Capacity-Constrained Online Learning with Delays: Scheduling Frameworks and Regret Trade-offs

Alexander Ryabchenko   ALEXANDER.RYABCHENKO@MAIL.UTORONTO.CA
University of Toronto and Vector Institute

Idan Attias   IDANATTIAS88@GMAIL.COM
University of Illinois at Chicago and Toyota Technological Institute at Chicago [*]

Daniel M. Roy   DANIEL.ROY@UTORONTO.CA
University of Toronto and Vector Institute

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study online learning with oblivious losses and delays under a novel "capacity constraint" that limits how many past rounds can be tracked simultaneously for delayed feedback. Under "clairvoyance" (i.e., delay durations are revealed upfront each round) and/or "preemptibility" (i.e., we have ability to stop tracking previously chosen round feedback), we establish matching upper and lower bounds (up to logarithmic terms) on achievable regret, characterizing the "optimal capacity" needed to match the minimax rates of classical delayed online learning, which implicitly assume unlimited capacity. Our algorithms achieve minimax-optimal regret across all capacity levels, with performance gracefully degrading under suboptimal capacity. For K actions and total delay D over T rounds, under clairvoyance and assuming capacity $C = \Omega(\log(T))$, we achieve regret $\widetilde{\Theta}(\sqrt{TK + DK/C + D\log(K)})$ for bandits and $\widetilde{\Theta}(\sqrt{(D+T)\log(K)})$ for full-information feedback. When replacing clairvoyance with preemptibility, we require a known maximum delay bound $d_{\max}$, adding $\widetilde{O}(d_{\max})$ to the regret. For fixed delays d (i.e., $D = Td$), the minimax regret is $\Theta(\sqrt{TK(1 + d/C) + Td\log(K)})$ and the optimal capacity is $\Theta(\min\{K/\log(K), d\})$ in the bandit setting, while in the full-information feedback setting, the minimax regret is $\Theta(\sqrt{T(d+1)\log(K)})$ and the optimal capacity is $\Theta(1)$. For round-dependent and fixed delays, our upper bounds are achieved using novel preemptive and non-preemptive scheduling policies, based on Pareto-distributed proxy delays, and batching techniques, respectively. Crucially, our work unifies delayed bandits, label-efficient learning, and online scheduling frameworks, demonstrating that robust online learning under delayed feedback is possible with surprisingly modest tracking capacity.

1. Introduction

Online learning is a fundamental sequential decision-making problem in which a player repeatedly selects actions, each with some associated loss. By exploiting feedback after each action, the player aims to minimize some notion of regret, i.e., cumulative loss, compared to that of a class of alternative choices (Cesa-Bianchi and Lugosi, 2006). In this work, we study external regret, comparing the player's cumulative loss to that of the best single action in hindsight.

The type of feedback the player receives is an important aspect of the problem. One way in which feedback can vary is by how much information is revealed about the losses. Two important types of feedback are bandit feedback (Lattimore and Szepesvári, 2020; Slivkins, 2019), where the player learns only the loss for the action they took, and full information, where the player learns the

---

[*] Part of this work was completed while IA was at the Vector Institute.

loss for all actions. Another way in which feedback can vary is by when the feedback arrives. A well-studied variant considers delayed feedback, where action losses are revealed only after several rounds, forcing the player to act without immediate information about losses (Mesterharm, 2005; Joulani et al., 2013; Cesa-Bianchi et al., 2016). For example, in most recommendation systems, a platform suggests content or products to users but receives feedback (such as clicks or purchases) only after the interaction ends, requiring it to make new recommendations while relying on delayed and possibly outdated feedback.

Previous studies of online learning with delays typically assume that feedback from every round is eventually observed, even if only at the end of the game. In practice, however, resource constraints often limit how many rounds can be tracked concurrently for delayed feedback. This is particularly relevant in human-in-the-loop systems (e.g., see Lykouris and Weng (2024)). Consider a healthcare model where the system evaluates the long-term effects of K approved drugs across a continuous stream of patients. Each treatment may result in delayed outcomes, such as recovery or side effects, that take days or weeks to manifest. Due to a limited number of highly qualified and in-demand doctors, only C patients at any given time can be actively monitored via home visits, follow-ups, or remote diagnostics. As treatments progress and multiple drugs remain in use, tracking all T patients until their outcomes becomes infeasible, requiring up to $\Omega(T)$ tracking slots in the worst case. With only C slots, the system must reallocate attention dynamically, releasing patients with slow outcomes to accommodate new ones, in order to ensure a steady flow of informative feedback. This underscores the need for capacity-aware monitoring policies that manage exploration under resource constraints.

As another example, consider a recommendation system operating on a massive user stream, where delays arise from concurrently handling many users. Observing user behavior often requires maintaining an open session for each user until a conversion event (e.g., a purchase) occurs. However, since only a limited number of sessions can be maintained at once, it becomes impossible to track every user. Thus, resource constraints naturally arise in this scenario as well.

Motivated by these examples, we propose a resource-efficient version of online learning with delays, where in order to observe feedback from a particular round, that round must be continuously tracked until its delay period ends, and the number of rounds tracked simultaneously for delayed feedback is capped by a specified limit C, which we term *capacity*. We refer to this broader framework as *Delay Scheduling*, drawing an analogy to Online Job Scheduling (e.g., see Borodin and El-Yaniv (1998)), a problem that involves assigning sequentially arriving jobs between multiple resources with the goal of optimizing specific objective, such as maximizing the number of completed jobs. In analogy to Online Job Scheduling, several variations arise naturally in our Delay Scheduling:

- **Clairvoyant VS Non-Clairvoyant:** If at the start of each round, the player observes this round's delay (d_t), we refer to this as the *clairvoyant* framework. In contrast, if the player discovers this delay only when the feedback actually arrives (at the end of round $t + d_t$), we call it *non-clairvoyant*. Delayed bandits under clairvoyance were previously studied by Thune et al. (2019), who leveraged this upfront information about delays to eliminate the need for prior knowledge of the horizon T and total delay D, while also removing the assumption of bounded delays.
- **Preemptive VS Non-Preemptive:** If the framework allows the player to stop tracking rounds before receiving their feedback, we define it as *preemptive*; otherwise, it is *non-preemptive*. Importantly, once we stop tracking a round (i.e., preempt it), we cannot resume tracking it later,[1] aligning our framework with Online Interval Scheduling (Lipton, 1994; Woeginger, 1994).

---

1. In this paper, "preemption" refers to permanently stopping tracking a round (Cf. "revoking" Borodin and Karavasilis (2023)). In Online Scheduling, preemption may refer to pausing a task with the possibility of resumption or restarting.

While job scheduling typically optimizes metrics such as throughput, makespan, or latency, Delay Scheduling is an online learning problem where the goal is to minimize regret by strategically allocating resources in order to observe representative feedback. This introduces a novel challenge: balancing exploration and timely feedback collection while deciding which rounds to track under limited capacity. Consequently, standard scheduling techniques do not directly apply to our setting.

Prior work on Delayed Online Learning implicitly leveraged scheduling concepts to enhance algorithm performance, with Thune et al. (2019); Zimmert and Seldin (2020) proposing "skipping schemes" to exclude rounds with excessive delays to improve regret bounds. However, the impact of limited capacity remained unexplored. For additional related work, see Appendix A.

1.1. Problem Setting

In the Delay Scheduling game (Figure 1), the player interacts with an environment determined by an oblivious adversary over T rounds. Before the game begins, the adversary sets delays and losses for each round. Without loss of generality, we assume that $d_{\max} \leqslant T$ because the feedback of any round t with $t + d_t > T$ will not be observed. The player repeatedly selects actions from a fixed set while maintaining a *tracking set* of at most C round indices (the "capacity constraint"). The player receives feedback only for rounds that are currently in the tracking set and may modify this set according to the rules of the specific game variation being played.

Delay Scheduling Game

- *Visible Parameters:* number of actions K and capacity C.
- *Latent Parameters:* number of rounds T.
- *Pre-game:* adversary selects losses $l_t \in [0, 1]^K$ and delays $d_t \in \mathbb{Z}_{\geqslant 0}$ for all $t \in [T]$.

Player initializes empty tracking set S of maximum size C.
For each round $t = 1, 2, \ldots, T$:
   0. If the framework is **clairvoyant**, then the environment reveals delay d_t.
   1. The player selects action $A_t \in \{1, ..., K\}$, plays it, and incurs corresponding loss l_{t,A_t}.
   2. The player may add round index t to the tracking set S, provided $|S| < C$.
   3. For all $s \leqslant t$ such that $s + d_s = t$ and $s \in S$, the environment reveals
      • round-value pair (s, l_{s,A_s}) in the **multi-armed bandit** game,
      • round-vector pair (s, l_s) in the **full-information** game,
   and s is automatically removed from the tracking set S.
   4. If the framework is **preemptive**, the player may remove elements (possibly none) from S.

Figure 1 The Delay Scheduling Game in all variations: clairvoyant vs. non-clairvoyant, preemptive vs. non-preemptive, and full-information vs. bandit feedback.

Since delays are assigned to rounds rather than round-action pairs, we track rounds using C units of *round-based* capacity. In the full-information regime, each unit of capacity tracks all K losses from the corresponding round. In the bandit regime, it tracks only the loss of the selected action.

The player's objective is to minimize *expected regret*, $\mathfrak{R}_T = \mathbb{E}[\sum_{t=1}^{T} l_{t,A_t}] - \min_{i \in [K]} \sum_{t=1}^{T} l_{t,i}$, i.e., the player's expected cumulative loss in excess of that of the best single action in hindsight, where the expectation is taken over the player's actions.

Notation. For all $n \in \mathbb{N}$, define $[n] = \{1, \ldots, n\}$. For all $z \in \mathbb{Z}$, let $\mathbb{Z}_{\geqslant z} = \mathbb{Z} \cap [z, \infty)$. For each $i \in [K]$, let $\boldsymbol{e}_i \in \mathbb{R}^K$ denote the standard basis vector, where $(\boldsymbol{e}_i)_j = \mathbb{I}(i = j)$ for all $j \in [K]$. Let $\boldsymbol{0}_K, \boldsymbol{1}_K \in \mathbb{R}^K$ be the zero and one vectors, i.e., $(\boldsymbol{0}_K)_j = 0$ and $(\boldsymbol{1}_K)_j = 1$ for all $j \in [K]$. Define the probability simplex over $[K]$ as $\Delta([K]) = \{x \in \mathbb{R}^K_+ : \|x\|_1 = 1\}$.

1.2. Our Contributions

Our key innovation is a capacity-efficient approach to tracking delayed rounds, without compromising regret performance. While prior work in Delayed Online Learning requires tracking all delayed rounds (i.e., the utilized capacity, C_{util}, adjusts to demand and can be arbitrarily large), we introduce selective sampling policies that maintain at most a constant-sized set of rounds to track when delays are fixed (i.e., all $d_t = d$), and at most a logarithmic-sized (in T) set of rounds, when delays are round-dependent (i.e., d_t can be arbitrary), achieving a significant reduction in tracking-set size, without degrading regret guarantees. This addresses the fundamental question of the optimal capacity, C_{opt}, sufficient to match the asymptotic regret of Delayed Online Learning. We analyze the Delay Scheduling problem across various settings characterized by three dimensions: delay structure (fixed or round-dependent), delay knowledge at action time (clairvoyant or non-clairvoyant), and scheduling flexibility (preemptive or non-preemptive), considering bandit and full-information feedback regimes.

As is standard in Delayed Online Learning, our regret bounds depend on the number of actions K, the time horizon T, and the total delay $D = \sum_{t=1}^{T} d_t$. In this work, we additionally study the dependence of regret on the capacity C. Another important quantity we consider is the number of outstanding delays, $\sigma_t = \sum_{s=1}^{t-1} \mathbb{I}(s + d_s \geqslant t)$ for each round t. Letting $\sigma_{\max} = \max_t \sigma_t$ and $d_{\max} = \max_t d_t$, we note that a capacity of order $\Omega(\sigma_{\max})$ is sufficient to observe feedback from every round. While $\sigma_{\max}$ can be as large as $\Omega(\min\{\sqrt{D}, d_{\max}\})$, we show that, in most cases, capacity of this order is unnecessary

Delay Scheduling with Fixed Delays. We first consider fixed delays, introduced in the bandit setting by Cesa-Bianchi et al. (2016) and in the full-information setting by Weinberger and Ordentlich (2002). Here, all delays are equal, i.e., $d_t = d$, and known in advance, naturally corresponding to the clairvoyant framework. We study both preemptive and non-preemptive versions of this setting and show that there is no benefit in allowing preemption: our lower bound holds for the preemptive case, and our upper bound algorithm applies to both frameworks. We determine the minimax expected regret in both bandit and full-information feedback regimes (Table 1).

Delayed Online Learning with fixed delays			
Regime	Regret Bounds	Utilized capacities	Reference
Bandit	$\Theta\left(\sqrt{TK} + \sqrt{Td\log(K)}\right)$	$C_{\text{util}} = \Theta(d)$	2016; 2020
Full-info	$\Theta\left(\sqrt{T(d+1)\log(K)}\right)$	$C_{\text{util}} = \Theta(d)$	2002
Delay Scheduling with fixed delays			
Regime	Regret Bounds	Optimal capacities	Reference
Bandit	$\Theta\left(\sqrt{TK(1 + d/C)} + \sqrt{Td\log(K)}\right)$	$C_{\text{opt}} = \Theta\left(\min\{\frac{K}{\log(K)}, d\}\right)$	Theorem 3
Full-info	$\Theta\left(\sqrt{T(d+1)\log(K)}\right)$	$C_{\text{opt}} = \Theta(1)$	

Table 1 Minimax regret bounds for Delay Scheduling compared to Delayed Online Learning under the assumption of fixed delays.

While $C = d + 1$ is the exact capacity required to observe feedback from every round under fixed delays, we establish that capacities of $C = \Omega(\min\{K/\log(K), d\})$ and $C = \Omega(1)$ are both sufficient and necessary to eliminate the impact of the capacity constraint in the bandit and full-information settings, respectively. This stands in contrast to previously studied delayed algorithms, which implicitly required a capacity of $\Omega(d)$.

Delay Scheduling: Clairvoyant and Non-Preemptive. In this setting, the player observes the delay d_t at the start of each round t and must decide whether to track it, with no option to preempt once committed. When $C = \Omega(\log T)$, we establish minimax-optimal upper bounds for this setting (up to logarithmic factors), matching the fixed-delay case with $D = Td$ (Table 2).

Delayed Online Learning with round-dependent delays			
Regime	Regret Bounds	Utilized capacities	Reference
Bandit	$\Theta\left(\sqrt{TK} + \sqrt{D\log(K)}\right)$	$C_{\text{util}} = \Theta(\sigma_{\max})$	2016; 2020
Full-info	$\Theta\left(\sqrt{(D+T)\log(K)}\right)$	$C_{\text{util}} = \Theta(\sigma_{\max})$	2002; 2016
Clairvoyant Non-Preemptive Delay Scheduling with round-dependent delays			
Regime	Regret Bounds	Optimal capacities	Reference
Bandit	$O\left(\sqrt{TK + \frac{\log(T)}{C}(D+T)K} + \sqrt{D\log(K)}\right)$	$C_{\text{opt}} = O\left(\log(T) \cdot \min\{\frac{K}{\log(K)}, \frac{D}{T}+1\}\right)$	Theorem 10
	$\Omega\left(\sqrt{TK + DK/C} + \sqrt{D\log(K)}\right)$	$C_{\text{opt}} = \Omega\left(\min\{\frac{K}{\log(K)}, \frac{D}{T}\}\right)$	Theorem 3
Full-info	$O\left(\sqrt{(1 + \frac{\log(T)}{C})(D+T)\log(K)}\right)$	$C_{\text{opt}} = O(\log(T))$	Theorem 10
	$\Omega\left(\sqrt{(D+T)\log(K)}\right)$	$C_{\text{opt}} = \Omega(1)$	Theorem 3

Table 2 Minimax regret bounds for Delay Scheduling, assuming $C = \Omega(\log(T))$ for upper bounds, compared to Delayed Online Learning.

While $C = \sigma_{\max} + 1$ is the exact capacity required to observe feedback from every round under round-dependent delays, we establish that capacities $C = \Omega(K\log(T)/\log(K))$, and $C = \Omega(\log(T))$ are sufficient to avoid the impact of the capacity constraint in the bandit and full-information settings, respectively. In general, these capacity requirements are significantly smaller than $\Theta(\sigma_{\max})$, which can range from $\Omega(D/T)$ to $O(\sqrt{D})$ for round-dependent delays.

The results in Table 2 are derived from a more general bound that holds for any $C \geqslant 1$; however, to achieve this bound our algorithm requires prior knowledge of the magnitude of $T^{1/C}$ in order to set its parameters (see Table 6). In particular, when $C = \Omega(\log(T))$, we have $T^{1/C} = O(1)$.

