where $\boldsymbol{u}_t$ is on the unit sphere $\mathbb{S}_1$, and define

$$h_t(\boldsymbol{x}) \triangleq \hat{l}_t(\boldsymbol{x})+\left(\boldsymbol{g}_t-\nabla\hat{l}_t(\boldsymbol{x}_t)\right)^T\boldsymbol{x} \tag{73}$$

for which $\nabla h_t(\boldsymbol{x}_t)=\boldsymbol{g}_t$ and $\mathbb{E}\{h_t(\boldsymbol{x})\}=\mathbb{E}\left\{\hat{l}_t(\boldsymbol{x})\right\}$ for all $\boldsymbol{x}$, and also $\mathbb{E}\{h_t(\boldsymbol{x}_t)\}=\mathbb{E}\left\{\hat{l}_t(\boldsymbol{x}_t)\right\}$ since $\boldsymbol{u}_t$ is independent of $\boldsymbol{x}_t$ and l_t (see Lemma 2). Next note that

$$\begin{aligned}&\left\|\boldsymbol{x}_{s_+}-\boldsymbol{a}_\delta^*\right\|^2=\left\|\prod_{\mathcal{K}_\delta}\left(\boldsymbol{x}_{s_-}-\eta_s\nabla h_s(\boldsymbol{x}_s)\right)-\boldsymbol{a}_\delta^*\right\|^2\underset{(a)}{\leq}\left\|\left(\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*\right)-\eta_s\nabla h_s(\boldsymbol{x}_s)\right\|^2\\ &=\left\|\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*\right\|^2-2\eta_s\left\langle\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*,\nabla h_s(\boldsymbol{x}_s)\right\rangle+\eta_s^2\|\nabla h_s(\boldsymbol{x}_s)\|^2=\left\|\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*\right\|^2+\eta_s^2\|\nabla h_s(\boldsymbol{x}_s)\|^2\\ &\quad-2\eta_s\left\langle\boldsymbol{x}_s-\boldsymbol{a}_\delta^*,\nabla h_s(\boldsymbol{x}_s)\right\rangle-2\eta_s\left\langle\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right)+\sum_{q\in\mathcal{S}_t,q<s}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right),\nabla h_s(\boldsymbol{x}_s)\right\rangle\\ &\underset{(b)}{\leq}\left\|\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*\right\|^2-2\eta_s\left(h_s(\boldsymbol{x}_s)-h_s(\boldsymbol{a}_\delta^*)\right)+\eta_s^2\|\boldsymbol{g}_s\|^2\\ &\quad-2\eta_s\left\langle\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right)+\sum_{q\in\mathcal{S}_t,q<s}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right),\nabla h_s(\boldsymbol{x}_s)\right\rangle\end{aligned} \tag{74}$$

where (a) follows since $\prod_{\mathcal{K}_\delta}$ is the projection of $\boldsymbol{x}_{s_-}-\eta_s\nabla h_s(\boldsymbol{x}_s)$ onto the convex $\mathcal{K}_\delta$. Inequality (b) uses the convexity and differentiability of h_s on $\mathcal{K}_\delta$, so $h_s(\boldsymbol{a}_\delta^*)\geq h_s(\boldsymbol{x}_s)+\langle\boldsymbol{a}_\delta^*-\boldsymbol{x}_s,\nabla h_s(\boldsymbol{x}_s)\rangle$.

12.1 Adaptive Adversary

First note that for any $\boldsymbol{x}\in\mathcal{K}$

$$\begin{aligned}\left|\sum_{t\notin\mathcal{M}}\eta_t\left(\hat{l}_t(\boldsymbol{x})-h_t(\boldsymbol{x})\right)\right| &\underset{(a)}{=}\left|\sum_{t\notin\mathcal{M}}\eta_t\left\langle\boldsymbol{g}_t-\nabla\hat{l}_t(\boldsymbol{x}_t),\boldsymbol{x}\right\rangle\right|\\ &\leq\|\boldsymbol{x}\|\left\|\sum_{t\notin\mathcal{M}}\eta_t\left(\boldsymbol{g}_t-\nabla\hat{l}_t(\boldsymbol{x}_t)\right)\right\|\leq|\mathcal{K}|\left\|\sum_{t\notin\mathcal{M}}\eta_t\left(\boldsymbol{g}_t-\nabla\hat{l}_t(\boldsymbol{x}_t)\right)\right\|\end{aligned} \tag{75}$$

where (a) uses (73). The expectation of the last term can be bounded as follows

$$\mathbb{E}^2\left\{\left\|\sum_{t\notin\mathcal{M}}\eta_t\left(\boldsymbol{g}_t-\nabla\hat{l}_t\left(\boldsymbol{x}_t\right)\right)\right\|\right\}\leq\mathbb{E}\left\{\left\|\sum_{t\notin\mathcal{M}}\eta_t\left(\boldsymbol{g}_t-\nabla\hat{l}_t\left(\boldsymbol{x}_t\right)\right)\right\|^2\right\}=$$
$$\sum_{t\notin\mathcal{M}}\eta_t^2\mathbb{E}\left\{\left\|\boldsymbol{g}_t-\nabla\hat{l}_t\left(\boldsymbol{x}_t\right)\right\|^2\right\}+\sum_{t_1\notin\mathcal{M}}\sum_{t_1\neq t_2}\eta_{t_1}\eta_{t_2}\mathbb{E}\left\{\left\langle\boldsymbol{g}_{t_1}-\nabla\hat{l}_{t_1}\left(\boldsymbol{x}_{t_1}\right),\boldsymbol{g}_{t_2}-\nabla\hat{l}_{t_2}\left(\boldsymbol{x}_{t_2}\right)\right\rangle\right\}\underset{(a)}{\leq}$$
$$2\sum_{t\notin\mathcal{M}}\eta_t^2\mathbb{E}\left\{\left\|\boldsymbol{g}_t\right\|^2+\left\|\nabla\hat{l}_t\left(\boldsymbol{x}_t\right)\right\|^2\right\}\underset{(b)}{\leq}2\sum_{t\notin\mathcal{M}}\eta_t^2\left(\frac{n^2}{\delta^2}+L^2\right)\quad(76)$$

where (a) uses that $\left\langle\boldsymbol{g}_{t_1}-\nabla\hat{l}_{t_1}\left(\boldsymbol{x}_{t_1}\right),\mathbb{E}\left\{\boldsymbol{g}_{t_2}-\nabla\hat{l}_{t_2}\left(\boldsymbol{x}_{t_2}\right)\mid\sigma\left(\mathcal{F}_{t_2},\boldsymbol{u}_{t_1}\right)\right\}\right\rangle=0$ for all $t_1<t_2$, which follows from Lemma 2. Inequality (b) uses that $\hat{l}_t$ is differentiable and L-Lipschitz continuous.