Delay Scheduling: Non-Clairvoyant and Preemptive. In this setting, the player can preempt rounds, but delays remain hidden at action times. The player observes each delay only for as long as it stays in the tracking set up to the current time. This is more restrictive than standard Delayed Online Learning (without clairvoyance), where all delays are continuously observed up to the current time. Without prior knowledge of T and D, but assuming that an upper bound on the maximum delay, $d_{\max}$, is known, we establish bounds identical to those in the Clairvoyant Non-Preemptive setting, up to an additional $\widetilde{O}(d_{\max})$ term. Specifically, when $C = \Omega(\log(T))$, we establish the regret bounds in Table 3. As in the Clairvoyant Non-Preemptive framework, a more general bound exists that requires prior knowledge of the magnitude of $T^{1/C}$, up to an additional $\widetilde{O}(d_{\max})$ term.

Assuming prior knowledge of D, preempting rounds with delays exceeding $\sqrt{D}$ removes dependence on $d_{\max}$, matching the Clairvoyant Non-Preemptive regret bound in Table 2. Prior work has explored various adaptive "skipping schemes" to mitigate the impact of highly unbalanced

Non-Clairvoyant Preemptive Delay Scheduling for $C = \Omega(\log(T))$ with known $d_{\max}$		
Regime	Regret Bounds	Reference
Bandit	$O\left(\sqrt{TK + \frac{\log(T)}{C}(D+T)K} + \sqrt{D\log(K)}\right) + \tilde{O}\left(d_{\max}\sqrt{1 + \frac{K}{C}}\right)$	Theorem 11
Full-info	$O\left(\sqrt{(1 + \frac{\log(T)}{C})(D+T)\log(K)}\right) + \tilde{O}(d_{\max})$	

Table 3   Regret upper bounds for Non-Clairvoyant Preemptive Delay Scheduling with round-dependent delays when $C = \Omega(\log(T))$, assuming prior knowledge of $d_{\max}$.

delays, treating skipped rounds as contributing at most 1 to regret while ignoring their delays. For example, such adaptive techniques may optimize regret by selecting the optimal skipping threshold (e.g., Thune et al. (2019) under clairvoyance) or by choosing the best subset of rounds to skip (e.g., Zimmert and Seldin (2020)). However, Non-Clairvoyant Preemptive framework imposes strict constraints on observing information about delays, preventing direct application of these techniques.

Delay Scheduling: Non-Clairvoyant and Non-Preemptive. For completeness, we also consider the most restrictive setting, where the player has to commit to tracking rounds without the ability to preempt and without any clairvoyant knowledge of delays. Assuming prior knowledge of T and D, we are able to achieve sublinear regret in both feedback regimes (Table 4) for all $C \geqslant 1$. Thus, given

Non-Clairvoyant Non-Preemptive Delay Scheduling with known T, D		
Regime	Regret Bounds	Optimal capacities
Bandit	$O\left(\sqrt[3]{\frac{T(D+T)K}{C}} + \sqrt{TK + D\log(K)}\right)$	$C_{\text{opt}} = O\left(\frac{K}{\log(K)} \cdot \frac{T}{\sqrt{(D+T)\log(K)}}\right)$
Full-info	$O\left(\sqrt[3]{\frac{T(D+T)\log(K)}{C}} + \sqrt{(D+T)\log(K)}\right)$	$C_{\text{opt}} = O\left(\frac{T}{\sqrt{(D+T)\log(K)}}\right)$

Table 4   Regret upper bounds for Non-Clairvoyant Non-Preemptive Delay Scheduling when capacity $C \geqslant 1$, assuming prior knowledge of T and D. Derived from Corollary 33 in Appendix H.

prior knowledge of T and D, a capacity of order $\Omega(\sqrt{T})$ with respect to T is always sufficient to avoid the effects of limited resources. Furthermore, as the total delay D approaches T^2, the upper bound on the optimal capacity decreases, reaching $O(1)$ for $D = \Omega(T^2)$[2]. Alternatively, when only an upper bound on the maximum delay, $d_{\max}$, is available, we establish different regret bounds (Table 5). In this case, ensuring sublinear regret may require C to grow polynomially with T when $d_{\max} = \Omega(T)$. For proofs and additional details about the setting, see Appendix H.

Non-Clairvoyant Non-Preemptive Delay Scheduling with known $d_{\max}$		
Regime	Regret Bounds	Optimal capacities
Bandit	$O\left(\sqrt{\frac{Td_{\max}K}{C}} + \sqrt{TK + D\log(K)}\right)$	$C_{\text{opt}} = O\left(\min\left\{\frac{K}{\log(K)} \cdot \frac{Td_{\max}}{D+T}, d_{\max}\right\}\right)$
Full-info	$O\left(\sqrt{\frac{Td_{\max}\log(K)}{C}} + \sqrt{(D+T)\log(K)}\right)$	$C_{\text{opt}} = O\left(\frac{Td_{\max}}{D+T}\right)$

Table 5   Regret upper bounds for Non-Clairvoyant Non-Preemptive Delay Scheduling when capacity $C \geqslant 1$, assuming prior knowledge of $d_{\max}$. Derived from Corollary 34 in Appendix H.

---

2. This may seem counterintuitive; however, when D is of the order T^2, linear regret becomes unavoidable in Delayed Online Learning, indicating that the optimal capacity should be of the smallest order $O(1)$.

Delay Scheduling under the expectation-capacity constraint. As an alternative approach to Delay Scheduling, we consider a setting where only the *expected size* of the tracking set is required to remain bounded by the expectation-capacity $C_E \in (0, \infty)$ at each round. We refer to this constraint as the "expectation-capacity constraint" and explore it further in Appendix I. Notably, we establish minimax bounds on achievable regret for all values of C_E (up to logarithmic factors). With prior knowledge of $\log(T)$ up to constant multiplicative factors, our algorithms for the standard capacity constraint can be adapted to the expectation-capacity setting, achieving similar regret bounds formula-wise as in Tables 2 and 3, but with C_E replacing C in the bounds and without assuming $C_E = \Omega(\log(T))$ (see Table 8). We prove matching lower bounds in Theorem 35 of Appendix I, completing the theoretical characterization of the problem.

1.3. Technique Highlights

Our paper introduces several key technical advances, listed here in the order they appear in the text.

The core learning component in our algorithms is an FTRL-based framework for delayed online learning. We extend the Delayed FTRL algorithm of Zimmert and Seldin (2020) to accommodate loss scales that vary between rounds (Section 2). When applied to Delay Scheduling, these scales reflect the weighting of losses with respect to probabilities of observing them.

We then introduce several scheduling techniques integrated with the learning algorithm. In Section 3, we present a natural Batch Partitioning method (Algorithm 1), which achieves minimax regret in Delay Scheduling under fixed delays, with matching lower bounds established via a reduction from Label-Efficient and Delayed Online Learning. This lower bound extends to round-dependent delays, which we match up to logarithmic factors in the following sections.

Section 4 introduces schedulers as autonomous subroutines, thereby externalizing scheduling from learning. Within this framework, we establish general regret bounds in Theorem 7 for a specific class of *precommitted* schedulers paired with Delayed FTRL (Algorithm 2). For this class of precommitted schedulers, we introduce preemptive and non-preemptive variants (Schedulers 3 and 4), with corresponding regret bounds established in Theorems 10 and 11. Notably, the preemptive Scheduler 4 introduces a novel technique of sampling *proxy delays* to balance the trade-off between observing long-delay feedback and controlling the size of the tracking set.

2. Delayed Follow the Regularized Leader (FTRL) with Time-Varying Loss Scales

As the first algorithm achieving minimax regret for oblivious bandits with round-dependent delays, the Delayed FTRL algorithm by Zimmert and Seldin (2020) uses a hybrid, time-varying regularizer F_t, where each action A_t is sampled according to $x_t = \operatorname{argmin}_{x \in \Delta([K])} \langle x, \hat{L}_t^{\text{obs}} \rangle + F_t(x)$. Here, $\hat{L}_t^{\text{obs}}$ denotes the cumulative estimator of all previously observed loss vectors up to the start of round t, and regularizer $F_t(x) = \alpha_t^{-1} F_{\text{Ts}}(x) + \beta_t^{-1} F_{\text{NE}}(x)$ is a weighted sum of $\frac{1}{2}$-Tsallis entropy $F_{\text{Ts}}(x) = -\sum_{i=1}^{K} 2x_i^{1/2}$ and negative entropy $F_{\text{NE}}(x) = \sum_{i=1}^{K} x_i \log(x_i)$, with separate learning rates α_t and β_t for each part.[3]

We extend this Delayed FTRL to a setting where loss scales $(B_t)_{t=1}^{T} \in [0, \infty)$ vary across rounds, with an oblivious adversary selecting losses $l_t \in [0, B_t]^K$ and delays $d_t \in \mathbb{Z}_{\geq 0}$ before the

---

3. For the full-information regime, we disable the Tsallis component by setting $\alpha_t = \infty$.

game begins.[4] This framework serves as the foundation for reductions from other algorithms in the following sections, providing a unified approach to bounding regret across various settings. For each round t, let $\mathcal{W}_t = \{s \in [t-1] : s + d_s \geqslant t\}$ denote the *working set* of rounds with pending feedback.

Theorem 1 *Consider Delayed FTRL with time-varying loss scales $l_t \in [0, B_t]^K$ (formally, Algorithm 5) running with arbitrary non-increasing sequences of learning rates $(\alpha_t)_{t=1}^T$ and $(\beta_t)_{t=1}^T$. Then, the regret in the bandit regime satisfies:*

$$\mathfrak{R}_T \leqslant \sum_{t=1}^T \left(\sqrt{K} \alpha_t B_t^2 + \beta_t B_t \sum_{s \in \mathcal{W}_t} B_s \right) + 2\sqrt{K} \alpha_T^{-1} + \log(K) \beta_T^{-1}.$$

And in the full-information regime ($\alpha_t = \infty$):

$$\mathfrak{R}_T \leqslant \sum_{t=1}^T \left(\beta_t B_t^2 + \beta_t B_t \sum_{s \in \mathcal{W}_t} B_s \right) + \log(K) \beta_T^{-1}.$$

We prove Theorem 1 in Appendix C by expanding the proof of Theorem 3 in Zimmert and Seldin (2020) to handle time-varying loss scales in both bandit and full-information regimes.

3. Batch Partitioning Algorithm with a Notable Application for Fixed Delays

A natural approach to managing the tracking set under the capacity constraint is to partition rounds into contiguous batches of equal size and track a single uniformly selected *representative* round per batch. If batch size b is sufficiently large (e.g., $b \geqslant \frac{d_{\max}}{C-1}$), then capacity C is never exceeded.

The batching technique has been explored in various online learning settings (e.g., Dekel et al. (2012); Arora et al. (2012)). In the presence of delayed feedback, batching is particularly effective, as it decouples the impact of one-per-batch feedback observation from the effect of delays on the regret. A similar decoupling phenomenon was independently leveraged in recent work by Wan et al. (2024) to obtain improved regret bounds for bandit convex optimization. We introduce a novel use of batching as a delay scheduling mechanism to satisfy the capacity constraint throughout the game.

In Algorithm 1, we run Delayed FTRL at the batch level: selecting one action per batch to be used in all its rounds and updating the player's decision rule using aggregated loss estimates from the observed representative rounds. Notably, this algorithm can be run in the most restrictive non-clairvoyant and non-preemptive framework.

The following batch-level notation arises naturally. The number of batches is $T' = \lceil \frac{T}{b} \rceil$. Each batch[5] $\tau \in [T']$ has a representative u_τ, batch loss $l_\tau^b = l_{u_\tau}$, batch delay $d_\tau^b = \lceil \frac{u_\tau + d_{u_\tau}}{b} \rceil - \lceil \frac{u_\tau}{b} \rceil$, number of outstanding batch delays $\sigma_\tau^b = |\{s < \tau : s + d_s^b \geqslant \tau\}|$, and total delay $D_\tau^b = \sum_{s \in [\tau]} \sigma_s^b$.

Theorem 2 *Let Algorithm 1 be run with batch size $b \geqslant \frac{d_{\max}}{C-1}$. Then, with learning rates $\alpha_\tau = \sqrt{1/\tau}, \beta_\tau = \sqrt{\frac{\log(K)}{D_\tau^b}}$, the regret in the bandit regime satisfies: $\mathfrak{R}_T \leqslant 14\sqrt{TbK} + 3\sqrt{D \log(K)}$. In the full-information regime, with $\beta_\tau = \sqrt{\frac{\log(K)}{\tau + D_\tau^b}}$: $\mathfrak{R}_T \leqslant 12\sqrt{Tb \log(K)} + 3\sqrt{D \log(K)}$.*

---

4. Unlike in the scale-free online learning literature (e.g., see Orabona and Pal (2016); Putta and Agrawal (2021)), our analysis is independent of how the player observes B_t, as we study this Delayed FTRL only with fixed, pre-determined sequences of learning rates.

5. The final batch is extended to b elements by padding with rounds of zero loss and delay.

Algorithm 1 Delay FTRL with Batch Partitioning

Input: Number of actions K and capacity C.

Parameters: Batch size b.

1. Initialize empty tracking set S of maximum size C. Initialize $\widehat{L}_1^{\text{obs}} = \mathbf{0}_K$.
2. **For** batch $\tau = 1, 2, ..., \lceil T/b \rceil$:
   - (a) Sample batch-representative $u_\tau \sim \text{Unif}\{(\tau - 1)b + 1, ..., \tau b\}$.
   - (b) Calculate learning rates α_τ, β_τ using available information and sample action $A_\tau^b \in [K]$ according to $x_\tau^b = \operatorname{argmin}_{x \in \Delta([K])} \langle x, \widehat{L}_\tau^{\text{obs}} \rangle + \alpha_\tau^{-1} F_{\text{Ts}}(x) + \beta_\tau^{-1} F_{\text{NE}}(x)$.
   - (c) **For** round $t = (\tau - 1)b + 1, ..., \min\{\tau b, T\}$:
     - Play action $A_t = A_\tau^b$. If round t is a representative u_τ, then add t to S.
     - For each expired $u_s \in S$ (i.e., $u_s + d_{u_s} = t$), observe feedback, and set estimator $\hat{l}_{u_s}$:
       - In the bandit regime, observe $(u_s, l_{u_s, A_{u_s}})$ and set $\hat{l}_{u_s} = l_{u_s, A_{u_s}} x_{u_s, A_{u_s}}^{-1} \boldsymbol{e}_{A_{u_s}}$.
       - In the full-information regime, observe (u_s, l_{u_s}) and set $\hat{l}_{u_s} = l_{u_s}$.
   - (d) Update $\widehat{L}_{\tau+1}^{\text{obs}} = \widehat{L}_\tau^{\text{obs}} + \sum_{s:(\tau-1)b < u_s + d_{u_s} \leqslant \tau b} \hat{l}_{u_s}$.

Theorem 3 *Consider Delay Scheduling with fixed delays $d_t = d$. Across all scheduling frameworks regardless of preemptibility and clairvoyance, the minimax regret is $\Theta(\sqrt{TK(1 + d/C)} + \sqrt{Td\log(K)})$ for the bandit regime and $\Theta(\sqrt{T(d + 1)\log(K)})$ for the full-information regime.*

The proofs of Theorems 2 and 3 are in Appendix D. For Theorem 2, the proof reduces the original problem to a batch-level game by conditioning on representative round selection. Upon conditioning, we obtain a T'-round game with losses l_τ^b and delays d_τ^b on which Algorithm 1 effectively runs the Delayed FTRL from Section 2, for which we apply Theorem 1 to obtain the stated bounds. The upper bound for Theorem 3 follows from Theorem 2 by setting $b = \max\{1, \lceil \frac{d}{C-1} \rceil\}$, while the lower bound uses reductions from classical label efficient and delayed online learning settings.

4. General Paradigm for Scheduling and Learning

We propose a general paradigm that separates the *scheduling policy*, which manages how rounds are tracked for delayed feedback, from the *learning algorithm*, which is responsible for action selection. This approach is inspired by the principle of separating high-level policies from low-level mechanisms Levin et al. (1975), a widely used concept in operating system design. In our framework, the scheduler operates autonomously, managing the tracking set and supplying the learning algorithm with observations, as captured by the following abstract interface:

Abstract Scheduler Interface

Input: Empty *tracking set* S of maximum size C.

For each round $t = 1, 2, \ldots, T$:
1. Decide (deterministically or randomly) whether to add current round t to S.
2. For each round $s \in S$ whose delay expires this round (i.e., $s + d_s = t$), observe feedback from round s, and send this feedback to the learning algorithm.
3. If the setting is preemptive, decide which rounds, if any, to remove from S.

We focus on randomized schedulers that operate autonomously from the learning algorithm, determining the observation schedule without relying on losses or actions. The following definitions formalize the aspects of the modularity that we will rely upon in our proofs.

4.1. Precommitted Schedulers and Observation-Independent Algorithms

Let S_t^0 denote the state of the tracking set immediately before round t, while S_t^1 denotes the state after the decision whether to include round t in the tracking set has been fully processed and before removing any elements. The observation indicator $Z_t = \mathbb{I}(t \in S_{t+d_t}^1)$ denotes whether feedback from t is observed at time $t + d_t$. See Table 7 for a summary of the notation.

Definition 4 *A scheduler $\mathcal{S}$ is **precommitted**, relative to a filtration $(\mathcal{F}_t^{\mathcal{S}})$, if there exists an i.i.d. sequence $(X_t^{\mathcal{S}})_{t=1}^T$ such that $\mathcal{F}_t^{\mathcal{S}} = \sigma(X_1^{\mathcal{S}}, \dots, X_{t-1}^{\mathcal{S}})$, each tracking set S_t^0 is $\mathcal{F}_t^{\mathcal{S}}$-measurable, and each observation indicator Z_t is $\mathcal{F}_{t+1}^{\mathcal{S}}$-measurable.*

Hence, the tracking set at round t is determined by randomness up to the start of round t, while feedback from round t is observed at round $t + d_t$ based on the scheduler's randomness up to the start of the next round.

Definition 5 *A precommitted scheduler $\mathcal{S}$ is **quantified** by a sequence of non-zero probabilities $(p_t)_{t=1}^T$ if, for all $t \in [T]$, the observation indicator Z_t satisfies $\mathbb{E}[Z_t \mid \mathcal{F}_t^{\mathcal{S}}] = p_t \mathbb{I}(|S_t^0| < C)$.*

Hence, conditional on the tracking set not being full at the start of round t, feedback from round t is observed with probability p_t.

Definition 6 *Let $\mathcal{A}$ be a delay scheduling algorithm and let $\mathcal{S}$ be its scheduler. We say $\mathcal{A}$ is **observation-independent** if $\mathcal{S}$ is precommitted relative to some filtration $(\mathcal{F}_t^{\mathcal{S}})$ and there exists an independent i.i.d. sequence $(X_t^{\mathcal{A}})_{t=1}^T$ such that the action A_t at round t is $\mathcal{F}_t^{\mathcal{A}}$-measurable, where $\mathcal{F}_t^{\mathcal{A}} = \sigma(\mathcal{F}_t; X_t^{\mathcal{A}})$ and $\mathcal{F}_t = \sigma(\mathcal{F}_t^{\mathcal{S}}; X_1^{\mathcal{A}}, \dots, X_{t-1}^{\mathcal{A}})$.*

Hence, the learner does not influence the scheduler, and the learner at round t only depends on the scheduler's randomness up to the start of that round, implying that feedback for round t is received at $t + d_t$ independently of A_t, i.e., $Z_t \perp A_t \mid \mathcal{F}_t$.

Algorithm 2 is a delayed scheduling algorithm where the learner is Delayed FTRL and the precommitted scheduler $\mathcal{S}$ is quantified by a sequence $(p_t)_{t=1}^T$, which determines the weights of observations in the loss estimators.

Theorem 7 *Consider Algorithm 2 as $\mathcal{A}$, where a Delayed FTRL algorithm is run with access to a precommitted scheduler $\mathcal{S}$, quantified by a sequence $(p_t)_{t=1}^T$. If the learning rates α_t and β_t are non-increasing and $\mathcal{F}_t^{\mathcal{S}}$-measurable, then $\mathcal{A}$ is an observation-independent delay scheduling algorithm whose regret, in the bandit regime, satisfies:*

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\sqrt{K} \alpha_t \frac{Z_t}{p_t^2} + \beta_t \frac{Z_t}{p_t} \sum_{s \in \mathcal{W}_t} \frac{Z_s}{p_s} \right) + 2\sqrt{K}\alpha_T^{-1} + \log(K)\beta_T^{-1} \right] + \sum_{t=1}^T \mathbb{P}(|S_t^0| = C),$$

and in the full-information regime:

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\beta_t \frac{Z_t}{p_t^2} + \beta_t \frac{Z_t}{p_t} \sum_{s \in \mathcal{W}_t} \frac{Z_s}{p_s} \right) + \log(K)\beta_T^{-1} \right] + \sum_{t=1}^T \mathbb{P}(|S_t^0| = C).$$

Algorithm 2 Delayed FTRL with access to a precommitted scheduler $\mathcal{S}$

Input: Number of arms K, Capacity C.

Access: Precommitted scheduler $\mathcal{S}$ quantified by a sequence $(p_t)_{t=1}^T$

1. Initialize scheduler $\mathcal{S}$ together with the *tracking set* S of maximum size C. Initialize $\widehat{L}_1^{\text{obs}} = \mathbf{0}_K$.

2. **For** round $t = 1, 2, \ldots, T$:

    (a) Calculate learning rates α_t, β_t using available information. Draw an action A_t according to $x_t = \mathrm{argmin}_{x \in \Delta([K])} \langle x, \widehat{L}_t^{\text{obs}} \rangle + \alpha_t^{-1} F_{\text{Ts}}(x) + \beta_t^{-1} F_{\text{NE}}(x)$ and play it.

    (b) Scheduler $\mathcal{S}$ makes a decision whether to start tracking round t or not.

    (c) For each round $s \in S$ whose delay expires this round (i.e., $s + d_s = t$), observe feedback from round s (via scheduler $\mathcal{S}$), calculate probability p_s, and construct loss estimator $\hat{l}_s$:

        • In the bandit regime, observe (s, l_{s,A_s}), and set $\hat{l}_s = l_{s,A_s} x_{s,A_s}^{-1} \boldsymbol{e}_{A_s} \cdot p_s^{-1}$.

        • In the full-information regime, observe (s, l_s), and set $\hat{l}_s = l_s \cdot p_s^{-1}$.

    (d) Update $\widehat{L}_{t+1}^{\text{obs}} = \widehat{L}_t^{\text{obs}} + \sum_{\text{observed } s : s+d_s=t} \hat{l}_s$.

    (e) Scheduler $\mathcal{S}$ makes preemption decisions.

The proof of Theorem 7 is in Appendix E. We show that $\mathcal{A}$ is observation-independent by constructing the sequence $(X_t^{\mathcal{A}})_{t=1}^T$ via induction. In the regret analysis, we forfeit rounds where $|S_t^0| = C$, incurring regret of 1 per round, and use the fact that the scheduler and algorithm are precommitted and observation-independent respectively in order to condition on the scheduler's randomness, reducing the analysis to Delayed FTRL with loss scales $B_t = \frac{Z_t}{p_t}$, for which we apply Theorem 1.

4.2. Precommitted Non-Preemptive Bernoulli Scheduler

From a practical standpoint, non-preemptive schedulers are fairly straightforward as they only have to decide whether to start tracking a given round without managing preemption decisions. In that vein, the non-preemptive precommitted scheduler quantified by a sequence $(p_t)_{t=1}^T$ can only be implemented as the Bernoulli scheduler (Scheduler 3). This scheduler, independently decides at each round t whether to track that round with probability p_t if the tracking set is not full.

Scheduler 3 Non-Preemptive Bernoulli Scheduler

Inputs: Empty tracking set S of maximum size C.

For round $t = 1, 2, \ldots, T$:

    1. Calculate p_t using available information. If $|S| < C$, then add t to S with probability p_t.

    2. For each expired $s \in S$ (i.e., $s + d_s = t$), observe its feedback, and pass it to the learner.

For Scheduler 3, p_t only needs to be deterministically computable[6] at each round t. By carefully selecting these probabilities, we can ensure representative feedback is captured throughout the game while keeping the probability of reaching the full capacity C at any round under control.

Under clairvoyance, one viable sequence class is given by $p_t = \min\{1, \frac{C}{(1+\alpha)\nu_t} \cdot \frac{1}{d_t+1}\}$, governed by two components: a normalizer sequence $(\nu_t)_{t=1}^T$ and a tunable Chernoff parameter $\alpha > 0$. The sequence ν_t helps to control the expected tracking set size $\mathbb{E}[|S_t^0|]$, while a sufficiently large α

---

6. "Deterministically computable" means that the value remains the same in every run, computed without any randomness.

ensures a high probability bound on $|S_t^0|$. We set $\nu_t = 2H_t$, where $H_t = \sum_{t=1}^{T} \frac{1}{t}$ is a harmonic number. The value of α is taken large enough to satisfy the following "Overflow condition":

$$\ln(1+\alpha) - \frac{\alpha}{1+\alpha} \geqslant \frac{\ln(\delta^{-1})}{C}, \tag{$\star$}$$

for some overflow probability $\delta \in (0,1)$. These choices are used in Lemma 8 and later Lemma 9.

For this choice of p_t, Lemma 8 provides a high-probability guarantee that the tracking set is not full at any round. The proof, given in Appendix F, follows from a multiplicative Chernoff bound.

Lemma 8 (Bernoulli Scheduler: Capacity Control) *Let $\delta \in (0,1)$. Suppose Chernoff parameter $\alpha > 0$ satisfies ($\star$), then Scheduler 3 with probabilities $p_t = \min\{1, \frac{C}{2(1+\alpha)H_t} \cdot \frac{1}{d_t+1}\}$ guarantees that $\mathbb{P}(|S_t^0| = C) \leqslant \delta$ for every $t \in [T]$.*

4.3. Precommitted Preemptive Scheduling via Proxy Delays

We introduce **proxy delays**, a novel approach to preemptive scheduling in which, at the beginning of each round t, the scheduler selects a proxy delay $\widetilde{d}_t \in \mathbb{Z}_{\geqslant -1}$ independently of previous rounds, determining how long round t will be tracked. Thus, if the tracking set is not full and $\widetilde{d}_t \geqslant 0$[7], round t remains in S until the end of round $t + \min\{d_t, \widetilde{d}_t\}$. This ensures tracking durations are determined at round t, keeping the scheduler precommitted, as $Z_t = \mathbb{I}(|S_t^0| < C, \widetilde{d}_t \geqslant d_t)$. Proxy delays are effective in non-clairvoyant settings, as they can be sampled without knowledge of d_t.

A natural choice for the proxy delay distribution is the *Pareto distribution*, whose heavy tail ensures long delays are observed with non-negligible probability, while its inverse-polynomial decay helps control capacity. Formally, for the *scale* and *shape* parameters $c, \beta > 0$, the CDF is defined as $F(x) = \mathbb{I}(x \geqslant c)(1 - (\frac{c}{x})^\beta)$. Setting the shape parameter $\beta = 1$ ensures that the probability of observing feedback from a round with delay d is proportional to $1/d$, aligning with the optimal sampling rate for fixed delays in Theorem 3.

The full scheduling policy using Pareto-distributed proxy delays is given in Scheduler 4. Here, the scheduler samples $\widetilde{d}_t$ from the distribution $\mathfrak{D}_t = \lfloor \mathrm{Pareto}(\frac{C}{(1+\alpha)\nu_t}, 1) - 1 \rfloor$, where the normalizer sequence $\nu_t = 2H_t$ and Chernoff parameter α are defined analogously to those in Subsection 4.2. This parameter choice ensures that Scheduler 4 is quantified by the same probability sequence as Scheduler 3 in Lemma 8, whose preemptive analogue is covered in Lemma 9 below.

Scheduler 4 Preemptive Scheduler with Pareto Proxy Delays

Inputs : Empty tracking set S of maximum size C.
Parameters : Chernoff parameter α.
For round $t = 1, 2, \ldots, T$:
    1. Sample proxy delay $\widetilde{d}_t \sim \mathfrak{D}_t$. Add round t to S if $|S| < C$ and $\widetilde{d}_t \geqslant 0$.
    2. For each expired $s \in S$ (i.e., $s + d_s = t$), observe its feedback, and pass it to the learner.
    3. For each $s \in S$ whose proxy delay expires this round (i.e., $s + \widetilde{d}_s = t$), remove s from S.

Lemma 9 (Proxy Delay Scheduler: Observation Probability and Capacity Control) *Let $\delta \in (0,1)$. If the Chernoff parameter α satisfies ($\star$), then Scheduler 4 ensures that $\mathbb{E}[Z_t \mid |S_t^0| < C] = \mathbb{P}(\widetilde{d}_t \geqslant d_t) = \min\{1, \frac{C}{2(1+\alpha)H_t} \cdot \frac{1}{d_t+1}\}$ and $\mathbb{P}(|S_t^0| = C) \leqslant \delta$ for every $t \in [T]$.*

---

7. A proxy delay of $\widetilde{d}_t = -1$ indicates that round t is never added to the tracking set S.

By construction, Scheduler 4 ensures that $Z_t = \mathbb{I}(|S_t^0| < C, \widetilde{d}_t \geqslant d_t)$, yielding the first result by evaluating tails of $\mathfrak{D}_t$. To analyze the capacity constraint, we let $\widetilde{\sigma}_t = \sum_{s=1}^{t-1} \mathbb{I}(s + \widetilde{d}_s \geqslant t)$ denote the number of outstanding proxy delays at round t, which bounds the tracking set size at the start of round t. The normalizer sequence ν_t controls $\mathbb{E}[\widetilde{\sigma}_t]$, while condition ($\star$) guarantees that α is large enough to obtain a strong multiplicative Chernoff bound on $\mathbb{P}(\widetilde{\sigma}_t \geqslant C)$. The proof of Lemma 9 is in Appendix F, where we also include an illustration (Figure 2) of how the tracking set evolves.

5. Upper Bounds for Clairvoyant or Preemptive Settings

In this section, we apply the techniques developed in Section 4 to conclude the regret bounds for Clairvoyant Non-Preemptive and Non-Clairvoyant Preemptive settings. In order to prove the results below, we apply Theorem 7 (Delayed FTRL with precommitted schedulers) with Scheduler 3 for Theorem 10, and with Scheduler 4 for Theorem 11, and choosing learning rates that are $\mathcal{F}_t^{\mathcal{S}}$-measurable and computable from the information available at each round t. (See Appendix G for proofs.) We use $\mu_t = (\mathbb{P}(\widetilde{d}_t \geqslant d_t))^{-1} = \max\{1, \frac{(1+\alpha)\nu_t}{C} \cdot (d_t + 1)\}$ and $z_t = Z_t \mu_t$ for tuning the learning rates in the following corollaries.

Theorem 10 (Clairvoyant and Non-Preemptive) *Let $\mathcal{S}$ be Scheduler 3 with probabilities $p_t = \min\{1, \frac{C}{(1+\alpha)\nu_t} \cdot \frac{1}{d_t+1}\}$ such that α satisfies ($\star$) for some $\delta \in (0,1)$. Then, in the bandit regime, Algorithm 2 with access to $\mathcal{S}$ has the following expected regret bound, when run with learning rates $\alpha_t = \sqrt{\frac{1}{\sum_{s \in [t]} \mu_s}}, \beta_t = \sqrt{\frac{\log(K)}{\sum_{s \in [t]} d_s}}$: $\mathfrak{R}_T \leqslant 4\sqrt{K}\sqrt{T + \frac{(1+\alpha)\nu_T}{C}(D+T)} + 3\sqrt{D\log(K)} + \delta T$.*

In the full-info regime, with $\beta_t = \sqrt{\frac{\log(K)}{\sum_{s \in [t]}(\mu_s + d_s)}}$: $\mathfrak{R}_T \leqslant 3\sqrt{\log(K)}\sqrt{(D+T)(1 + \frac{(1+\alpha)\nu_T}{C})} + \delta T$.

For each round t, let $O_t = \{s \in [t-1] : s + d_s < t\}$ denote the *observation set* of rounds whose feedback might be available. In Delay Scheduling, the player only observes its subset $\widetilde{O}_t = \{s \in O_t : Z_s = 1\}$ by the start of round t. Note that for all $s \in O_t \backslash \widetilde{O}_t$, $Z_s = z_s = 0$.

Theorem 11 (Non-Clairvoyant and Preemptive with known $d_{\max}$) *Let $\mu_{\max,t} = \max\{1, \frac{(1+\alpha)\nu_t}{C} \cdot (d_{\max} + 1)\}$ and $\mu_{\max} = \mu_{\max,T}$. Let $\mathcal{S}$ be Scheduler 4 with parameter α satisfying ($\star$) for some $\delta \in (0,1)$. Then, in the bandit regime, Algorithm 2 with access to $\mathcal{S}$ has the following regret bound, when run with learning rates $\alpha_t = \sqrt{\frac{1}{\sum_{s \in \widetilde{O}_t} z_s^2 + C\mu_{\max,t}^2}}, \beta_t = \sqrt{\frac{\log(K)}{\sum_{s \in \widetilde{O}_t} z_s d_s + C\mu_{\max,t} d_{\max}}}$:*

$$\mathfrak{R}_T \leqslant 4\sqrt{K}\sqrt{T + \frac{(1+\alpha)\nu_T}{C}(D+T)} + 3\sqrt{D\log(K)} + 7\sqrt{C\mu_{\max}(K\mu_{\max} + \log(K)d_{\max})} + \delta T.$$

And in the full-information regime, with learning rate $\beta_t = \sqrt{\frac{\log(K)}{\sum_{s \in \widetilde{O}_t} z_s(z_s + d_s) + C\mu_{\max,t}(\mu_{\max,t} + d_{\max})}}$:

$$\mathfrak{R}_T \leqslant 3\sqrt{\log(K)}\sqrt{(D+T)(1 + \frac{(1+\alpha)\nu_T}{C})} + 3\sqrt{C\mu_{\max}\log(K)(\mu_{\max} + d_{\max})} + \delta T.$$

Parameter Choices and Result Summary: To provide context for these results, we need to set parameter α large enough so that ($\star$) is satisfied for some $\delta \in (0,1)$. Assuming $C \geqslant 3\log(T)$, we can set $\alpha = 1$ without prior knowledge of T, ensuring that ($\star$) holds for $\delta = T^{-0.5}$. In this case, Theorems 10 and 11 yield the results presented in Tables 2 and 3, respectively.