Then

$$\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\hat{l}_s\left(\boldsymbol{x}_s\right)-\hat{l}_s\left(\boldsymbol{a}_\delta^*\right)\right)\right\}\underset{(a)}{\leq}$$
$$\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(h_s\left(\boldsymbol{x}_s\right)-h_s\left(\boldsymbol{a}_\delta^*\right)\right)\right\}+|\mathcal{K}|\sqrt{2\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\left(\frac{n^2}{\delta^2}+L^2\right)}\underset{(b)}{\leq}$$
$$\frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\mathbb{E}\left\{\left\|\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*\right\|^2-\left\|\boldsymbol{x}_{s_+}-\boldsymbol{a}_\delta^*\right\|^2\right\}+\frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\mathbb{E}\left\{\left\|\boldsymbol{g}_s\right\|^2\right\}+|\mathcal{K}|\sqrt{2\sum_{t\notin\mathcal{M}}\eta_t^2\left(\frac{n^2}{\delta^2}+L^2\right)}$$
$$-\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\mathbb{E}\left\{\left\langle\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right)+\sum_{q\in\mathcal{S}_t,q<s}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right),\mathbb{E}\left\{\nabla h_s\left(\boldsymbol{x}_s\right)\mid\mathcal{F}_{s_-}\right\}\right\rangle\right\}\underset{(c)}{\leq}$$
$$\frac{|\mathcal{K}|^2}{2}+\frac{n^2}{2\delta^2}\sum_{t\notin\mathcal{M}}\eta_t^2+|\mathcal{K}|\sqrt{2\sum_{t\notin\mathcal{M}}\eta_t^2\left(\frac{n^2}{\delta^2}+L^2\right)}$$
$$+\frac{n}{\delta}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}+\sum_{q\in\mathcal{S}_t,q<s}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}\right)\quad(77)$$

where (a) uses (75) and (76) on $\boldsymbol{x}=\boldsymbol{a}_\delta^*$ and $\mathbb{E}\left\{h_s\left(\boldsymbol{x}_s\right)\right\}=\mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{x}_s\right)\right\}$, (b) uses (74) and (c) uses the telescopic sum with $\left\|\boldsymbol{x}_0-\boldsymbol{a}_\delta^*\right\|^2-\left\|\boldsymbol{x}_T-\boldsymbol{a}_\delta^*\right\|^2\leq|\mathcal{K}|^2$, that $\boldsymbol{g}_s=\frac{n}{\delta}l_s\left(\boldsymbol{x}_s+\delta\boldsymbol{u}_s\right)\boldsymbol{u}_s$, and also Cauchy-Schwarz and then applying Lemma 2 to obtain $\left\|\mathbb{E}\left\{\nabla h_s\left(\boldsymbol{x}_s\right)\mid\mathcal{F}_{s_-}\right\}\right\|\leq\mathbb{E}\left\{\left\|\boldsymbol{g}_s\right\|\mid\mathcal{F}_{s_-}\right\}\leq\frac{n}{\delta}$. Finally we bound the last term by bounding

$$\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|_2=\left\|\prod_{\mathcal{K}_\delta}\left(\boldsymbol{x}_{q_-}-\eta_q\boldsymbol{g}_q\right)-\boldsymbol{x}_{q_-}\right\|_2\underset{(a)}{\leq}\eta_q\left\|\boldsymbol{g}_q\right\|_2\leq n\frac{\eta_q}{\delta}\quad(78)$$

where (a) follows since $\prod_{\mathcal{K}_\delta}$ is the projection of $\boldsymbol{x}_{q_-}-\eta_q\boldsymbol{g}_q$ onto the convex $\mathcal{K}_\delta$. We obtain

$$\frac{n}{\delta}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}+\sum_{q\in\mathcal{S}_t,q<s}\mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}\right)$$
$$\leq\frac{n^2}{\delta^2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q+\sum_{q\in\mathcal{S}_t,q<s}\eta_q\right)\underset{(a)}{\leq}2\frac{n^2}{\delta^2}\sum_{t\notin\mathcal{M}}\eta_t^2d_t\quad(79)$$

where (a) uses Lemma 8. We conclude by applying (79) on (77) and adding (71) and (70).

12.2 Oblivious Adversary

For an oblivious adversary, l_q is not random and for $q > s$ does not depend on $\boldsymbol{a}_s$, so by Lemma 2

$$\left\|\mathbb{E}\left\{\nabla h_s\left(\boldsymbol{x}_s\right) \mid \mathcal{F}_{s_-}\right\}\right\| \underset{(a)}{=} \left\|\mathbb{E}\left\{\frac{n}{\delta} l_s\left(\boldsymbol{x}_s+\delta \boldsymbol{u}_s\right) \boldsymbol{u}_s \mid \mathcal{F}_{s_-}\right\}\right\| = \left\|\nabla \hat{l}_s\left(\boldsymbol{x}_s\right)\right\| \underset{(b)}{\leq} L \tag{80}$$

where (a) uses (73) and (b) follows since $\hat{l}_s$ is differentiable and L-Lipschitz continuous. Then

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{x}_s\right)-\hat{l}_s\left(\boldsymbol{a}_\delta^*\right)\right\} \underset{(a)}{=} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \mathbb{E}\left\{h_s\left(\boldsymbol{x}_s\right)-h_s\left(\boldsymbol{a}_\delta^*\right)\right\} \\
&\qquad \underset{(b)}{\leq} \frac{1}{2} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \mathbb{E}\left\{\left\|\boldsymbol{x}_{s_-}-\boldsymbol{a}_\delta^*\right\|^2-\left\|\boldsymbol{x}_{s_+}-\boldsymbol{a}_\delta^*\right\|^2\right\}+\frac{1}{2} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s^2 \mathbb{E}\left\{\left\|\boldsymbol{g}_s\right\|^2\right\}- \\
&\qquad \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s \mathbb{E}\left\{\left\langle \sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right)+\sum_{q \in \mathcal{S}_t, q<s}\left(\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right), \mathbb{E}\left\{\nabla h_s\left(\boldsymbol{x}_s\right) \mid \mathcal{F}_{s_-}\right\}\right\rangle\right\} \\
&\underset{(c)}{\leq} \frac{|\mathcal{K}|^2}{2}+\frac{n^2}{2 \delta^2} \sum_{t \notin \mathcal{M}} \eta_t^2+L \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s\left(\sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}+\sum_{q \in \mathcal{S}_t, q<s} \mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}\right)
\end{aligned} \tag{81}$$