More generally, for any $C \geqslant 1$, if we were to set $\alpha = eT^{0.5/C} - 1$ and consider $\delta = T^{-0.5}$, then we would have the regret bounds summarized in Table 6.

Framework	Regime	Regret Bounds
Clairvoyant Non-Preemptive	Bandit	$O\left(\sqrt{TK + \frac{T^{0.5/C}\log(T)}{C}(D+T)K} + \sqrt{D\log(K)}\right)$
	Full-info	$O\left(\sqrt{(1 + \frac{T^{0.5/C}\log(T)}{C})(D+T)\log(K)}\right)$
Non-Clairvoyant Preemptive	Bandit	$O\left(\sqrt{TK + \frac{T^{0.5/C}\log(T)}{C}(D+T)K} + \sqrt{D\log(K)}\right) + \widetilde{O}\left(d_{\max}\sqrt{T^{0.5/C} + \frac{T^{1/C}K}{C}}\right)$
	Full-info	$O\left(\sqrt{(1 + \frac{T^{0.5/C}\log(T)}{C})(D+T)\log(K)}\right) + \widetilde{O}\left(d_{\max}\sqrt{T^{0.5/C} + \frac{T^{1/C}}{C}}\right)$

Table 6   Regret upper bounds derived from Theorems 10 and 11, when $\alpha = eT^{0.5/C} - 1$.

When $C = \Omega(\log T)$, we have $T^{1/C} = O(1)$ and the bounds match those in Tables 2 and 3.

6. Discussion and Future Work

We introduce the Delay Scheduling setting, a novel framework for online learning with delayed feedback under a capacity constraint. Our analysis spans various settings characterized by distinct delay structures, clairvoyance, and preemptibility, covering both bandit and full-information feedback regimes. A key finding reveals that, in many cases, remarkably modest capacity suffices to achieve regret comparable to that of unconstrained Delayed Online Learning. Building on these results, we identify several critical directions for future inquiry.

A key open question in our work is determining the minimax regret when $C = O(\log T)$. Although we establish minimax regret bounds for $C = \Omega(\log T)$ under clairvoyance or preemptibility, the precise dependence of regret on C in the small-capacity regime remains unclear. Tightening these bounds would refine our understanding of the fundamental capacity requirements for efficient learning. Moreover, designing algorithms for Non-Clairvoyant, Non-Preemptive Delay Scheduling that require no prior knowledge of T, D, or $d_{\max}$ represents a significant challenge. Additionally, since existing lower bounds are derived via reductions from Delay Scheduling with fixed delays, establishing setting-specific lower bounds remains an important open problem.

Recent work has explored delayed bandits under various assumptions that bring the setting closer to real-world applications. It is promising to extend these frameworks by incorporating the capacity constraint introduced in this paper. For instance, Delay Scheduling with action-dependent delays, where delay duration varies based on chosen actions, as in Van Der Hoeven and Cesa-Bianchi (2022), represents a promising research direction. Such action-dependent delays naturally occur in applications like dynamic pricing and medical trials, necessitating novel algorithmic approaches, particularly in non-clairvoyant settings. Investigating Delay Scheduling with scale-free losses, where delay correlates with the incurred loss, as in Huang et al. (2023), offers another exciting extension. If longer delays correspond to systematically different loss distributions, dynamically adjusting sampling rates could enhance performance. Finally, contextual bandits with delayed feedback, as in Erez et al. (2024), considered under the capacity constraint, present an important open direction for future work. In applications such as personalized recommendations, where context-action pairs must be tracked, capacity constraint introduces new challenges for exploration-exploitation trade-offs.

Overall, our results establish a foundation for learning with capacity-constrained delayed feedback, suggesting many promising directions for future research.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1747–1754, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. No weighted-regret learning in adversarial bandits with delays, 2022.

Allan Borodin and Ran El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 1998.

Allan Borodin and Christodoulos Karavasilis. Any-order online interval selection. In *Approximation and Online Algorithms: 21st International Workshop, WAOA 2023, Amsterdam, The Netherlands, September 7-8, 2023, Proceedings*, pages 175–189, Berlin, Heidelberg, 2023. Springer-Verlag.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 605–622, Columbia University, New York, New York, USA, 2016. PMLR.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005a.

Graham Cormode and Shan Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. *ACM Transactions on Database Systems (TODS)*, 30(1):249–278, 2005b.

Mayur Datar, Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Maintaining stream statistics over sliding windows. *SIAM journal on computing*, 31(6):1794–1813, 2002.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(6):165–202, 2012.

Liad Erez, Orin Levy, and Yishay Mansour. Regret guarantees for adversarial contextual bandits with delayed feedback. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk van der Hoeven, Nicolo' Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help?, 2023.

Philippe Flajolet and G Nigel Martin. Probabilistic counting algorithms for data base applications. *Journal of computer and system sciences*, 31(2):182–209, 1985.

Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. *Discrete mathematics & theoretical computer science*, Proceedings, 2007.

Amelie Heliou, Panayotis Mertikopoulos, and Zhengyuan Zhou. Gradient-free online learning in games with delayed rewards, 2020.

David Helmbold and Sandra Panizza. Some label efficient learning results. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, pages 218–230, New York, NY, USA, 1997. Association for Computing Machinery.

Jiatai Huang, Yan Dai, and Longbo Huang. Banker online mirror descent: A universal approach for delayed online bandit learning, 2023.

Piotr Indyk and David Woodruff. Optimal approximations of the frequency moments of data streams. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 202–208, 2005.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1453–1461. PMLR, 2013.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016.

Daniel M Kane, Jelani Nelson, and David P Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 41–52, 2010.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5969–5978. PMLR, 2021.

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

Roy Levin, Ellis S. Cohen, William M. Corwin, Fred J. Pollack, and William A. Wulf. Policy/mechanism separation in hydra. *Proceedings of the fifth ACM symposium on Operating systems principles*, 1975.

R. Lipton. Online interval scheduling. In *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '94, pages 302–311, USA, 1994. Society for Industrial and Applied Mathematics.

Thodoris Lykouris and Wentao Weng. Learning to defer in content moderation: The human-ai interplay, 2024.

Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*, pages 346–357. Elsevier, 2002.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback, 2022.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays, 2024.

Chris Mesterharm. On-line learning with delayed label feedback. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, pages 399–413, Berlin, Heidelberg, 2005. Springer-Verlag.

Shanmugavelayutham Muthukrishnan et al. Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2):117–236, 2005.

Francesco Orabona. A modern introduction to online learning, 2023.

Francesco Orabona and David Pal. Scale-free online learning, 2016.

Michael L. Pinedo. *Scheduling: Theory, Algorithms, and Systems*. Mathematics and Statistics, Mathematics and Statistics (R0). Springer, Cham, 6 edition, 2022.

Sudeep Raja Putta and Shipra Agrawal. Scale free adversarial multi armed bandits, 2021.

Hao Qiu, Emmanuel Esposito, and Mengxiao Zhang. Exploiting curvature in online convex optimization with delayed feedback, 2025.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 280–287, Bejing, China, 6 2014. PMLR.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1–2):1–286, 2019.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dirk Van Der Hoeven and Nicolò Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2022–2044. PMLR, 2022.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Yuanyu Wan, Chang Yao, Mingli Song, and Lijun Zhang. Improved regret for bandit convex optimization with delayed feedback, 2024.

M.J. Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48:1959–1976, 07 2002.

Gerhard J. Woeginger. On-line scheduling of jobs with fixed start and end times. *Theoretical Computer Science*, 130(1):5–16, 1994.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3285–3294. PMLR, 2020.

Appendix A. Additional Related Work

Online learning with delays. One of the earliest works in online learning with delays against an oblivious adversary was by Weinberger and Ordentlich (2002), where they determined minimax regret of $\Theta(\sqrt{T(d+1)\log(K)})$ for the full-information feedback under fixed delays. For the full-information regime with round-dependent delays, Joulani et al. (2016); Quanrud and Khashabi (2015) proved the minimax regret bound of $\Theta(\sqrt{(D+T)\log(K)})$. The delayed bandit setting was later introduced by Cesa-Bianchi et al. (2016), who established a regret lower bound of $\Omega(\max\{\sqrt{TK}, \sqrt{Td\log(K)}\})$ for the fixed delay case, where delays are fixed as $d_t = d$. For the round-dependent delay case, Thune et al. (2019) proposed the Delayed Exponential Weights algorithm, achieving an almost matching regret bound of $O(\sqrt{(TK+D)\log(K)})$. In the following year, Zimmert and Seldin (2020) proposed an FTRL-based algorithm, which had $O(\sqrt{TK + D\log(K)})$ regret, matching the lower bound from Cesa-Bianchi et al. (2016).

There has been ongoing interest in online and bandit convex optimization under delayed feedback, namely Quanrud and Khashabi (2015); Joulani et al. (2016); Heliou et al. (2020); Bistritz et al. (2022), and more recently Wan et al. (2024); Qiu et al. (2025). Notably, Wan et al. (2024) proposed an algorithm that also employs a batching technique similar in spirit to the one used in our Algorithm 1.

Significant attention has also been given to the stochastic version of the setting, notably by Joulani et al. (2013) and Lancewicki et al. (2021). Furthermore, the FTRL-based approach from Zimmert and Seldin (2020) has been extended to the best-of-both-worlds framework in Masoudian et al. (2022, 2024).

Recently, a number of studies have explored delayed bandits under assumptions that mirror real-life applications. For instance, Erez et al. (2024) examined contextual bandits with delayed feedback, Van Der Hoeven and Cesa-Bianchi (2022) considered bandits with arm-dependent delays, and Esposito et al. (2023) investigated delayed bandits with intermediate observations.

We further explore the connection of our Delay Scheduling to Delayed Online Learning in Appendix A.2.

Label-efficient online learning. In a label-efficient game, as proposed by Helmbold and Panizza (1997), it is assumed that the learner can query feedback from at most M rounds out of T. The full-information setting has been shown to have minimax regret of $\Theta(T\sqrt{\log(K)/M})$ by Cesa-Bianchi et al. (2005). Next, Audibert and Bubeck (2010) have solved the bandit feedback regime with minimax regret of $\Theta(T\sqrt{K/M})$.

Appendix A.1 provides additional insights into the relationship between our Delay Scheduling and label-efficient online learning.

Online job scheduling. Online Job Scheduling (e.g., see Borodin and El-Yaniv (1998); Pinedo (2022)) is a broad research area that involves studying the problem of assigning sequentially arriving jobs across multiple resources with the goal of optimizing specific objectives, such as minimizing maximum tardiness or maximizing weighted throughput. Online Interval Scheduling is a variant of Online Job Scheduling where jobs have fixed starting and end times, which is precisely our case in Delay Scheduling. In *Online Interval Scheduling* with multiple resources, the algorithm is presented with a sequence of time intervals ordered by their starting times, and it must immediately assign each interval to one of the machines or reject it, ensuring no overlap on the same machine while optimizing the overall schedule with respect to some metric (e.g., number of intervals, total length). In the original paper by Lipton (1994), intervals could not be unscheduled and their lengths were

announced at starting times. Woeginger (1994) considered a modification of this setting where intervals can be unscheduled.

In the analogy between our Delay Scheduling setting and Online Job (Interval) Scheduling, rounds correspond to jobs with increasing arrival times (a_t) and arbitrary processing times (p_t), where capacity C represents the number of initially idle resources available for processing jobs. A job can only be assigned to an idle resource upon arrival, after which the resource becomes busy and processing continues until time $a_t + p_t$ or possible preemption, when resource becomes idle again. Then, the delay d_t of round t corresponds to the number of jobs arriving during the processing interval $[a_t, a_t + p_t)$ of job t. Note that long processing time does not necessarily imply long delay, as delays serve as the measure of how concurrent the jobs are.

Streaming. Streaming algorithms are specialized algorithms designed to process massive data streams using limited memory to answer queries approximately. Instead of storing and processing the entire dataset, these algorithms make a single pass (or very few passes) over the data, maintaining a small summary, usually referred to as a sketch, which captures essential information about the stream. By analogy, the limitation on space in streaming is replaced in our model by the constraint on how many rounds should be tracked in the online learning with delays model. In both models, a key technique is to use randomized sampling to select elements from the stream in a way that still allows for good performance: answering queries in streaming and minimizing regret in our setting.

The foundational work of Alon, Matias, and Szegedy Alon et al. (1996) introduced the streaming model and developed space-efficient algorithms for estimating frequency moments. Cormode and Muthukrishnan Cormode and Muthukrishnan (2005a) proposed the Count-Min Sketch, a widely used tool for approximate frequency estimation. In the domain of cardinality estimation, Flajolet et al. Flajolet et al. (2007) developed HyperLogLog, an optimal algorithm for counting distinct elements in a stream. This field continues to be extensively studied; see, for example, Charikar et al. (2002); Cormode and Muthukrishnan (2005b); Datar et al. (2002); Flajolet and Martin (1985); Indyk and Woodruff (2005); Kane et al. (2010); Manku and Motwani (2002); Muthukrishnan et al. (2005).

A.1. Related Setting: Delayed Online Learning

Delayed Online Learning considers scenarios with delayed feedback but no resource-driven constraints (i.e., $C = \infty$). It is straightforward to see that, for every selection of delays $\{d_t\}_{t=1}^T$, the regret in the Delay Scheduling game cannot exceed that of the corresponding Delayed Online Learning game with the same delays. Theorem 12 presents the lower bound for the fixed-delay case in the bandit regime.

Delayed Game

- *Visible Parameters:* number of actions K.
- *Latent Parameters:* number of rounds T.
- *Pre-game:* adversary selects losses $l_t \in [0, 1]^K$ and delays $d_t \in [T - t]$ for all $t \in [T]$.

For each round $t = 1, 2, \ldots, T$ repeat:

    0. If the setting is **clairvoyant**, then the environment reveals d_t.
    1. The player plays $A_t \in [K]$ and incurs corresponding loss l_{t,A_t}.
    2. For all $s \leq t$ such that $s + d_s = t$, the environment reveals:
        • index-value pair (s, l_{s,A_s}) in the **multi-armed bandit** game,
        • index-vector pair (s, l_s) in the **full-information** game.

Theorem 12 (Cesa-Bianchi et al. (2016), proof of Corollary 11, Appendix D) *The minimax regret in the multi-armed bandit setting with fixed delays $d_t = d$ is of the order*

$$\Omega\left(\max\left\{\sqrt{KT}, \sqrt{Td\log(K)}\right\}\right).$$

A.2. Related Setting: Label Efficient Online Learning

Since the total sum of delays for observed rounds in delay scheduling is bounded by CT, whereas the total delay D can grow quadratically in terms of T in the worst case, the capacity constraint may significantly limit the number of observations. Previously studied label efficient games have already explored scenarios where the number of observations is bounded.

Label Efficient Game

- *Visible Parameters:* number of arms K, number of rounds T, number of queries M.
- *Pre-game:* adversary selects losses $l_{t,i} \in [0,1]$ for all $t \in [T]$ and $i \in [K]$.

For each round $t = 1, 2, \ldots, T$ repeat:
1. The player plays $A_t \in [K]$ and incurs corresponding loss l_{t,A_t}.
2. The player observes:
   - value l_{s,A_s} in the **multi-armed bandit** game,
   - vector l_s in the **full-information** game,

   only if he asks for it with the global constraint that he is not allowed to ask it more than M times throughout the game.

In particular, it has been shown that when the number of observations is capped at M, the minimax regret of a label efficient game is $\Theta(T\sqrt{K/M})$ in the bandit regime and $\Theta(T\sqrt{\log(K)/M})$ in the full-information regime. It follows directly that if, for some selection of delays $\{d_t\}_{t=1}^T$ in delay scheduling, it is impossible to make more than M observations without violating the capacity constraint, then the regret in the delay scheduling game cannot exceed that of the corresponding label efficient game with M queries.

Theorem 13 (Audibert and Bubeck (2010), Theorem 30) *Let $M > K$. Consider a label efficient game where a player can query feedback from at most M rounds. Let* sup *be taken over all oblivious adversaries and* inf *over all players, then the following holds true in the label efficient full-information game:*

$$\inf\sup\mathfrak{R}_T \geqslant 0.03T\sqrt{\frac{\log(K)}{M}},$$

and in the label efficient bandit game we have:

$$\inf\sup\mathfrak{R}_T \geqslant 0.04T\sqrt{\frac{K}{M}}.$$

Appendix B. Additional Preliminaries

For the reader's convenience, we provide a consolidated summary of the notation used throughout the paper. Table 7 lists the key symbols along with their definitions. Additionally, we include the definition of the Pareto distribution (Definition 14) and a brief overview of harmonic numbers.

Symbol	Definition / Description
$L_{t,a}$	$\sum_{s=1}^{t-1} l_{t,a}$ (Cumulative loss for action a up to time t)
$\mathfrak{R}_{T,a}$	$\mathbb{E}\left[\sum_{t=1}^{T} l_{t,A_t}\right] - L_{T+1,a}$ (Expected regret w.r.t. action a)
$i*$	$\operatorname{argmin}_{a\in[K]}\{L_{T+1,a}\}$ (Best action in hindsight)
$\mathfrak{R}_T$	$\mathfrak{R}_{T,i*}$ (Expected regret w.r.t. the best action $i*$)
O_t	$\{s \in [t-1] : s + d_s < t\}$ (Observed set of round t)
$\mathcal{W}_t$	$\{s \in [t-1] : s + d_s \geqslant t\}$ (Working set of round t)
σ_t	$|\mathcal{W}_t|$ (Number of outstanding delays at round t)
D	$\sum_{t=1}^{T} \sigma_t = \sum_{t=1}^{T} d_t$ (Total delay across all rounds)
$\sigma_{\max}$	$\max_{t\in[T]} \sigma_t$ (Maximum number of outstanding delays across all rounds)
$d_{\max}$	$\max_{t\in[T]} d_t$ (Maximum delay across all rounds)
S_t^0	State of the tracking set S immediately before round t.
S_t^1	State of the tracking set S in round t, after the decision whether to include t in S has been fully processed, and before removing any elements.
Z_t	$\mathbb{I}\left(t \in S_{t+d_t}^1\right)$ (Indicator that feedback from round t is observed)

Table 7   Notation used throughout the paper.

Definition 14 (Pareto Distribution) *A random variable X follows a Pareto distribution with scale parameter $c > 0$ and shape parameter $\beta > 0$, denoted by $X \sim \operatorname{Pareto}(c, \beta)$, if its cumulative distribution function is given by $F_X(x) = \mathbb{I}(x \geqslant c)(1 - (\frac{c}{x})^{\beta})$.*

Let $H_t = \sum_{s=1}^{t} 1/s$ denote the t-th harmonic number. It is a well-known fact that $H_t = \log(t) + O(1)$, with $\gamma = \lim_{t\to\infty}(H_t - \log(t)) \approx 0.577$ known as the Euler–Mascheroni constant.

Appendix C. Delayed FTRL with Time-Varying Loss Scales: Proof

In this section, we prove Theorem 1. Algorithm 5 presents the Delayed FTRL algorithm, explored in Section 2, with learning rate sequences taken as parameters. In the full-information regime, the algorithm uses the full loss vector as the estimator, while in the bandit regime, it performs importance weighting for the observed loss.

Following Zimmert and Seldin (2020), the proof is structured into six facts and three lemmas. Before proceeding, we restate the notation from Zimmert and Seldin (2020) related to the algorithm's regularization structure. Given non-increasing sequences of learning rates α_t, β_t, the hybrid regularizer in round t is $F_t(x) = F_{t,1}(x) + F_{t,2}(x)$, where $F_{t,1}(x) = \alpha_t^{-1} F_{\mathrm{Ts}}(x)$ and $F_{t,2}(x) = \beta_t^{-1} F_{\mathrm{NE}}(x)$ denote the components. For each regularizer $F \in \{F_t\}_{t\in[T]}$, we define unconstrained and constrained convex conjugates:

$$F^*(\theta) = \sup_{x\in\mathbb{R}^k} \{\langle x, \theta \rangle - F(x)\},$$

$$\overline{F}^*(\theta) = \sup_{x\in\Delta([K])} \{\langle x, \theta \rangle - F(x)\}.$$

Algorithm 5 Generic Delayed FTRL

Input : Number of arms K.

Parameters : Learning rates $(\alpha_t)_{t=1}^T$ and $(\beta_t)_{t=1}^T$.

1. Initialize $\widehat{L}_1^{\text{obs}} = \mathbf{0}_K$.

2. **For** round $t = 1, 2, \ldots, T$:

   (a) Draw and play an action $A_t \sim x_t = \text{argmin}_{x \in \Delta([K])}\langle x, \widehat{L}_t^{\text{obs}}\rangle + \alpha_t^{-1}F_{\text{Ts}}(x) + \beta_t^{-1}F_{\text{NE}}(x)$.

   (b) For each round $s \in [t]$ whose delays expires this round (i.e., $s + d_s = t$):
   - **Bandit:** Observe (s, l_{s,A_s}) and construct estimator $\hat{l}_s = l_{s,A_s}x_{s,A_s}^{-1}\boldsymbol{e}_{A_s}$.
   - **Full-information:** Observe (s, l_s) and construct estimator $\hat{l}_s = l_s$.

   (c) Update $\widehat{L}_{t+1}^{\text{obs}} = \widehat{L}_t^{\text{obs}} + \sum_{s:s+d_s=t} \hat{l}_s$.

Define $f_t : \mathbb{R} \to \mathbb{R}$ as $f_t(x) = -2\alpha_t^{-1}\sqrt{x} + \beta_t^{-1}x\log(x)$, decomposed as $f_t = f_{t,1} + f_{t,2}$ with $f_{t,1}(x) = -2\alpha_t^{-1}\sqrt{x}$ and $f_{t,2}(x) = \beta_t^{-1}x\log(x)$. Then $F_t(x) = \sum_{i=1}^K f_t(x_i)$. The algorithm's update rule for the weights can be written as $x_t = \nabla\overline{F}_t^*(-\widehat{L}_t^{\text{obs}})$ (e.g., see Theorems 5.7 and 6.8 from Orabona (2023)). Also, as F_t^* considers maximization over $\mathbb{R}^K$, it holds that $F_t^*(\theta) = \sum_{i=1}^K f_t^*(\theta_i)$.

Proof (Theorem 1) Let $\widehat{L}_{t+1} = \sum_{s=1}^t \hat{l}_s$ for each $t \in [T]$. Let $i^* = \text{argmin}_{i \in [K]} L_{t,i}$. Expand $\mathfrak{R}_T$ as follows

$$
\begin{aligned}
\mathfrak{R}_T &= \mathbb{E}\left[\sum_{t=1}^T l_{t,A_t}\right] - L_{T+1,i^*} \\
&= \mathbb{E}\left[\sum_{t=1}^T \left(\overline{F}_t^*(-\widehat{L}_t^{\text{obs}} - \hat{l}_t) - \overline{F}_t^*(-\widehat{L}_t^{\text{obs}}) + \langle x_t, \hat{l}_t\rangle\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^T \left(\overline{F}_t^*(-\widehat{L}_t) - \overline{F}_t^*(-\widehat{L}_{t+1})\right) - \widehat{L}_{T+1,i^*}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^T \left(\overline{F}_t^*(-\widehat{L}_t^{\text{obs}}) - \overline{F}_t^*(-\widehat{L}_t^{\text{obs}} - \hat{l}_t) - \overline{F}_t^*(-\widehat{L}_t) + \overline{F}_t^*(-\widehat{L}_{t+1})\right)\right].
\end{aligned}
$$

The resulting three terms can be bounded for both regimes using Lemmas 21, 22, and 23. ∎

Fact 15 $f_t''(x) : \mathbb{R}_+ \to \mathbb{R}_+$ *are monotonically decreasing positive functions and* $f_t^{*\prime} : \mathbb{R} \to \mathbb{R}_+$ *are convex and monotonically increasing*

Proof By definition $f_t''(x) = \frac{1}{2}\alpha_t^{-1}x^{-3/2} + \beta_t^{-1}x^{-1} > 0$, proving the first statement.

Since f_t are Legendre functions, we have $f_t^{*\prime\prime}(x^*) = (f_t''(f_t^{*\prime}(x^*)))^{-1} > 0$, showing that functions $f_t^{*\prime}$ are monotonically increasing. As both $f_t''(x)^{-1}$ and $f_t^{*\prime}(x^*)$ are increasing, their composition is also increasing, so $f_t^{*\prime\prime\prime} > 0$, showing that $f_t^{*\prime}$ are convex. ∎

Fact 16 *For every convex F, for $L \in \mathbb{R}^K$ and $c \in \mathbb{R}$:*

$$\overline{F}^*(L + c\mathbf{1}_K) = \overline{F}^*(L) + c.$$

Proof By definition $\overline{F}^*(L + c\mathbf{1}_K) = \sup_{x \in \Delta([K])}\langle x, L + c\mathbf{1}_K\rangle - F(x) = \sup_{x \in \Delta([K])}\langle x, L\rangle - F(x) + c$. ∎

Fact 17 *For every x_t, there exists $c \in \mathbb{R}$ such that:*

$$x_t = \nabla \overline{F}_t^*(-\widehat{L}_t^{obs}) = \nabla F_t^*(-\widehat{L}_t^{obs} + c\mathbf{1}_K) = \nabla F_t^*(\nabla F_t(x_t)).$$

Proof By the KKT conditions, there exists $c \in \mathbb{R}$ such that $x_t = \arg\sup_{x \in \Delta([K])}\langle x, -\widehat{L}_t^{obs}\rangle - F_t(x)$ satisfies $\nabla F_t(x_t) = -\widehat{L}_t^{obs} + c\mathbf{1}_K$. The rest follows by the standard property $\nabla F = (\nabla F^*)^{-1}$ for Legendre F. ∎

Fact 18 *For every Legendre function F and $L \in \mathbb{R}^K$, it holds that*

$$\overline{F}^*(L) \leqslant F^*(L),$$

with equality if and only if there exists $x \in \Delta([K])$ such that $L = \nabla F(x)$.

Proof The first statement follows from the definitions of convex conjugates as $\Delta([K]) \subset \mathbb{R}^K$. For the second statement, equality $\overline{F}^*(L) = F^*(L)$ would be equivalent to

$$\nabla F^*(L) = \arg\sup_{x \in \mathbb{R}^K}\langle x, L\rangle - F(x) = \arg\sup_{x \in \Delta([K])}\langle x, L\rangle - F(x) \in \Delta([K]),$$

and equivalently $L = \nabla F(x)$ for $x = \nabla F^*(L) \in \Delta([K])$ ∎

Fact 19 *For every $x \in \Delta([K])$, $L \geqslant 0$, and $i \in [K]$ it holds that:*

$$\nabla \overline{F}_t^*(\nabla F_t(x) - L)_i \geqslant \nabla F_t^*(\nabla F_t(x) - L)_i.$$

Proof Using a similar argument as in the proof of Fact 17, by the KKT conditions, we can find $c \in \mathbb{R}$ such that $\nabla \overline{F}_t^*(\nabla F_t(x) - L) = \nabla F_t^*(\nabla F_t(x) - L + c\mathbf{1}_K)$. By Fact 15 $f_t^{*\prime}$ is monotonically increasing, so the statement is equivalent to $c \geqslant 0$. It cannot be that $c < 0$, because otherwise it would hold that

$$
\begin{aligned}
1 &= \sum_{i=1}^{K}(\nabla \overline{F}_t^*(\nabla F_t(x) - L))_i \\
&= \sum_{i=1}^{K}(\nabla F_t^*(\nabla F_t(x) - L + c\mathbf{1}_K))_i \\
&= \sum_{i=1}^{K} f_t^{*\prime}(f_t'(x_i) - L_i + c) \\
&< \sum_{i=1}^{K} f_t^{*\prime}(f_t'(x_i)) \\
&= 1.
\end{aligned}
$$

∎

Fact 20 *Let $D_F(x, y) = F(x) - F(y) - \langle x - y, \nabla F(y) \rangle$ denote the Bregman divergence of a function F. For every Legendre function f with a monotonically decreasing second derivative, $x \in dom(f)$, and $l \geq 0$, such that $f'(x) - l \in dom(f^*)$, it holds that*

$$D_{f*}(f'(x) - l, f'(x)) \leq \frac{l^2}{2f''(x)}.$$

Proof By Taylor's theorem, there exists $\tilde{x} \in [f^{*\prime}(f'(x) - l), x]$ such that

$$D_{f*}(f'(x) - \ell, f'(x)) = \frac{\ell^2}{2f''(\tilde{x})}.$$

Since $\tilde{x} \leq x$ and f'' is decreasing, we have $f''(\tilde{x})^{-1} \leq f''(x)^{-1}$, completing the proof. ∎

Lemma 21 *For every $t \in [T]$, the following holds true in the bandit setting:*

$$\mathbb{E}\left[\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{l}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{l}_t \rangle\right] \leq \sqrt{K}\alpha_t B_t^2.$$

And in the full-information setting:

$$\mathbb{E}\left[\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{l}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{l}_t \rangle\right] \leq \frac{1}{2}\beta_t B_t^2.$$

Proof Our proof builds on the argument from Zimmert and Seldin (2020) and extends it to handle general estimators. We then apply this general result to both the bandit and full-information settings, incorporating modifications to address losses of time-varying scales in both cases. For both settings, we write

$$
\begin{aligned}
\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{l}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{l}_t \rangle &\overset{(a)}{=} \overline{F}_t^*(\nabla F_t(x_t) - \hat{l}_t) - \overline{F}_t^*(\nabla F_t(x_t)) + \langle x_t, \hat{l}_t \rangle \\
&\overset{(b)}{\leq} F_t^*(\nabla F_t(x_t) - \hat{l}_t) - F_t^*(\nabla F_t(x_t)) + \langle x_t, \hat{l}_t \rangle \\
&= \sum_{i=1}^K D_{f_t^*}\left(f_t'(x_{t,i}) - \hat{l}_{t,i}, f_t'(x_{t,i})\right),
\end{aligned}
\tag{1}
$$

where (a) follows from Facts 17 and then 16, while (b) follows from both parts of Fact 18. In the bandit setting, with estimators $\hat{l}_t = l_{t,A_t} x_{t,A_t}^{-1} e_{A_t}$, (1) can be further bounded as

$$
\begin{aligned}
\sum_{i=1}^K D_{f_t^*}\left(f_t'(x_{t,i}) - \hat{l}_{t,i}, f_t'(x_{t,i})\right) &\leq D_{f_t^*}\left(f_t'(x_{t,A_t}) - \hat{l}_{t,A_t}, f_t'(x_{t,A_t})\right) \\
&\overset{(c)}{\leq} \frac{1}{2}(\hat{l}_{t,A_t})^2 f_{t,1}''(x_{t,A_t})^{-1} \\
&= \frac{1}{2}\left(l_{t,A_t} x_{t,A_t}^{-1}\right)^2 \left(2\alpha_t(x_{t,A_t})^{3/2}\right) \\
&\leq x_{t,A_t}^{-1/2} \alpha_t B_t^2,
\end{aligned}
$$

where (c) follows from Fact 20 and then bounding f_t'' with $f_{t,1}''$ from below. By taking expectation over each $A_t \sim x_t$, we obtain the bandit result:

$$\mathbb{E}\left[\overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}} - \hat{l}_t) - \overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}}) + \langle x_t, \hat{l}_t \rangle\right] \leqslant \mathbb{E}\left[\sum_{i=1}^{K} (x_{t,i})^{1/2} \cdot \alpha_t B_t^2\right] \leqslant \sqrt{K}\alpha_t B_t^2.$$

In the full-information setting, with estimators $\hat{l}_t = l_t$, (1) can be bounded as

$$
\begin{aligned}
\sum_{i=1}^{K} D_{f_t^*}\left(f_t'(x_{t,i}) - \hat{l}_{t,i}, f_t'(x_{t,i})\right) &\overset{(d)}{\leqslant} \sum_{i=1}^{K} \frac{1}{2}(\hat{l}_t)^2 f_{t,2}''(x_{t,i})^{-1} \\
&= \frac{1}{2} l_t^2 \sum_{i=1}^{K} \beta_t x_{t,i} \\
&\leqslant \frac{1}{2} \beta_t B_t^2,
\end{aligned}
$$

where (d) follows from Fact 20 and bounding f_t'' with $f_{t,2}''$ from below. This concludes the proof of the full-information result. ∎

Lemma 22 *For non-increasing learning rates α_t, β_t, in both bandit and full-information settings, it holds almost surely that*

$$\sum_{t=1}^{T} \left(\overline{F}_t^*(-\widehat{L}_t) - \overline{F}_t^*(-\widehat{L}_{t+1})\right) - \widehat{L}_{T+1,i*} \leqslant 2\sqrt{K}\alpha_T^{-1} + \log(K)\beta_T^{-1}.$$

Proof This proof repeats the argument from Zimmert and Seldin (2020) without any notable changes. Let $\bar{x}_t = \arg\sup_{x \in \Delta([K])} \langle x, -\widehat{L}_t \rangle - F_t(x)$, so that

$$\overline{F}_t^*(-\widehat{L}_t) = \langle \bar{x}_t, -\widehat{L}_t \rangle - F_t(\bar{x}_t) = \sup_{x \in \Delta([K])} \langle x, -\widehat{L}_t \rangle - F_t(x).$$