where (a) uses $\mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{x}_s\right)\right\} = \mathbb{E}\left\{h_s\left(\boldsymbol{x}_s\right)\right\}$ and $\mathbb{E}\left\{\hat{l}_s\left(\boldsymbol{a}_\delta^*\right)\right\} = \mathbb{E}\left\{h_s\left(\boldsymbol{a}_\delta^*\right)\right\}$ as explained below (73) (since with an oblivious adversary, $\boldsymbol{a}_\delta^*$ is not random), (b) uses (74) and (c) uses the telescopic sum with $\left\|\boldsymbol{x}_0-\boldsymbol{a}_\delta^*\right\|^2-\left\|\boldsymbol{x}_T-\boldsymbol{a}_\delta^*\right\|^2 \leq |\mathcal{K}|^2$, $\left\|\boldsymbol{g}_s\right\| \leq \frac{n}{\delta}$, and Cauchy-Schwarz with (80). Finally, we use (78) to bound the last term in (81):

$$\begin{aligned}
&L \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s\left(\sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}+\sum_{q \in \mathcal{S}_t, q<s} \mathbb{E}\left\{\left\|\boldsymbol{x}_{q_+}-\boldsymbol{x}_{q_-}\right\|\right\}\right) \\
&\qquad \leq \frac{L n}{\delta} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \eta_s\left(\sum_{r=s}^{t-1} \sum_{q \in \mathcal{S}_r} \eta_q+\sum_{q \in \mathcal{S}_t, q<s} \eta_q\right) \underset{(a)}{\leq} 2 L \frac{n}{\delta} \sum_{t \notin \mathcal{M}} \eta_t^2 d_t
\end{aligned} \tag{82}$$

where (a) uses Lemma 8. We conclude by applying (82) on (81) and adding (71) and (70).

13. Proof of Lemma 6: Weighted-Regret of EXP3 with Delays

Recall that $\mathcal{M}^*$ is the set of missing or discarded samples. Let s_-, s_+ as the step a moment before and a moment after the algorithm uses the feedback from round s, which updates the mixed action from $\boldsymbol{p}_{s_-}$ to $\boldsymbol{p}_{s_+}$. Both s_- and s_+ are algorithm update steps that take place in round t of the game if $s \in \mathcal{S}_t$. Let s_T be the last feedback to be updated.

We begin by the standard EXP3 analysis for arbitrary $\tilde{\boldsymbol{l}}_s$ (Lattimore and Szepesvári, 2020), but with careful consideration to both the interleaved arrivals and the weight η_s. Recall that $\tilde{L}_t^{(i)} = \sum_{t \notin \mathcal{M}^*} \eta_t \frac{l_t^{(i)} 1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t}$, as defined in Algorithm 3. Define $\Phi(t) = -\log\left(\sum_{i=1}^K e^{-\tilde{L}_t^{(i)}}\right)$ and $\tilde{\boldsymbol{l}}_t = \left(0, ..., \frac{l_t^{(a_t)}}{p_t^{(a_t)}+\gamma_t}, ..., 0\right)$. Then

$$\begin{aligned}
\Phi(s_+) - \Phi(s_-) = -\log\left(\frac{\sum_{i=1}^K e^{-\tilde{L}_{s_-}^{(i)}} e^{-\eta_s \tilde{l}_s^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}\right) &= -\log\left(\sum_{i=1}^K p_{s_-}^{(i)} e^{-\eta_s \tilde{l}_s^{(i)}}\right) \\
&\underset{(a)}{\geq} -\log\left(\sum_{i=1}^K p_{s_-}^{(i)}\left(1 - \eta_s \tilde{l}_s^{(i)} + \frac{1}{2}\eta_s^2\left(\tilde{l}_s^{(i)}\right)^2\right)\right) \\
&= -\log\left(1 - \sum_{i=1}^K p_{s_-}^{(i)}\left(\eta_s \tilde{l}_s^{(i)} - \frac{1}{2}\eta_s^2\left(\tilde{l}_s^{(i)}\right)^2\right)\right) \\
&\underset{(b)}{\geq} \eta_s \sum_{i=1}^K p_{s_-}^{(i)} \tilde{l}_s^{(i)} - \frac{\eta_s^2}{2}\sum_{i=1}^K p_{s_-}^{(i)}\left(\tilde{l}_s^{(i)}\right)^2 \qquad (83)
\end{aligned}$$

where (a) is $e^{-x} \leq 1 - x + \frac{1}{2}x^2$ and (b) is $\log(1-x) \leq -x$. Hence, iterating (83) over s yields

$$\Phi\left(s_T^+\right) - \Phi(1) = \sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \left(\Phi(s_+) - \Phi(s_-)\right) \geq \sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \eta_s \sum_{i=1}^K p_{s_-}^{(i)} \tilde{l}_s^{(i)} - \frac{1}{2}\sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \eta_s^2 \sum_{i=1}^K p_{s_-}^{(i)}\left(\tilde{l}_s^{(i)}\right)^2. \qquad (84)$$

Next we upper bound $\Phi\left(s_T^+\right) - \Phi(1)$. We have for $i^* \triangleq \arg\min_i \sum_{t=1}^T \eta_t l_t^{(i)}$ that

$$\Phi\left(s_T^+\right) - \Phi(1) = -\log\left(\sum_{j=1}^K e^{-\tilde{L}_{s_T^+}^{(j)}}\right) + \log K \underset{(a)}{\leq} \tilde{L}_{s_T^+}^{(i^*)} + \log K = \sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \eta_s \tilde{l}_s^{(i^*)} + \log K \qquad (85)$$

where (a) omits positive terms from $\sum_{j=1}^K e^{-\tilde{L}_{s_T^+}^{(j)}}$. We conclude from (84) and (85) that

$$\sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \eta_s \sum_{i=1}^K p_{s_-}^{(i)} \tilde{l}_s^{(i)} - \sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \eta_s \tilde{l}_s^{(i^*)} \leq \log K + \frac{1}{2}\sum_{t=1}^T \sum_{s \in \mathcal{S}_t} \eta_s^2 \sum_{i=1}^K p_{s_-}^{(i)}\left(\tilde{l}_s^{(i)}\right)^2. \qquad (86)$$

Now observe that

$$\begin{aligned}
\sum_{i=1}^K p_{s_-}^{(i)} \tilde{l}_s^{(i)} = \sum_{i=1}^K p_{s_-}^{(i)} \frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s} &= \sum_{i=1}^K \left(p_s^{(i)} + \gamma_s\right)\frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s} - \sum_{i=1}^K \left(p_s^{(i)} + \gamma_s - p_{s_-}^{(i)}\right)\frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s} \\
&= l_s^{(a_s)} - \gamma_s \sum_{i=1}^K \tilde{l}_s^{(i)} + \sum_{i=1}^K \left(p_{s_-}^{(i)} - p_s^{(i)}\right)\tilde{l}_s^{(i)}. \qquad (87)
\end{aligned}$$

Using $\frac{p_{s_-}^{(i)}}{p_s^{(i)}} \leq e^2$ from Lemma 9, we obtain

$$\sum_{i=1}^{K} p_{s_-}^{(i)} \left(\tilde{l}_s^{(i)}\right)^2 = \sum_{i=1}^{K} p_{s_-}^{(i)} \left(\frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}\right)^2 \leq e^2 \sum_{i=1}^{K} \frac{p_s^{(i)}}{p_s^{(i)}+\gamma_s} \frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s} \leq e^2 \sum_{i=1}^{K} \tilde{l}_s^{(i)}. \tag{88}$$

Taking the weighted sum of (87) over $t \notin \mathcal{M}^*$ and subtracting $\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \tilde{l}_s^{(i^*)}$ from both sides:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s l_s^{(a_s)} - \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \gamma_s\eta_s \sum_{i=1}^{K} \tilde{l}_s^{(i)} + \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \sum_{i=1}^{K} \left(p_{s_-}^{(i)} - p_s^{(i)}\right) \tilde{l}_s^{(i)} - \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \tilde{l}_s^{(i^*)} \\
&= \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \sum_{i=1}^{K} p_{s_-}^{(i)} \tilde{l}_s^{(i)} - \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \tilde{l}_s^{(i^*)} \\
&\underset{(a)}{\leq} \log K + \frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s^2 \sum_{i=1}^{K} p_{s_-}^{(i)} \left(\tilde{l}_s^{(i)}\right)^2 \\
&\underset{(b)}{\leq} \log K + \frac{1}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} e^2\eta_s^2 \sum_{i=1}^{K} \tilde{l}_s^{(i)}
\end{aligned} \tag{89}$$

where (a) uses (86) and (b) uses (88). Rearranging (89) and subtracting $\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s l_s^{(i^*)}$ from both sides gives

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \left(l_s^{(a_s)} - l_s^{(i^*)}\right) \leq \\
&\underbrace{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \sum_{i=1}^{K} \left(p_s^{(i)} - p_{s_-}^{(i)}\right) \tilde{l}_s^{(i)}}_{A} + \log K + \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \left(\tilde{l}_s^{(i^*)} - l_s^{(i^*)}\right) + \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \left(\gamma_s\eta_s + \frac{e^2}{2}\eta_s^2\right) \sum_{i=1}^{K} \tilde{l}_s^{(i)}.
\end{aligned} \tag{90}$$

13.1 Adaptive Adversary with $\gamma_t = \eta_t$

13.1.1 Taking the Expectation

Define $W_1^{(i)} = \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right) - \log K$. From Lemma 11 with $\delta \leftarrow \frac{\delta}{K}$ and $\alpha_s^{(i)} = \eta_s$ and $\alpha_s^{(j)} = 0$ for all $j \neq i$ we get by using the union bound that

$$\mathbb{P}\left(\max_i W_1^{(i)} \geq \log\frac{1}{\delta}\right) \leq K\mathbb{P}\left(\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \eta_s \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right) \geq \log\frac{K}{\delta}\right) \leq \delta \tag{91}$$

so by substituting $x = \log\frac{1}{\delta}$ so $dx = -\frac{d\delta}{\delta}$

$$\mathbb{E}\left\{\max_i W_1^{(i)}\right\} \leq \int_0^{\infty} \mathbb{P}\left(\max_i W_1^{(i)} \geq x\right) dx = \int_0^{1} \frac{1}{\delta}\mathbb{P}\left(\max_i W_1^{(i)} \geq \log\frac{1}{\delta}\right) d\delta \leq 1. \tag{92}$$

Define $W_2 = \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \left(\gamma_s\eta_s + \frac{e^2}{2}\eta_s^2\right) \sum_{i=1}^{K} \left(\tilde{l}_s^{(i)} - l_s^{(i)}\right)$. From Lemma 11 with $\alpha_s^{(i)} = \gamma_s\eta_s + \frac{e^2}{2}\eta_s^2 \leq 2\eta_s$ for all i we get $\mathbb{P}\left(W_2 \geq \log\frac{1}{\delta}\right) \leq \delta$, so using (92) on W_2 we obtain $\mathbb{E}\{W_2\} \leq 1$.

From $\mathbb{E}\left\{\max_i W_1^{(i)}\right\} \leq 1$ and $\mathbb{E}\left\{W_2\right\} \leq 1$ we conclude that

$$\mathbb{E}\left\{\max_i \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\tilde{l}_s^{(i)}-l_s^{(i)}\right)+\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right)\sum_{i=1}^{K}\tilde{l}_s^{(i)}\right\} \leq$$
$$\log K+2+\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right)\sum_{i=1}^{K}l_s^{(i)} \leq \log K+2+K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right). \quad (93)$$

13.1.2 The Effect of Delays

Next we bound the A term in (90), which quantifies the effect of the delays. Let $s\in\mathcal{S}_t$, and let q be a round for which the feedback is used after or at round s, but before the feedback from round s is used. Define $h_i\left(\tilde{\boldsymbol{L}}_{q_-}\right)\triangleq p_{q_-}^{(i)}=\frac{e^{-\tilde{L}_{q_-}^{(i)}}}{\sum_{j=1}^{K}e^{-\tilde{L}_{q_-}^{(j)}}}$, so $p_{q^+}^{(i)}=h_i\left(\tilde{\boldsymbol{L}}_{q_-}+\eta_q\tilde{\boldsymbol{l}}_q\right)$. Using Lemma 10 with $\boldsymbol{x}=\tilde{\boldsymbol{L}}_{q_-}$ and $\Delta=\eta_q\tilde{\boldsymbol{l}}_q$ (so $\boldsymbol{h}\left(\boldsymbol{x}\right)=\boldsymbol{p}_{q_-}$) yields

$$\left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1 \leq 2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}\tilde{l}_q^{(i)}=2\eta_q\sum_{i=1}^{K}p_{q_-}^{(i)}\frac{l_q^{(i)}1_{\{a_q=i\}}}{p_q^{(i)}+\gamma_q}$$
$$\underset{(a)}{\leq} 2e^2\eta_q\sum_{i=1}^{K}p_q^{(i)}\frac{l_q^{(i)}1_{\{a_q=i\}}}{p_q^{(i)}+\gamma_q} \leq 2e^2\eta_q\sum_{i=1}^{K}l_q^{(i)}1_{\{a_q=i\}}=2e^2\eta_q l_q^{(q_q)} \leq 2e^2\eta_q \quad (94)$$