By the definition of constrained convex conjugate, it holds that

$$
\begin{aligned}
\overline{F}_T^*(-\widehat{L}_{T+1}) &\geqslant \langle e_{i*}, -\widehat{L}_{T+1} \rangle - F_T(e_{i*}) \geqslant -\widehat{L}_{T+1,i*}, \\
\overline{F}_{t-1}^*(-\widehat{L}_t) &\geqslant \langle \bar{x}_t, -\widehat{L}_t \rangle - F_{t-1}(\bar{x}_t).
\end{aligned}
$$

Plugging these inequalities into the LHS gives us

$$\sum_{t=1}^{T} \left(\overline{F}_t^*(-\hat{L}_t) - \overline{F}_t^*(-\hat{L}_{t+1}) \right) - \hat{L}_{T+1,i*}$$

$$\leqslant \overline{F}_1^*(-\hat{L}_1) + \sum_{t=2}^{T} \left(-\overline{F}_{t-1}^*(-\hat{L}_t) + \overline{F}_t^*(-\hat{L}_t) \right)$$

$$\leqslant -F_1(\bar{x}_1) + \sum_{t=2}^{T} (F_{t-1}(\bar{x}_t) - F_t(\bar{x}_t))$$

$$\leqslant \sup_{x \in \Delta([K])} -F_1(x) + \sum_{t=2}^{T} \sup_{x \in \Delta([K])} (F_{t-1}(x) - F_t(x))$$

$$= \sup_{x \in \Delta([K])} \left((-\alpha_1^{-1})F_{0,1} + (-\beta_1^{-1})F_{0,2} \right)(x)$$

$$+ \sum_{t=2}^{T} \sup_{x \in \Delta([K])} \left((\alpha_{t-1}^{-1} - \alpha_t^{-1})F_{0,1} + (\beta_{t-1}^{-1} - \beta_t^{-1})F_{0,2} \right)(x)$$

$$\overset{(a)}{=} -F_1(\mathbf{1}_K/K) + \sum_{t=2}^{T} (F_{t-1}(\mathbf{1}_K/K) - F_t(\mathbf{1}_K/K))$$

$$= -F_T(\mathbf{1}_K/K)$$

$$= 2\sqrt{K}\alpha_T^{-1} + \log(K)\beta_T^{-1},$$

where (a) follows from the fact that both learning rates are non-increasing and that $\mathbf{1}_K/K$ minimizes both $F_{0,1}$ and $F_{0,2}$ on $\Delta([K])$. ∎

Lemma 23 *For every $t \in [T]$, in both bandit and full-information settings, it holds that*

$$\mathbb{E}\left[\overline{F}_t^*(-\hat{L}_t^{obs}) - \overline{F}_t^*(-\hat{L}_t^{obs} - \hat{l}_t) - \overline{F}_t^*(-\hat{L}_t) + \overline{F}_t^*(-\hat{L}_{t+1}) \right] \leqslant \beta_t B_t \sum_{s \in \mathcal{W}_t} B_s.$$

Proof Similar to the proof of Lemma 21, here our proof extends the argument from Zimmert and Seldin (2020) to analyze general estimators. Building on this generalized framework, we incorporate modifications to address losses of time-varying scales and apply this result to both bandit and full-information settings.

Let $\hat{L}_t^{\text{miss}} = \hat{L}_t - \hat{L}_t^{\text{obs}} = \sum_{s \in \mathcal{W}_t} \hat{l}_s$ denote the sum of estimators whose values were determined but not observed by the start of round t. Consider function $\bar{x}(z) = \nabla \overline{F}_t^*(-\hat{L}_t^{\text{obs}} - z\hat{l}_t)$. Then, for

both bandits and full-information regimes, we write

$$\overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}}) - \overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}} - \hat{l}_t) - \overline{F}_t^*(-\widehat{L}_t) + \overline{F}_t^*(-\widehat{L}_{t+1})$$

$$\overset{\text{(a)}}{=} \int_0^1 \langle \hat{l}_t, \nabla \overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}} - z\hat{l}_t) \rangle dz - \int_0^1 \langle \hat{l}_t, \nabla \overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}} - \widehat{L}_t^{\mathrm{miss}} - z\hat{l}_t) \rangle dz$$

$$\overset{\text{(b)}}{=} \int_0^1 \langle \hat{l}_t, \bar{x}(z) - \nabla \overline{F}_t^*(\nabla F_t(\bar{x}(z)) - \widehat{L}_t^{\mathrm{miss}}) \rangle dz$$

$$\overset{\text{(c)}}{\leqslant} \int_0^1 \langle \hat{l}_t, \bar{x}(z) - \nabla F_t^*(\nabla F_t(\bar{x}(z)) - \widehat{L}_t^{\mathrm{miss}}) \rangle dz$$

$$= \sum_{i=1}^K \int_0^1 \hat{l}_{t,i}(\bar{x}_i(z) - f_t^{*\prime}(f_t'(\bar{x}_i(z) - \widehat{L}_{t,i}^{\mathrm{miss}}))) dz$$

$$\overset{\text{(d)}}{\leqslant} \sum_{i=1}^K \int_0^1 \hat{l}_{t,i}(f_t^{*\prime\prime}(f_t'(\bar{x}_i(z)))) \widehat{L}_{t,i}^{\mathrm{miss}} dz$$

$$= \sum_{i=1}^K \int_0^1 \hat{l}_{t,i}((f_t'' \circ f_t^{*\prime})(f_t'(\bar{x}_i(z))))^{-1} \widehat{L}_{t,i}^{\mathrm{miss}} dz$$

$$= \sum_{i=1}^K \int_0^1 \hat{l}_{t,i} f_t''(\bar{x}_i(z))^{-1} \widehat{L}_{t,i}^{\mathrm{miss}} dz, \tag{2}$$

where (a) follows from the fundamental theorem of calculus, (b) substitutes $\bar{x}(z)$ and applies Fact 17, (c) applies Fact 19, and (d) follows from the convexity of $f_t^{*\prime}$ by Fact 15. In the bandit setting, with estimators $\hat{l}_t = l_{t,A_t} x_{t,A_t}^{-1} \boldsymbol{e}_{A_t}$, (2) can be further bounded as

$$\sum_{i=1}^K \int_0^1 \hat{l}_{t,i} f_t''(\bar{x}(z))^{-1} \widehat{L}_{t,i}^{\mathrm{miss}} dz = \int_0^1 \hat{l}_{t,A_t} f_t''(\bar{x}_{A_t}(z))^{-1} \widehat{L}_{t,A_t}^{\mathrm{miss}} dz$$

$$\overset{\text{(e)}}{\leqslant} \int_0^1 \hat{l}_{t,A_t} f_t''(x_{t,A_t})^{-1} \widehat{L}_{t,A_t}^{\mathrm{miss}} dz$$

$$\leqslant \int_0^1 \hat{l}_{t,A_t} f_{t,2}''(x_{t,A_t})^{-1} \widehat{L}_{t,A_t}^{\mathrm{miss}} dz$$

$$= \int_0^1 (l_{t,A_t} x_{t,A_t}^{-1})(\beta_t x_{t,A_t}) \widehat{L}_{t,A_t}^{\mathrm{miss}} dz$$

$$\leqslant \beta_t B_t \widehat{L}_{t,A_t}^{\mathrm{miss}},$$

where (e) follows because $f_t''(x)$ is monotonically increasing by Fact 15 and for every $z \geqslant 0$ it holds that

$$\bar{x}_{A_t}(z) = (\nabla \overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}} - z l_{t,A_t} x_{t,A_t}^{-1} \boldsymbol{e}_{A_t}))_{A_t} \leqslant (\nabla \overline{F}_t^*(-\widehat{L}_t^{\mathrm{obs}}))_{A_t} = x_{t,A_t}. \tag{3}$$

Equation (3) holds because $\nabla \overline{F}_t^*(-L)_{A_t}$ decreases when the loss increases only in coordinate A_t. As A_t and $\{A_s : s \in \mathcal{W}_t\}$ are independent given $\{A_s : s \in O_t\}$, we have in expectation

$$\mathbb{E}\left[\widehat{L}_{t,A_t}^{\mathrm{miss}}\right] = \sum_{s \in \mathcal{W}_t} \mathbb{E}\left[\hat{l}_{s,A_t}\right] = \sum_{s \in \mathcal{W}_t} \sum_{i=1}^K \mathbb{E}\left[\hat{l}_{s,i} \mathbb{I}(A_t = i)\right] = \sum_{s \in \mathcal{W}_t} \sum_{i=1}^K \mathbb{E}\left[\hat{l}_{s,i}\right] \mathbb{P}\{A_t = i\} \leqslant \sum_{s \in \mathcal{W}_t} B_s,$$

which concludes the proof for the bandit result. In the full-information setting, with estimators $\hat{l}_t = l_t$, (2) can be bounded as

$$
\begin{aligned}
\sum_{i=1}^{K} \int_0^1 \hat{l}_{t,i} f_t''(\bar{x}_i(z))^{-1} \widehat{L}_{t,i}^{\text{miss}} dz &\leqslant \sum_{i=1}^{K} \int_0^1 \hat{l}_{t,i} f_{t,2}''(\bar{x}_i(z))^{-1} \widehat{L}_{t,i}^{\text{miss}} dz \\
&= \sum_{i=1}^{K} \int_0^1 l_{t,i}(\beta_t \bar{x}_i(z)) \widehat{L}_{t,i}^{\text{miss}} dz \\
&\leqslant \beta_t B_t \sum_{i=1}^{K} \int_0^1 \bar{x}_i(z) \widehat{L}_{t,i}^{\text{miss}} dz \\
&\leqslant \beta_t B_t \sum_{s \in \mathcal{W}_t} \int_0^1 \sum_{i=1}^{K} \bar{x}_i(z) l_{s,i} dz \\
&\overset{\text{(f)}}{\leqslant} \beta_t B_t \sum_{s \in \mathcal{W}_t} B_s,
\end{aligned}
$$

where (f) follows from the fact that $\bar{x}(z) \in \Delta([K])$ and each $l_{s,i} \leqslant B_s$. This concludes the proof for the full-information result. ∎

Appendix D. Batch Partitioning Algorithm for Delay Scheduling: Proof

In this section, we prove Theorem 2. To do so, we first establish the more general Theorem 25, which applies to a broader class of learning rates. Theorem 2 then follows as a corollary.

We introduce a bit more batch-level. Extend the final batch with zero losses $l_t = 0$ and delays $d_t = 0$ for $t \in [T'b] \setminus [T]$. Then, for each batch $\tau \in [T']$, let $\mathcal{B}_\tau = \{(\tau-1)b+1, \ldots, \tau b\}$, $L_\tau^b = \sum_{t \in \mathcal{B}_\tau} l_t$, and $l_\tau^b = l_{u_\tau}$. Also, let $\hat{l}_\tau^b = \hat{l}_{u_\tau}$ as generated by algorithm. So, $\widehat{L}_\tau^{\text{obs}} = \sum_{s:s+d_s^b < \tau} \hat{l}_s^b$.

Note that sequence of independently sampled representatives $(u_\tau)_{\tau=1}^{T'}$ determines scheduling behavior of Algorithm 1. Consider filtration $\mathcal{H}_\tau = \sigma(u_1, \ldots, u_{\tau-1})$ for $\tau \in [T']$.

Fact 24 *Suppose $C \geqslant 2$ and $b \geqslant \frac{d_{\max}}{C-1}$. Then, Algorithm 1 never exceeds maximum capacity C.*

Proof This trivially holds because, at any round, the tracking set can contain representative rounds from at most $\lceil d_{\max}/b \rceil$ previous batches, so the size of the tracking set at any point in time is at most $1 + \lceil d_{\max}/b \rceil \leqslant 1 + (C-1) = C$. ∎

Theorem 25 *Suppose that $b \geqslant \frac{d_{\max}}{C-1}$ and learning rates α_τ and β_τ are $\mathcal{H}_\tau$-measurable. Then, for the bandit regime, Algorithm 1 ensures that the expected regret satisfies:*

$$
\frac{\mathfrak{R}_T}{b} \leqslant \mathbb{E}\left[\sum_{\tau=1}^{T'} \left(\sqrt{K}\alpha_\tau + \beta_\tau \sigma_\tau^b \right) + 2\sqrt{K}\alpha_{T'}^{-1} + \log(K)\beta_{T'}^{-1} \right].
$$

And for the full-information regime ($\alpha_\tau = \infty$):

$$
\frac{\mathfrak{R}_T}{b} \leqslant \mathbb{E}\left[\sum_{\tau=1}^{T'} \beta_\tau(\sigma_\tau^b + 1) + \log(K)\beta_{T'}^{-1} \right].
$$

Proof First of all, Fact 24 confirms that for this batch size b, capacity C is never exceeded.

Next, using the fact that learning rates α_τ, β_τ are $\mathcal{H}_\tau$-measurable, we will construct a randomization sequence $(X_\tau)_{\tau=1}^{T'}$ of i.i.d. random variables independent of $\mathcal{H}_{T'+1}$ such that each A_τ^b will be measurable with respect to $\sigma(\mathcal{F}_\tau; X_\tau)$, where $\mathcal{F}_\tau = \sigma(\mathcal{H}_\tau; X_1, \ldots, X_{\tau-1})$ encapsulates all the randomness of the algorithm up to the start of batch τ.

- For the base case $\tau = 1$, $\mathcal{H}_1 = \mathcal{F}_1 = \{\varnothing, \Omega\}$, so $\alpha_1, \beta_1, \widehat{L}_1^{\text{obs}}$, and x_1 are constants. Thus, we can just take $X_1 \sim \text{Unif}(0, 1)$ independent of $\mathcal{H}_{T'+1}$ that would determine A_1^b, i.e. A_1^b would be $\sigma(\mathcal{F}_1; X_1)$ measurable.
- For the induction step, suppose that we constructed the first $\tau \in [T' - 1]$ random variables $(X_\tau)_{s=1}^\tau$. As $\alpha_{\tau+1}$, $\beta_{\tau+1}$, and $\widehat{L}_{\tau+1}^{\text{obs}}$ are $\mathcal{F}_{\tau+1}$-measurable, $x_{\tau+1}$ is as well. Thus, we can just take $X_{\tau+1} \sim \text{Unif}(0, 1)$ independently of both $\mathcal{H}_{T'+1}$ and $(X_s)_{s=1}^\tau$, so that it would determine randomness of $A_{\tau+1}^b$ based on $\mathcal{F}_{\tau+1}$-measurable $x_{\tau+1}$. Then, $A_{\tau+1}$ would be indeed $\sigma(\mathcal{F}_{\tau+1}; X_{\tau+1})$-measurable.

Therefore, we have conditional independence $A_\tau^b \perp u_\tau \mid \mathcal{F}_\tau$ for all $\tau \in [T']$ since A_τ^b is $\sigma(\mathcal{F}_\tau; X_\tau)$-measurable and u_τ is independent of random variables $\{u_s\}_{s=1}^{\tau-1} \cup \{X_s\}_{s=1}^\tau$.

For action $a \in [K]$, let $R_{T',a}^b = \sum_{\tau=1}^{T'}(l_{\tau,A_\tau^b}^b - l_{\tau,a}^b)$. Then, we can write

$$
\begin{aligned}
\mathfrak{R}_T &= \mathbb{E}\left[\sum_{\tau=1}^{T'} \sum_{t \in \mathcal{B}_\tau}(l_{t,A_t} - l_{t,i*})\right] \\
&= \sum_{\tau=1}^{T'} \mathbb{E}\left[\mathbb{E}[\langle e_{A_\tau^b} - e_{i*}, L_\tau^b\rangle \mid \mathcal{F}_\tau]\right] \\
&= \sum_{\tau=1}^{T'} \mathbb{E}\left[\langle \mathbb{E}[e_{A_\tau^b} - e_{i*} \mid \mathcal{F}_t], L_\tau^b\rangle\right] \\
&\overset{(a)}{=} \sum_{\tau=1}^{T'} \mathbb{E}\left[\langle \mathbb{E}[e_{A_\tau^b} - e_{i*} \mid \mathcal{F}_t], \mathbb{E}[l_\tau^b|\mathcal{F}_t]b\rangle\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{\tau=1}^{T'}(l_{\tau,A_\tau^b}^b - l_{\tau,i*}^b)b\right] \\
&= \mathbb{E}\left[\mathbb{E}[R_{T',i*}^b \mid \mathcal{H}_{T'+1}]\right]b,
\end{aligned}
$$

where (a) uses the fact that $L_\tau^b = \mathbb{E}[l_\tau^b b] = \mathbb{E}[l_\tau^b b|\mathcal{F}_t]$ as u_τ is independent of $\mathcal{F}_t$ and (b) applies conditional independence $A_\tau^b \perp u_\tau|\mathcal{F}_t$.