where (a) uses $\frac{p_{q_-}^{(i)}}{p_q^{(i)}}\leq e^2$ from Lemma 9. Hence

$$\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle=\sum_{i=1}^{K}\left(p_{q_-}^{(i)}-p_{q_+}^{(i)}\right)\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}\underset{(a)}{\leq}2e^2\eta_q\sum_{i=1}^{K}\frac{l_s^{(i)}1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}\leq 2e^2\eta_q\sum_{i=1}^{K}\frac{1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s} \quad (95)$$

where (a) follows since $p_{q_-}^{(i)}-p_{q_+}^{(i)}\leq\left|p_{q_-}^{(i)}-p_{q_+}^{(i)}\right|\leq\left\|\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\|_1\leq 2e^2\eta_q$. Using (95) we can write

$$\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\mid\mathcal{F}_s\right\}\leq 2e^2\eta_q\sum_{i=1}^{K}\mathbb{E}\left\{\frac{1_{\{a_s=i\}}}{p_s^{(i)}+\gamma_s}\mid\mathcal{F}_s\right\}\underset{(a)}{=}2e^2\eta_q\sum_{i=1}^{K}\frac{p_s^{(i)}}{p_s^{(i)}+\gamma_s}\leq 2e^2\eta_q K \quad (96)$$

where (a) uses that $p_s^{(i)}$ is $\mathcal{F}_s$-measurable and that $a_s = i$ with probability $p_s^{(i)}$. Hence the A term in (90) can be bounded as

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_s-\boldsymbol{p}_{s_-}\right\rangle\right\}\\
&=\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_t-\boldsymbol{p}_{s_-}\right\rangle+\sum_{r=s}^{t-1}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_r-\boldsymbol{p}_{r+1}\right\rangle\right)\right\}\\
&=\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right)\right\}\\
&=\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\langle\tilde{\boldsymbol{l}}_s,\boldsymbol{p}_{q_-}-\boldsymbol{p}_{q_+}\right\rangle\right\}\right)\\
&\underset{(a)}{\leq}2e^2K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t,q<s}\eta_q+\sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q\right)\underset{(b)}{\leq}4e^2K\sum_{t\notin\mathcal{M}^*}\eta_t^2d_t
\end{aligned}\tag{97}$$

where (a) uses (96) with the tower rule, and (b) uses Lemma 8.

13.1.3 Concluding the Proof

We conclude that for $i^*\triangleq\arg\min_i\sum_{t=1}^{T}\eta_t l_t^{(i)}$:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{t=1}^{T}\eta_t\left(l_t^{(a_t)}-l_t^{(i^*)}\right)\right\}&\underset{(a)}{\leq}\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(l_s^{(a_s)}-l_s^{(i^*)}\right)\right\}+\sum_{t\in\mathcal{M}^*}\eta_t\\
&\underset{(b)}{\leq}\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\sum_{i=1}^{K}\left(p_s^{(i)}-p_{s_-}^{(i)}\right)\tilde{l}_s^{(i)}\right\}+\log K\\
&\quad+\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\left(\eta_s\left(\tilde{l}_s^{(i^*)}-l_s^{(i^*)}\right)+\left(\gamma_s\eta_s+\frac{e^2}{2}\eta_s^2\right)\sum_{i=1}^{K}\tilde{l}_s^{(i)}\right)\right\}+\sum_{t\in\mathcal{M}^*}\eta_t\\
&\underset{(c)}{\leq}4e^2K\sum_{t\notin\mathcal{M}^*}\eta_t^2d_t+2+\left(1+\frac{e^2}{2}\right)K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2+2\log K+\sum_{t\in\mathcal{M}^*}\eta_t
\end{aligned}\tag{98}$$

where (a) uses that $0\leq l_t^{(i)}\leq 1$ for every i and t, (b) is (90) and (c) is (93) and (97) for $\gamma_s=\eta_s$.

13.2 Oblivious Adversary with $\gamma_t = 0$

13.2.1 Taking the Expectation

With an oblivious adversary $\mathbb{E}\left\{\tilde{l}_s^{(i)}\right\}=l_s^{(i)}\mathbb{E}\left\{\frac{1_{\{a_s=i\}}}{p_s^{(i)}}\right\}=l_s^{(i)}$ for each s and i, since $l_s^{(i)}$ is not random. Then for any i, in particular $i^*\triangleq\arg\min_i\sum_{t=1}^{T}\eta_t l_t^{(i)}$, we have

$$\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\tilde{l}_s^{(i^*)}-l_s^{(i^*)}\right)+\frac{e^2}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\sum_{i=1}^{K}\tilde{l}_s^{(i)}\right\}=\frac{e^2}{2}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2\sum_{i=1}^{K}l_s^{(i)}\leq\frac{e^2}{2}K\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s^2.\tag{99}$$

13.2.2 The Effect of Delays

Let $s \in \mathcal{S}_t$, and let q be a round for which the feedback is used after or at round s, but before the feedback from round s is used. Using Lemma 10 with $\boldsymbol{x} = \tilde{\boldsymbol{L}}_{q_-}$ and $\Delta = \eta_q \tilde{\boldsymbol{l}}_q$, so $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{p}_{q_-}$ yields

$$\begin{aligned}\mathbb{E}\left\{\left\|\boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\|_1 \mid \mathcal{F}_{q_-}\right\} &\leq 2\eta_q \mathbb{E}\left\{\sum_{i=1}^{K} p_{q_-}^{(i)} \tilde{l}_q^{(i)} \mid \mathcal{F}_{q_-}\right\} \\ &\underset{(a)}{=} 2\eta_q \sum_{i=1}^{K} p_{q_-}^{(i)} \mathbb{E}\left\{\tilde{l}_q^{(i)} \mid \mathcal{F}_{q_-}\right\} \underset{(b)}{=} 2\eta_q \sum_{i=1}^{K} p_{q_-}^{(i)} l_q^{(i)} \leq 2\eta_q \sum_{i=1}^{K} p_{q_-}^{(i)} = 2\eta_q \quad (100)\end{aligned}$$

where (a) uses that $p_{q_-}^{(i)}$ is $\mathcal{F}_{q_-}$-measurable and (b) uses that $p_q^{(i)}$ is $\mathcal{F}_{q_-}$-measurable (since $q < q_-$) and that $\tilde{l}_q^{(i)}$ is $\frac{l_q^{(i)}}{p_q^{(i)}}$ with probability $p_q^{(i)}$ and zero otherwise. Note that a_q given $\mathcal{F}_q$ is independent of $\mathcal{F}_{q_-}$ since by definition the feedback from a_q was not received until round q_-. This is unique to the oblivious adversary case when $\boldsymbol{l}_r$ for $r > s$ is not a random variable that depends on a_q, which the adversary observes already at the end of round q. Then for every $q \in \mathcal{S}_r$ for $r < t$ or $q \in \mathcal{S}_t$ such that $q < s$ we have

$$\begin{aligned}\mathbb{E}\left\{\left\langle \tilde{\boldsymbol{l}}_s, \boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\rangle\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\sum_{i=1}^{K}\left(p_{q_-}^{(i)} - p_{q_+}^{(i)}\right)\frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}} \mid \mathcal{F}_{s_-}\right\}\right\} \\ &= \mathbb{E}\left\{\sum_{i=1}^{K}\left(p_{q_-}^{(i)} - p_{q_+}^{(i)}\right)\mathbb{E}\left\{\frac{l_s^{(i)} 1_{\{a_s=i\}}}{p_s^{(i)}} \mid \mathcal{F}_{s_-}\right\}\right\} \\ &= \mathbb{E}\left\{\sum_{i=1}^{K}\left(p_{q_-}^{(i)} - p_{q_+}^{(i)}\right) l_s^{(i)}\right\} \\ &= \mathbb{E}\left\{\left\langle \boldsymbol{l}_s, \boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\rangle\right\} \underset{(a)}{\leq} \mathbb{E}\left\{\|\boldsymbol{l}_s\|_\infty \left\|\boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\|_1\right\} \\ &\underset{(b)}{\leq} \mathbb{E}\left\{\left\|\boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\|_1\right\} \underset{(c)}{\leq} 2\eta_q \quad (101)\end{aligned}$$

where (a) is Hölder's inequality, (b) uses $0 \leq l_t^{(i)} \leq 1$ and (c) uses (100) and the tower rule. Therefore

$$\begin{aligned}&\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left\langle \tilde{\boldsymbol{l}}_s, \boldsymbol{p}_s - \boldsymbol{p}_{s_-}\right\rangle\right\} \\ &= \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\left\langle \tilde{\boldsymbol{l}}_s, \boldsymbol{p}_t - \boldsymbol{p}_{s_-}\right\rangle + \sum_{r=s}^{t-1}\left\langle \tilde{\boldsymbol{l}}_s, \boldsymbol{p}_r - \boldsymbol{p}_{r+1}\right\rangle\right)\right\} \\ &= \sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t, q<s}\mathbb{E}\left\{\left\langle \tilde{\boldsymbol{l}}_s, \boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\rangle\right\} + \sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\mathbb{E}\left\{\left\langle \tilde{\boldsymbol{l}}_s, \boldsymbol{p}_{q_-} - \boldsymbol{p}_{q_+}\right\rangle\right\}\right) \\ &\underset{(a)}{\leq} 2\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t}\eta_s\left(\sum_{q\in\mathcal{S}_t, q<s}\eta_q + \sum_{r=s}^{t-1}\sum_{q\in\mathcal{S}_r}\eta_q\right) \underset{(b)}{\leq} 4\sum_{t\notin\mathcal{M}^*}\eta_t^2 d_t \quad (102)\end{aligned}$$

where (a) follows from (101) and (b) follows from Lemma 8.

13.2.3 Concluding the Proof

We conclude that for $i^* \triangleq \arg\min_i \sum_{t=1}^T \eta_t l_t^{(i)}$:

$$\mathbb{E}\left\{\sum_{t=1}^T \eta_t l_t^{(a_t)} - \sum_{t=1}^T \eta_t l_t^{(i^*)}\right\} \underset{(a)}{\leq} \mathbb{E}\left\{\sum_{t=1}^T \sum_{s\in\mathcal{S}_t} \eta_s \left(l_s^{(a_s)} - l_s^{(i^*)}\right)\right\} + \sum_{t\in\mathcal{M}^*} \eta_t \underset{(b)}{\leq} \log K + \frac{e^2}{2} K \sum_{t=1}^T \eta_t^2 + 4 \sum_{t\notin\mathcal{M}^*} \eta_t^2 d_t + \sum_{t\in\mathcal{M}^*} \eta_t \quad (103)$$

where (a) uses that $0 \leq l_t^{(i)} \leq 1$ for every i and t, and (b) uses (90), (99) and (102).

13.3 EXP3 Auxiliary Lemmas

The following lemma generalizes Lemma 2 from Cesa-Bianchi et al. (2019) to a sequence of delays $\{d_t\}$ and a sequence of step-sizes $\{\eta_t\}$.

Lemma 9. *Let $\{\eta_t\}$ be a positive non-increasing step-size sequence such that $\eta_t \leq \frac{1}{2}e^{-2}$ for all t. Let $\mathcal{D} = \left\{t \,|\, d_t \geq \frac{1}{e^2\eta_t} - 1\right\}$. Then for every s,t such that $s \in \mathcal{S}_t$ (so $s \notin \mathcal{D}$) Algorithm 3 maintains for all $i = 1,...,K$ both $\frac{p_{s_+}^{(i)}}{p_{s_-}^{(i)}} \leq \frac{1}{1-e^2\eta_s}$ and $\frac{p_{s_-}^{(i)}}{p_s^{(i)}} \leq e^2$.*

Proof The proof follows by induction on the feedback arrival index. Let s be the first feedback to arrive. Before that, at s_-, we have $p_{s_-}^{(i)} = \frac{1}{K}$ and $\frac{p_{s_-}^{(i)}}{p_s^{(i)}} = 1$ for all i. Then, the first update satisfies

$$\frac{p_{s_-}^{(i)}}{p_{s_+}^{(i)}} = \frac{\frac{1}{K}}{\frac{\frac{1}{K}e^{-\eta_s \tilde{l}_s^{(i)}}}{\sum_{j=1}^K \frac{1}{K} e^{-\eta_s \tilde{l}_s^{(j)}}}} \geq 1 - \frac{1}{K} + \frac{1}{K} e^{-\eta_s \frac{l_s^{(a_s)}}{\frac{1}{K}+\gamma_s}} \geq 1 + \frac{1}{K}\left(e^{-\eta_s K l_s^{(a_s)}} - 1\right) \geq 1 - \eta_s l_s^{(a_s)} \geq 1 - e^2\eta_s. \quad (104)$$