For the fixed choice of representatives u_τ, our algorithm effectively runs Delayed FTRL from Zimmert and Seldin (2020) over T' rounds with oblivious losses l_τ^b and delays d_τ^b. Hence, we can bound this expected regret conditioned on the choice of representatives $\mathbb{E}[R_{T',i*}^b|\mathcal{H}_{T'+1}]$ via Theorem 1 for T' rounds and loss scales $B_\tau = 1$. For the bandit regime, we have:

$$
\mathbb{E}\left[R_{T',i*}^b \mid \mathcal{H}_{T'+1}\right] \leqslant \sum_{\tau=1}^{T'}\left(\sqrt{K}\alpha_\tau + \beta_\tau \sigma_\tau^b\right) + 2\sqrt{K}\alpha_{T'}^{-1} + \log(K)\beta_{T'}^{-1},
$$

and for the full-information regime:

$$
\mathbb{E}\left[R_{T',i*}^b \mid \mathcal{H}_{T'+1}\right] \leqslant \sum_{\tau=1}^{T'} \beta_\tau(\sigma_\tau^b + 1) + \log(K)\beta_{T'}^{-1}.
$$

By taking expectation, we obtain the stated bound on $\mathfrak{R}_T$. ∎

Lemma 26 *For every batch $\tau \in [T']$, number of outstanding batch delays σ_τ^b is $\mathcal{H}_\tau$-measurable. It almost surely holds that $\sum_{\tau=1}^{T'} \sigma_\tau^b = \sum_{\tau=1}^{T'} d_\tau^b$. Plus, it holds that $\mathbb{E}\left[\sum_{\tau=1}^{T'} d_\tau^b\right] \leqslant \frac{D}{b^2} + T'$.*

Proof The first two statements are trivial. To prove the third one, we write

$$\sum_{\tau=1}^{T'} \mathbb{E}[d_\tau^b] = \sum_{\tau=1}^{T'} \mathbb{E}\left[\lceil \tfrac{u_\tau + d_{u_\tau}}{b} \rceil - \lceil \tfrac{u_\tau}{b} \rceil \right] \leqslant \sum_{\tau=1}^{T'} \mathbb{E}\left[\lceil \tfrac{d_{u_\tau}}{b} \rceil \right] \leqslant \sum_{\tau=1}^{T'} \mathbb{E}\left[d_{u_\tau}/b + 1\right] = \frac{D}{b^2} + T'.$$

∎

Lemma 27 ([Seldin et al. (2014)](), Lemma 8) *For a sequence $(x_t)_{t=1}^T$ on $[0,\infty)$, let $\eta_t = (\sum_{s=1}^t x_s)^{-0.5} \in (0,\infty]$. Then, with the convention that $x_t\eta_t = 0$ when $x_t = 0$, it holds that $\sum_{t=1}^T x_t\eta_t \leqslant 2\eta_T^{-1}$.*

Proof (Theorem 2) In both bandit and full-information regimes, the chosen learning rates are $\mathcal{H}_\tau$-measurable because each σ_τ^b is $\mathcal{H}_\tau$-measurable. Then, in the bandit regime, by Theorem 25, we have

$$\frac{\mathfrak{R}_T}{b} \leqslant \mathbb{E}\left[\sum_{\tau=1}^{T'}\left(\sqrt{K}\alpha_\tau + \beta_\tau\sigma_\tau^b\right) + 2\sqrt{K}\alpha_{T'}^{-1} + \log(K)\beta_{T'}^{-1}\right]$$

$$\overset{(a)}{\leqslant} 4\sqrt{T'K} + 3\mathbb{E}\left[\sqrt{\sum_{\tau\in[T']}\sigma_\tau^b}\right]\sqrt{\log(K)}$$

$$\overset{(b)}{\leqslant} 4\sqrt{T'K} + 3\sqrt{D/b^2 + T'}\sqrt{\log(K)}.$$

where (a) applies Lemma 27 for our choice of the learning rates, (b) applies Jensen's inequality and Lemma 26. Consequently, the expected regret for the bandit regime satisfies:

$$\mathfrak{R}_T \leqslant \left(4\sqrt{T'K} + 3\sqrt{(D/b^2 + T')\log(K)}\right)b$$

$$= 4\sqrt{\lceil T/b \rceil b^2 K} + 3\sqrt{(D + \lceil T/b \rceil b^2)\log(K)}$$

$$\leqslant 8\sqrt{TbK} + 3\sqrt{D\log(K)} + 6\sqrt{Tb\log(K)}$$

$$\leqslant 14\sqrt{TbK} + 3\sqrt{D\log(K)}.$$

Similarly, in the full-information regime, by Theorem 25, we have

$$\frac{\mathfrak{R}_T}{b} \leqslant \mathbb{E}\left[\sum_{\tau=1}^{T'}\beta_\tau(\sigma_\tau^b + 1) + \log(K)\beta_{T'}^{-1}\right]$$

$$\overset{(c)}{\leqslant} 3\mathbb{E}\left[\sqrt{\sum_{\tau\in[T']}(\sigma_\tau^b + 1)}\right]\sqrt{\log(K)}$$

$$\overset{(d)}{\leqslant} 3\sqrt{D/b^2 + 2T'}\sqrt{\log(K)}.$$

where (c) applies Lemma 27 for our choice of the learning rates (d) applies Jensen's inequality and Lemma 26. Consequently, the expected regret for the full-information regime satisfies:

$$\mathfrak{R}_T \leqslant \left(3\sqrt{D/b^2 + 2T'}\sqrt{\log(K)}\right)b$$

$$= 3\sqrt{(D + 2\lceil T/b \rceil b^2)\log(K)}$$

$$\leqslant 12\sqrt{Tb\log(K)} + 3\sqrt{D\log(K)}.$$

∎

D.1. Application to Fixed Delays

In this subsection, we consider Delay Scheduling under the assumption of fixed delays, which are all equal to the same $0 \leqslant d \leqslant T$. As the player knows d before the start of each round[8], the most restrictive scheduling framework to consider in this setting is Clairvoyant Non-Preemptive.

Proof (Theorem 3) As per the comment above, it will suffice to consider the clairvoyant preemptive framework for the lower bound and the clairvoyant non-preemptive one for the upper bound. See Theorem 28 for the lower bound. The upper bound follows from Theorem 2 by applying non-preemptive Algorithm 1 with batch size $b = \max\{1, \lceil \frac{d}{C-1} \rceil\} \geqslant \lceil \frac{d_{\max}}{C-1} \rceil$. Then, for the bandit regime, we have

$$\mathfrak{R}_T = O(\sqrt{TbK + D \log(K)}) = O(\sqrt{TK(1 + d/C) + Td \log(K)}),$$

and for the full-information regime:

$$\mathfrak{R}_T = O(\sqrt{(Tb + D)\log(K)}) = O(\sqrt{T(1 + d/C + d)\log(K)}) = O(\sqrt{T(d+1)\log(K)}).$$

■

Theorem 28 (Fixed delays lower bound) *For $K \leqslant \lfloor \frac{CT}{d+1} \rfloor$ in the bandit regime, the minimax regret of Delay Scheduling with fixed delays is of the order*

$$\Omega\left(\sqrt{TK(1 + \tfrac{d}{C})} + \sqrt{Td \log(K)}\right).$$

And for arbitrary K in the full-information regime, regret is of the order

$$\Omega\left(\sqrt{T(d+1)\log(K)}\right).$$

Proof We begin with the full-information case. Since the regret of Delay Scheduling with fixed delays is lower bounded by that of Delayed Online Learning with the same delays, the minimax result of Weinberger and Ordentlich (2002) for the full-information regime implies that

$$\mathfrak{R}_T = \Omega\left(\sqrt{T(d+1)\log(K)}\right).$$

To derive a lower bound on the regret in the bandit regime, we reduce from both Delayed Bandits with fixed delays and Label-Efficient Bandits. Notably, for feedback from round t to be observed, it must satisfy $t \in S_\tau^1$ for all $\tau \in \{t, t+1, \ldots, t+d\}$. Consequently, with probability one, we have

$$\sum_{t=1}^T Z_t(d+1) \leqslant \sum_{t=1}^T \sum_{\tau=t}^{\min\{t+d,T\}} \mathbb{I}(t \in S_\tau^1) = \sum_{t=1}^T |S_t^1| \leqslant CT.$$

Therefore, the player can observe losses from no more than $M = \lfloor \frac{CT}{d+1} \rfloor$ different rounds. Note that $K \leqslant M$ by assumption. Thus, from Theorem 13 and 12 for the bandit regime, we have

$$\mathfrak{R}_T = \Omega\left(\max\left\{\sqrt{\tfrac{T^2K}{M}}, \sqrt{TK} + \sqrt{Td \log(K)}\right\}\right) = \Omega\left(\sqrt{TK(1 + \tfrac{d}{C})} + \sqrt{Td \log(K)}\right),$$

via the reductions from Label Efficient Bandits and Delayed Bandits. ■

Remark: The regret in Theorem 28 is already linear for $K = \lfloor \frac{CT}{d+1} \rfloor$. This shows that considering $K \geqslant \frac{CT}{d+1}$ is unnecessary, as regret remains linear.

---

8. We assume that d is known to the player at the start of the game. Otherwise, it can be inferred within the first d rounds, during which no feedback is received.

Appendix E. General Scheme for Scheduling and Learning

In this section, we prove Theorem 7 and, as a corollary, present Theorem 29, which replaces dependence on $\mathcal{W}_t$ with d_t.

Proof (Theorem 7) To show that $\mathcal{A}$ is an observation-independent delay scheduling algorithm, we will construct a randomization sequence $(X_t^{\mathcal{A}})_{t=1}^T$ of i.i.d. random variables satisfying Theorem 6.

- For the base case $t = 1$, $\mathcal{F}_1 = \{\varnothing, \Omega\}$, so $\alpha_1, \beta_1, \widehat{L}_1^{\text{obs}}$, and x_1 would be constants. Thus, we can just take $X_1^{\mathcal{A}} \sim \text{Unif}(0,1)$ independent of the sequence $(X_t^{\mathcal{S}})_{t=1}^T$ that would determine A_1, i.e. A_1 would be $\mathcal{F}_t^{\mathcal{A}} = \sigma(\mathcal{F}_1; X_1^{\mathcal{A}})$ measurable.
- For the induction step, suppose that we constructed the first $t \in [T-1]$ random variables $(X_s^{\mathcal{A}})_{s=1}^t$. As $\alpha_{t+1}, \beta_{t+1}$, and $\widehat{L}_{t+1}^{\text{obs}}$ are $\mathcal{F}_{t+1}$-measurable, x_{t+1} is as well. Thus, we can just take $X_{t+1}^{\mathcal{A}} \sim \text{Unif}(0,1)$ independently of both $(X_s^{\mathcal{S}})_{s=1}^T$ and $(X_s^{\mathcal{A}})_{s=1}^t$, so that it would determine randomness of A_{t+1} based on $\mathcal{F}_{t+1}$-measurable x_{t+1}. Then, A_{t+1} would be indeed $\mathcal{F}_{t+1}^{\mathcal{A}} = \sigma(\mathcal{F}_{t+1}; X_{t+1}^{\mathcal{A}})$ -measurable.

Thus, algorithm $\mathcal{A}$ can indeed be formalized as an observation-independent delay scheduling algorithm. Moreover, we have conditional independence $A_t \perp Z_t \mid \mathcal{F}_t$ for all $t \in [T]$ since A_t is $\sigma(\mathcal{F}_t; X_t^{\mathcal{A}})$-measurable and Z_t is $\sigma(\mathcal{F}_t; X_t^{\mathcal{S}})$-measurable, with $X_t^{\mathcal{A}}$ and $X_t^{\mathcal{S}}$ being independent random variables.

Let $z_t = Z_t/p_t$ and $e_t = \mathbb{I}(|S_t^0| < C)$, so that $\mathbb{E}[z_t \mid \mathcal{F}_t] = e_t$. Note that rounds where capacity is exceeded can be forfeited for a price of 1 per round in the regret bound, as follows:

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^T (l_{t,A_t} - l_{t,i*})\right]$$
$$\leqslant \mathbb{E}\left[\sum_{t=1}^T (e_t l_{t,A_t} - e_t l_{t,i*})\right] + \mathbb{E}\left[\sum_{t=1}^T (1 - e_t)\right]. \tag{4}$$

For the second term in (4), we have

$$\mathbb{E}\left[\sum_{t=1}^T (1 - e_t)\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}(|S_t^0| = C)\right] = \sum_{t=1}^T \mathbb{P}(|S_t^0| = C).$$

To analyze the first term in (4), let $\widetilde{l}_t = z_t l_t$ and $\widetilde{L}_t = \sum_{s=1}^{t-1} \widetilde{l}_s$. Then, write

$$\mathbb{E}[\widetilde{l}_t] = \mathbb{E}\left[\mathbb{E}[z_t l_t \mid \mathcal{F}_t]\right] = \mathbb{E}\left[\mathbb{E}[z_t \mid \mathcal{F}_t] l_t\right] = \mathbb{E}[e_t l_t].$$

Moreover, using the fact that Z_t and A_t are independent when conditioned on $\mathcal{F}_t$, we have

$$\begin{aligned}
\mathbb{E}[\widetilde{l}_{t,A_t}] &= \mathbb{E}\left[\mathbb{E}[\langle \boldsymbol{e}_{A_t}, l_t\rangle z_t \mid \mathcal{F}_t]\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\mathbb{E}[\langle \boldsymbol{e}_{A_t}, l_t\rangle \mid \mathcal{F}_t] \mathbb{E}[z_t | \mathcal{F}_t]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\langle \boldsymbol{e}_{A_t}, l_t\rangle \mid \mathcal{F}_t] e_t\right] \\
&\overset{(b)}{=} \mathbb{E}\left[e_t \langle \boldsymbol{e}_{A_t}, l_t\rangle\right] \\
&= \mathbb{E}[e_t l_{t,A_t}],
\end{aligned}$$

where (a) follows from the fact $A_t \perp Z_t \mid \mathcal{F}_t$ and (b) applies the defintion of conditional expectation for event $\{|S_t^0| < C\} \in \mathcal{F}_t^{\mathcal{S}} \subseteq \mathcal{F}_t$. This allows us to present the first term in (4) as follows:

$$\mathbb{E}\left[\sum_{t=1}^T (e_t l_{t,A_t} - e_t l_{t,i*})\right] = \mathbb{E}\left[\sum_{t=1}^T (\widetilde{l}_{t,A_t} - \widetilde{l}_{t,i*})\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^T \widetilde{l}_{t,A_t} - \widetilde{L}_{T+1,i*}\,\Big|\,\mathcal{F}_{T+1}^{\mathcal{S}}\right]\right]$$

$$=: \mathbb{E}\left[\widetilde{\mathfrak{R}}_{T,i*}\right].$$

Here, $\widetilde{\mathfrak{R}}_{T,i*}$ represents the expected regret against adversary i^* for Delayed FTRL with time-varying loss scales $B_t = z_t$, losses $\widetilde{l}_t = l_t z_t \in [0, B_t]$, delays d_t, and learning rates α_t, β_t. Importantly, even though the Delay Scheduling Algorithm 2 does not have access to z_t at round $t + d_t$ when no observation occurs ($Z_t = 0$), the reduction to the analysis of Delayed FTRL with time-varying loss scales still holds. This is because $\widetilde{l}_t = 0$ in such cases, and applying a zero loss to FTRL does not affect the algorithm's behavior.