Now let s be any arbitrary round for which the feedback arrives at time t. According to the inductive hypothesis, we have $\frac{p_{q_+}^{(i)}}{p_{q_-}^{(i)}} \leq \frac{1}{1-e^2\eta_q}$ for all $q \in \{r \in \mathcal{S}_t, r < s\} \cup \left\{\bigcup_{r=s}^{t-1} \mathcal{S}_r\right\}$. Define s_0 as the minimal $q < s$ such that $s \leq q + d_q \leq t$ and $q \notin \mathcal{D}$ (if it exists). Then for all $i = 1,...,K$

$$\frac{p_{s_-}^{(i)}}{p_s^{(i)}} = \prod_{r=s}^{t-1} \prod_{q\in\mathcal{S}_r} \prod_{q\in\mathcal{S}_t, q<s} \frac{p_{q_+}^{(i)}}{p_{q_-}^{(i)}} \underset{(a)}{\leq} \prod_{r=s}^{t-1} \prod_{q\in\mathcal{S}_r} \prod_{q\in\mathcal{S}_t, q<s} \left(1 + \frac{e^2\eta_q}{1-e^2\eta_q}\right) \underset{(b)}{\leq} \left(1 + \frac{1}{e^{-2}\eta_{s_0}^{-1} - 1}\right)^{d_{s_0}} \left(1 + \frac{1}{e^{-2}\eta_s^{-1} - 1}\right)^{d_s} \underset{(c)}{\leq} e^2 \quad (105)$$

where (a) uses the inductive hypothesis and (c) uses that by definition $d_{s_0} \leq e^{-2}\eta_{s_0}^{-1} - 1$ and $d_s \leq e^{-2}\eta_s^{-1} - 1$. If s_0 does not exist then the first factor is one (i.e., $d_{s_0} = 0$). Inequality (b) uses that the product runs over all rounds $q \notin \mathcal{D}$ for which the feedback is received between s and s_-. Feedback from $q \in \mathcal{D}$ is discarded and has no effect on $\frac{p_{s_-}^{(i)}}{p_s^{(i)}}$. The received feedback includes no more than d_{s_0} samples of rounds before s. This follows since there are at most d_{s_0} rounds between s_0 and s (since $s \leq s_0 + d_{s_0}$ by definition), and each of them contributes at most one feedback that is received between s and s_-. We have $\eta_q \leq \eta_{s_0}$ for each such round q, since η_t is non-increasing. It also includes no more than d_s feedback samples of rounds after s, since all these feedback samples are received before s_-, which occurs during round $t = s + d_s$. We have $\eta_r \leq \eta_s$ for each such round

r. We conclude that the update at s_-, occurring at time t using the feedback for a_s, satisfies:

$$\frac{p_{s_-}^{(i)}}{p_{s_+}^{(i)}} = \frac{\frac{e^{-\tilde{L}_{s_-}^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}}{\frac{e^{-\tilde{L}_{s_+}^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_+}^{(j)}}}} = \frac{\frac{e^{-\tilde{L}_{s_-}^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}}{\frac{e^{-\tilde{L}_{s_-}^{(i)}} e^{-\eta_s \tilde{l}_s^{(i)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}} e^{-\eta_s \tilde{l}_s^{(j)}}}} \geq \frac{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}} e^{-\eta_s \tilde{l}_s^{(j)}}}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}} \geq \frac{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}} \left(1 - \eta_s \tilde{l}_s^{(j)}\right)}{\sum_{j=1}^K e^{-\tilde{L}_{s_-}^{(j)}}}$$

$$= 1 - \eta_s \sum_{j=1}^K p_{s_-}^{(j)} \tilde{l}_s^{(j)} = 1 - \eta_s p_{s_-}^{(a_s)} \frac{l_s^{(a_s)}}{p_s^{(a_s)} + \gamma_s} \geq 1 - \eta_s \frac{p_{s_-}^{(a_s)}}{p_s^{(a_s)}} \underset{(a)}{\geq} 1 - e^2 \eta_s \quad (106)$$

where (a) follows from (105). Hence $\frac{p_{s_+}^{(i)}}{p_{s_-}^{(i)}} \leq \frac{1}{1-e^2\eta_s}$ and the proof is complete. ■

The next lemma shows standard smoothness properties of the softmax function, and we provide it here for completeness.

Lemma 10. *Let* $h_i(\boldsymbol{x}) = \frac{e^{-x_i}}{\sum_{j=1}^K e^{-x_j}}$ *and* $h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), ..., h_K(\boldsymbol{x}))$. *Then* $\forall \boldsymbol{x} \in \mathbb{R}^K$ *and* $\forall \Delta \in \mathbb{R}_+^K$

$$\|h(\boldsymbol{x}) - h(\boldsymbol{x} + \Delta)\|_1 \leq 2 \langle h(\boldsymbol{x}), \Delta \rangle. \quad (107)$$

Proof For all $\boldsymbol{x} \in \mathbb{R}^K$ and $\Delta \in \mathbb{R}_+^K$

$$h_i(\boldsymbol{x} + \Delta) - h_i(\boldsymbol{x}) = \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j - \Delta_j}} - \frac{e^{-x_i}}{\sum_{j=1}^K e^{-x_j}} \underset{(a)}{\geq} \left(e^{-\Delta_i} - 1\right) h_i(\boldsymbol{x}) \underset{(b)}{\geq} -\Delta_i h_i(\boldsymbol{x}) \quad (108)$$

where (a) follows since $\sum_{j=1}^K e^{-x_j - \Delta_j} \leq \sum_{j=1}^K e^{-x_j}$ and (b) since $1 - x \leq e^{-x}$ for all $x \geq 0$. We also have for all $\boldsymbol{x} \in \mathbb{R}^K$ and $\Delta \in \mathbb{R}_+^K$ that

$$h_i(\boldsymbol{x} + \Delta) - h_i(\boldsymbol{x}) = \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j - \Delta_j}} - \frac{e^{-x_i}}{\sum_{j=1}^K e^{-x_j}} \underset{(a)}{\leq} \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j - \Delta_j}} - \frac{e^{-x_i - \Delta_i}}{\sum_{j=1}^K e^{-x_j}} =$$

$$h_i(\boldsymbol{x} + \Delta) \left(1 - \frac{\sum_{j=1}^K e^{-x_j - \Delta_j}}{\sum_{l=1}^K e^{-x_l}}\right) = h_i(\boldsymbol{x} + \Delta) \frac{\sum_{j=1}^K e^{-x_j} \left(1 - e^{-\Delta_j}\right)}{\sum_{l=1}^K e^{-x_l}} \underset{(b)}{\leq} h_i(\boldsymbol{x} + \Delta) \frac{\sum_{j=1}^K \Delta_j e^{-x_j}}{\sum_{l=1}^K e^{-x_l}} \quad (109)$$

where (a) uses $e^{-x_j} \geq e^{-x_j - \Delta_j}$ and (b) uses $1 - x \leq e^{-x}$ for all $x \geq 0$. Combining (108) and (109):

$$\|h(\boldsymbol{x}) - h(\boldsymbol{x} + \Delta)\|_1 = \sum_{i=1}^K |h_i(\boldsymbol{x}) - h_i(\boldsymbol{x} + \Delta)| \underset{(a)}{\leq} \sum_{i=1}^K h_i(\boldsymbol{x} + \Delta) \left(\sum_{j=1}^K \frac{\Delta_j e^{-x_j}}{\sum_{l=1}^K e^{-x_l}}\right)$$

$$+ \sum_{i=1}^K \Delta_i h_i(\boldsymbol{x}) = \left(\sum_{j=1}^K \left(\Delta_j \frac{e^{-x_j}}{\sum_{l=1}^K e^{-x_l}}\right)\right) \sum_{i=1}^K h_i(\boldsymbol{x} + \Delta) + \langle h(\boldsymbol{x}), \Delta \rangle \underset{(b)}{=} 2 \langle h(\boldsymbol{x}), \Delta \rangle \quad (110)$$

where (a) uses $|h_i(\boldsymbol{x} + \Delta) - h_i(\boldsymbol{x})| \leq \max \left\{ \Delta_i h_i(\boldsymbol{x}), h_i(\boldsymbol{x} + \Delta) \frac{\sum_{j=1}^K \Delta_j e^{-x_j}}{\sum_{l=1}^K e^{-x_l}} \right\}$ for all i, due to (108) and (109). Equality (b) uses that $\sum_{i=1}^K h_i(\boldsymbol{x} + \Delta) = 1$ by definition. ■

The next Lemma is taken from Neu (2015), and we provide a (very) slightly modified proof to verify that the same result holds even when the order of arrivals changes as a result of the delayed feedback.

Lemma 11. *Let* $\tilde{l}_t^{(i)} = \frac{l_t^{(i)} 1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t}$. *If* $\left\{\alpha_t^{(i)}\right\}$ *is a non-negative sequence such that* $\alpha_t^{(i)} \leq 2\gamma_t$ *for all* t *and all* i, *then*

$$\mathbb{P}\left(\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^{K}\alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right) > \log\frac{1}{\delta}\right) \leq \delta. \tag{111}$$

Proof Define the filtration $\mathcal{G}_t = \sigma\left(\{a_\tau \,|\, \tau < t\}\right)$ and note that $\boldsymbol{l}_t$ is $\mathcal{G}_t$-measurable. We have

$$\tilde{l}_t^{(i)} = \frac{l_t^{(i)}1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t} \leq \frac{l_t^{(i)}1_{\{a_t=i\}}}{p_t^{(i)}+\gamma_t l_t^{(i)}1_{\{a_t=i\}}} = \frac{1}{2\gamma_t}\frac{2\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}}{1+\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}} \underset{(a)}{\leq} \frac{1}{2\gamma_t}\log\left(1+2\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right) \tag{112}$$

where (a) uses $\frac{x}{1+\frac{x}{2}} \leq \log(1+x)$ which holds for $x \geq 0$. Then

$$\begin{aligned}
\mathbb{E}\left\{e^{\sum_{i=1}^K \alpha_t^{(i)}\tilde{l}_t^{(i)}} \,|\, \mathcal{G}_t\right\} &\leq \mathbb{E}\left\{e^{\sum_{i=1}^K \frac{\alpha_t^{(i)}}{2\gamma_t}\log\left(1+2\gamma_t\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right)} \,|\, \mathcal{G}_t\right\} \\
&\underset{(a)}{\leq} \mathbb{E}\left\{e^{\sum_{i=1}^K \log\left(1+\alpha_t^{(i)}\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right)} \,|\, \mathcal{G}_t\right\} = \mathbb{E}\left\{\prod_{i=1}^K\left(1+\alpha_t^{(i)}\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}}\right) \,|\, \mathcal{G}_t\right\} \\
&\underset{(b)}{=} \mathbb{E}\left\{1+\sum_{i=1}^K \alpha_t^{(i)}\frac{l_t^{(i)}}{p_t^{(i)}}1_{\{a_t=i\}} \,|\, \mathcal{G}_t\right\} \underset{(c)}{=} 1+\sum_{i=1}^K \alpha_t^{(i)} l_t^{(i)} \leq e^{\sum_{i=1}^K \alpha_t^{(i)} l_t^{(i)}}
\end{aligned} \tag{113}$$

where (a) uses $x\log(1+y) \leq \log(1+xy)$ which holds for $y > -1$ and $0 \leq x \leq 1$, since $\alpha_t^{(i)} \leq 2\gamma_t$. Inequality (b) uses that $1_{\{a_t=i\}}1_{\{a_t=j\}} = 0$ for all $i \neq j$, and (c) uses that $p_t^{(i)}$ and $l_t^{(i)}$ are $\mathcal{G}_t$-measurable and that given $\mathcal{G}_t$, $a_t = i$ with probability $p_t^{(i)}$.

Since $l_t^{(i)}$ is $\mathcal{G}_t$-measurable then (113) yields $\mathbb{E}\left\{e^{\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)} \,|\, \mathcal{G}_t\right\} \leq 1$ for all i. Let $S = \sum_{t=1}^T |\mathcal{S}_t|$. Let τ_l be l-th round for which the feedback is not missing or discarded, so $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_S$. Then

$$\begin{aligned}
&\mathbb{E}\left\{e^{\sum_{l=1}^S\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)}\right\} = \mathbb{E}\left\{\mathbb{E}\left\{e^{\sum_{l=1}^S\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)} \,|\, \mathcal{G}_{\tau_S}\right\}\right\} \underset{(a)}{=} \\
&\quad \mathbb{E}\left\{e^{\sum_{l=1}^{S-1}\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)}\mathbb{E}\left\{e^{\sum_{i=1}^K \alpha_{\tau_S}^{(i)}\left(\tilde{l}_{\tau_S}^{(i)}-l_{\tau_S}^{(i)}\right)} \,|\, \mathcal{G}_{\tau_S}\right\}\right\} \leq \mathbb{E}\left\{e^{\sum_{l=1}^{S-1}\sum_{i=1}^K \alpha_{\tau_l}^{(i)}\left(\tilde{l}_{\tau_l}^{(i)}-l_{\tau_l}^{(i)}\right)}\right\}
\end{aligned} \tag{114}$$

where (a) uses that $a_{\tau_1}, \boldsymbol{l}_{\tau_1}, \boldsymbol{p}_{\tau_1}, \ldots, a_{\tau_{S-1}}, \boldsymbol{l}_{\tau_{\tau_{S-1}}}, \boldsymbol{p}_{\tau_{S-1}}$ are $\mathcal{G}_{\tau_S}$-measurable.

Iterating over (114) yields $\mathbb{E}\left\{e^{\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)}\right\} \leq 1$, so by Markov's inequality

$$\begin{aligned}
\mathbb{P}\left(\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^{K}\alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right) > \log\frac{1}{\delta}\right) &= \mathbb{P}\left(e^{\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)} > \frac{1}{\delta}\right) \\
&\leq \delta\mathbb{E}\left\{e^{\sum_{t\notin\mathcal{M}^*}\sum_{i=1}^K \alpha_t^{(i)}\left(\tilde{l}_t^{(i)}-l_t^{(i)}\right)}\right\} \leq \delta.
\end{aligned} \tag{115}$$

■