Since $\widetilde{l}_t$, α_t, and β_t are all $\mathcal{F}_{T+1}^{\mathcal{S}}$-measurable, they act as constants when conditioned on $\mathcal{F}_{T+1}^{\mathcal{S}}$. Applying Theorem 1, we can bound the first term in (4) as follows. For the bandit regime:

$$\mathbb{E}\left[\widetilde{\mathfrak{R}}_{T,i*}\right] \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\sqrt{K}\alpha_t z_t^2 + \beta_t z_t \sum_{s\in\mathcal{W}_t} z_s\right) + 2\sqrt{K}\alpha_T^{-1} + \log(K)\beta_T^{-1}\right]$$

and for the full-information regime:

$$\mathbb{E}\left[\widetilde{\mathfrak{R}}_{T,i*}\right] \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\beta_t z_t^2 + \beta_t z_t \sum_{s\in\mathcal{W}_t} z_s\right) + \log(K)\beta_T^{-1}\right].$$

This concludes the proof of Theorem 7. ∎

Theorem 29 *Under the same conditions as Theorem 7, the expected regret is also bounded in the bandit regime as:*

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\sqrt{K}\alpha_t \frac{Z_t}{p_t^2} + \beta_t \frac{Z_t}{p_t} d_t\right) + 2\sqrt{K}\alpha_T^{-1} + \log(K)\beta_T^{-1}\right] + \sum_{t=1}^T \mathbb{P}(|S_t^0| = C).$$

and in the full-information regime as:

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\beta_t \frac{Z_t}{p_t^2} + \beta_t \frac{Z_t}{p_t} d_t\right) + \log(K)\beta_T^{-1}\right] + \sum_{t=1}^T \mathbb{P}(|S_t^0| = C).$$

Proof Let $z_t = Z_t/p_t$ and $e_t = \mathbb{I}(|S_t^0| < C)$, so that $\mathbb{E}[z_t \mid \mathcal{F}_t] = e_t$. Based on the result of Theorem 7, for this theorem to hold, it will suffice to show $\mathbb{E}[\sum_{t=1}^T (\beta_t z_t \sum_{s\in\mathcal{W}_t} z_s)] \leqslant \mathbb{E}[\sum_{t=1}^T \beta_t z_t d_t]$. Using the fact that β_t is non-increasing and $\mathcal{W}_t \subseteq [t-1]$, we can write

$$\mathbb{E}\left[\sum_{t=1}^T \left(\beta_t z_t \sum_{s\in\mathcal{W}_t} z_s\right)\right] \leqslant \mathbb{E}\left[\sum_{t=1}^T \left(z_t \sum_{s\in\mathcal{W}_t} \beta_s z_s\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^T \left(\beta_t z_t \sum_{s=t+1}^{t+d_t} z_s\right)\right]$$

$$= \sum_{t=1}^T \sum_{s=t+1}^{t+d_t} \mathbb{E}[\beta_t z_t z_s].$$

Finally, for every $t, s \in [T]$ such that $s > t$, we have that

$$
\begin{aligned}
\mathbb{E}[\beta_t z_t z_s] &= \mathbb{E}[\mathbb{E}[\beta_t z_t z_s \mid \mathcal{F}_s^\mathcal{S}]] \\
&\overset{(a)}{=} \mathbb{E}[\beta_t z_t \mathbb{E}[z_s \mid \mathcal{F}_s^\mathcal{S}]] \\
&= \mathbb{E}[\beta_t z_t e_s] \\
&\leqslant \mathbb{E}[\beta_t z_t],
\end{aligned}
$$

where (a) follows from the fact that β_t and Z_t are $\mathcal{F}_s^\mathcal{S}$ measurable for $s > t$. In conclusion,

$$
\mathbb{E}\left[\sum_{t=1}^T \left(\beta_t z_t \sum_{s \in \mathcal{W}_t} z_s\right)\right] \leqslant \sum_{t=1}^T \sum_{s=t+1}^{t+d_t} \mathbb{E}[\beta_t z_t z_s] \leqslant \mathbb{E}\left[\sum_{t=1}^T \beta_t z_t d_t\right].
$$

∎

Appendix F. Analysis of Scheduling Policies

Fact 30 *For every $t \in \mathbb{N}$ and $d \in \mathbb{Z}_{\geqslant 0}$, $\mathbb{P}(\widetilde{d}_t \geqslant d) = \min\{1, \frac{C}{(1+\alpha)\nu_t} \cdot \frac{1}{d+1}\}$.*

Proof Sample $\bar{d} \sim \text{Pareto}\,(c, 1)$ for $c = \frac{C}{(1+\alpha)\nu_t}$ so that $\widetilde{d}_t$ has the same distribution as $\lfloor \bar{d} - 1 \rfloor$. Since $\bar{d}$ has cumulative distribution function $F_{\bar{d}}(x) = \mathbb{I}(x > c)(1 - \frac{c}{x})$ and $d \in \mathbb{Z}_{\geqslant 0}$, we can write

$$
\mathbb{P}(\widetilde{d}_t \geqslant d) = \mathbb{P}(\lfloor \bar{d} - 1 \rfloor \geqslant d) = \mathbb{P}(\bar{d} \geqslant d + 1) = \min\left\{1, \frac{c}{d+1}\right\}.
$$

∎

Fact 31 *For sequence $\nu_t = 2H_t$, it holds that $\sum_{s=1}^t \frac{1}{\nu_s(t-s+1)} \leqslant 1$ for every $t \in \mathbb{N}$.*

Proof For every $t \in \mathbb{N}$, we can write

$$
\begin{aligned}
\sum_{s=1}^t \frac{1}{\nu_s(t-s+1)} &= \sum_{s=1}^{\lceil t/2 \rceil} \frac{1}{2H_s(t-s+1)} + \sum_{s=\lceil t/2 \rceil+1}^t \frac{1}{2H_s(t-s+1)} \\
&\leqslant \sum_{s=1}^{\lceil t/2 \rceil} \frac{1}{2\lceil t/2 \rceil} + \sum_{s=\lceil t/2 \rceil+1}^t \frac{1}{2H_{\lceil t/2 \rceil}(t-s+1)} \\
&\leqslant \frac{\lceil t/2 \rceil}{2\lceil t/2 \rceil} + \frac{H_{\lceil t/2 \rceil}}{2H_{\lceil t/2 \rceil}} \\
&= 1.
\end{aligned}
$$

∎

Next, we prove Lemmas 8 and 9 via the Chernoff multiplicative bound.

Proof (Lemma 8) Clairvoyance allows us to run Scheduler 3 with $p_t = \min\{1, \frac{C}{(1+\alpha)\nu_t} \cdot \frac{1}{d_t+1}\}$. Suppose $b_t \sim \text{Ber}(p_t)$ is sampled every round and round t is scheduled if and only if $b_t = 1$ and $|S_t^0| < C$, i.e. $Z_t = b_t \mathbb{I}(|S_t^0| < C)$. Consider arbitrary round t. Due to the structure of this scheduler, the tracking set $S_t^0 = \{s \in [t-1] : s + d_s \geqslant t, Z_s = 1\}$. Thus, we can write

$$
|S_t^0| = \sum_{s=1}^{t-1} Z_s \, \mathbb{I}(s + d_s \geqslant t) \leqslant \sum_{s=1}^{t-1} b_s \, \mathbb{I}(d_s \geqslant t - s) =: \widetilde{\sigma}_t.
$$

The rest of the proof proceeds by bounding $\mathbb{P}(\widetilde{\sigma}_t \geqslant C)$ via the Chernoff bound. Note that $\widetilde{\sigma}_t$ can be written as a finite sum of independent Bernoulli random variables, because b_s are sampled independently. The choice of p_s and Fact 31 allows us to write

$$\mathbb{E}[\widetilde{\sigma}_t] = \sum_{s=1}^{t-1} p_s \, \mathbb{I}(d_s \geqslant t-s) = \sum_{s=1}^{t-1} \min\left\{1, \tfrac{C}{(1+\alpha)\nu_s} \cdot \tfrac{1}{t-s+1}\right\} \leqslant \tfrac{C}{(1+\alpha)}.$$

Let $\alpha' = \tfrac{C}{\mathbb{E}[\widetilde{\sigma}_t]} - 1 \geqslant \alpha$. Then, the Multiplicative Chernoff bound (e.g., see Theorem 2.3.1 from Vershynin (2018)) grants

$$\mathbb{P}(\widetilde{\sigma}_t \geqslant C) = \mathbb{P}(\widetilde{\sigma}_t \geqslant (1+\alpha')\mathbb{E}[\widetilde{\sigma}_t])$$

$$\leqslant \left(\frac{e^{\alpha'}}{(1+\alpha')^{1+\alpha'}}\right)^{\frac{C}{1+\alpha'}}$$

$$= \exp\left(C\left(\tfrac{\alpha'}{1+\alpha'} - \ln(1+\alpha')\right)\right)$$

$$\overset{(a)}{\leqslant} \exp\left(C\left(\tfrac{\alpha}{1+\alpha} - \ln(1+\alpha)\right)\right)$$

$$\overset{(b)}{\leqslant} \delta,$$

where (a) follows from the fact that function $f(x) = 1 - \tfrac{1}{1+x} - \ln(1+x)$ is decreasing on the domain $(0, \infty)$ and $0 < \alpha \leqslant \alpha'$ and (b) holds by $(\star)$. This concludes the proof. ∎

Proof (Lemma 9) The first part of the theorem follows from the fact that $\widetilde{d}_t$ is independent from S_t^0 and Fact 30. Here, we write

$$\mathbb{E}[Z_t \mid |S_t^0| < C] = \mathbb{E}[\mathbb{I}(\widetilde{d}_t \geqslant d_t)\,\mathbb{I}(|S_t^0| < C) \mid |S_t^0| < C] = \mathbb{E}[\mathbb{I}(\widetilde{d}_t \geqslant d_t)] = \mathbb{P}(\widetilde{d}_t \geqslant d_t).$$

Consider arbitrary $t \in [T]$ and let $\widetilde{\sigma}_t = \sum_{s=1}^{t-1} \mathbb{I}(s + \widetilde{d}_s \geqslant t)$ denote the number of outstanding proxy delays. Since the size of the tracking set S cannot exceed the number of outstanding proxy delays at the start of round t, it is sufficient to verify that $\mathbb{P}(\widetilde{\sigma}_t \geqslant C) \leqslant \delta$.

Note that $\widetilde{\sigma}_t$ can be written as a finite sum of independent Bernoulli random variables $\widetilde{\sigma}_t = \sum_{s=1}^{t-1} \mathbb{I}(\widetilde{d}_s \geqslant t-s)$, because proxy delays $\widetilde{d}_s$ are sampled independently from distributions $\mathfrak{D}_s$. From Facts 30 and 31, we can write

$$\mathbb{E}[\widetilde{\sigma}_t] = \sum_{s=1}^{t-1} \mathbb{P}(\widetilde{d}_s \geqslant t-s) = \sum_{s=1}^{t-1} \min\left\{1, \tfrac{C}{(1+\alpha)\nu_s} \cdot \tfrac{1}{t-s+1}\right\} \leqslant \tfrac{C}{(1+\alpha)}.$$

The rest of the proof follows via the Multiplicative Chernoff bound, exactly as was done in the proof of Lemma 8 above. ∎

Appendix G. Upper Bounds for Clairvoyant or Preemptive Settings: Proofs

To prove Theorems 10 and 11, note that both considered Schedulers 3 and 4 are precommitted and quantified by the same sequence $p_t = \mathbb{P}(\widetilde{d}_t \geqslant d_t) = \min\{1, \tfrac{C}{(1+\alpha)\nu_t} \cdot \tfrac{1}{d_t+1}\} = \mu_t^{-1}$, where $\widetilde{d}_t$ is
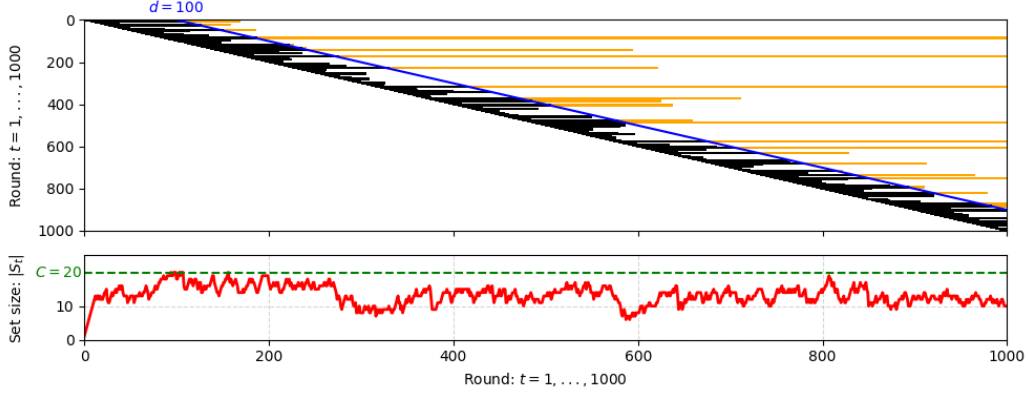
Figure 2　Example of the evolution of the tracking set: $T = 1000$, $C = 20$, $d = 100$. Black intervals $[t,\, t + \min\{d,\, \widetilde{d}_t\}]$ indicate when round t is in the tracking set, while orange intervals $[t + d,\, t + \widetilde{d}_t]$ represent the excess proxy delay for rounds where feedback is observed.

the proxy delay sampled from distribution $\mathfrak{D}_t$. Also, note that p_t (and μ_t) is computable whenever d_t is known, so in clairvoyant frameworks that occurs for every round t during round t, and in non-clairvoyant preemptive frameworks this occurs for every round t with observed feedback during round $t + d_t$. Therefore, as $z_t \neq 0$ if and only if feedback from round t arrived in round $t + d_t$ and d_t was observed, all of these learning rates are computable using only available information.

Additionally, all the considered learning rates are $\mathcal{F}_t^{\mathcal{S}}$-measurable. For Theorem 10 they are even constant, and for Theorem 11 they are determined by the $\mathcal{F}_t^{\mathcal{S}} = \sigma(\widetilde{d}_1, \ldots, \widetilde{d}_{t-1})$.

As we consider $z_t = Z_t/p_t$, note that $\mathbb{E}[z_t] = \mathbb{E}[Z_t \mu_t] \leqslant 1$ and $\mathbb{E}[z_t^2] = \mathbb{E}[Z_t \mu_t^2] \leqslant \mu_t$.

Proof (Theorem 10) From Theorem 29 (modified Theorem 7) for bandits, it follows that

$$
\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{t=1}^{T} \left(\sqrt{K} \alpha_t z_t^2 + \beta_t z_t d_t \right) + 2\sqrt{K} \alpha_T^{-1} + \log(K) \beta_T^{-1} \right] + \sum_{t=1}^{T} \mathbb{P}(|S_t^0| = C).
$$

$$
\overset{(a)}{\leqslant} \sqrt{K} \sum_{t=1}^{T} \alpha_t \mu_t + \sum_{t=1}^{T} \beta_t d_t + 2\sqrt{K} \alpha_T^{-1} + \log(K) \beta_T^{-1} + \delta T
$$

$$
\overset{(b)}{\leqslant} 4\sqrt{K} \sqrt{\sum_{t=1}^{T} \mu_t} + 3\sqrt{\log(K)} \sqrt{\sum_{t=1}^{T} d_t} + \delta T
$$

$$
\leqslant 4\sqrt{K} \sqrt{T + \frac{(1+\alpha)\nu_t}{C}(D + T)} + 3\sqrt{D \log(K)} + \delta T,
$$

where (a) substitutes $\mathbb{E}[z_t] \leqslant 1$ and $\mathbb{E}[z_t^2] \leqslant \mu_t$ and applies Lemma 8, (b) applies Lemma 27 for our choice of the learning rates. Similarly, in the full-information regime, from Theorem 29, we have

$$
\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{t=1}^{T} \left(\beta_t z_t^2 + \beta_t z_t d_t \right) + \log(K) \beta_T^{-1} \right] + \sum_{t=1}^{T} \mathbb{P}(|S_t^0| = C)
$$

$$
\overset{(c)}{\leqslant} \sum_{t=1}^{T} \beta_t (\mu_t + d_t) + \log(K) \beta_T^{-1} + \delta T
$$

$$
\overset{(d)}{\leqslant} 3\sqrt{\log(K)} \sqrt{\sum_{t=1}^{T} (\mu_t + d_t)} + \delta T
$$

$$
\leqslant 3\sqrt{\log(K)} \sqrt{(D + T)\left(1 + \frac{(1+\alpha)\nu_T}{C}\right)} + \delta T,
$$

37

where again (c) substitutes $\mathbb{E}[z_t] \leqslant 1$ and $\mathbb{E}[z_t^2] \leqslant \mu_t$ and applies Lemma 8, (d) applies Lemma 27 for our choice of the learning rates. ∎

Proof (Theorem 11) Note that for each round t, set S_t^1 contains at most C rounds from the set $\mathcal{W}_t \cup \{t\}$ and for each $s \in S_t^1 \subseteq [t]$, it holds that $z_s \leqslant \mu_{\max,t}$ and $d_s \leqslant d_{\max}$. Therefore, our choice of the learning rates uses only information available in non-clairvoyant, preemptive scheduling to guarantee that in the bandit regime:

$$\alpha_t^{-1} \geqslant \sqrt{\textstyle\sum_{s \in [t]} z_s^2}, \qquad \beta_t^{-1} \geqslant \log(K)^{-1/2} \sqrt{\textstyle\sum_{s \in [t]} z_s d_s},$$

and in the full-information regime:

$$\beta_t^{-1} \geqslant \log(K)^{-1/2} \sqrt{\textstyle\sum_{s \in [t]} z_s(z_s + d_s)}.$$

Then, in the bandit regime, from Theorem 29, it follows that

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\textstyle\sum_{t=1}^T \left(\sqrt{K}\alpha_t z_t^2 + \beta_t z_t d_t \right) + 2\sqrt{K}\alpha_T^{-1} + \log(K)\beta_T^{-1} \right] + \textstyle\sum_{t=1}^T \mathbb{P}(|S_t^0| = C).$$

$$\overset{\text{(a)}}{\leqslant} \mathbb{E}\left[4\sqrt{K}\sqrt{\textstyle\sum_{t=1}^T z_t^2 + C\mu_{\max}^2} + 3\sqrt{\log(K)}\sqrt{\textstyle\sum_{t=1}^T z_t d_t + C d_{\max}\mu_{\max}} \right] + \delta T$$

$$\overset{\text{(b)}}{\leqslant} 4\sqrt{K}\sqrt{\textstyle\sum_{t=1}^T \mu_t + C\mu_{\max}^2} + 3\sqrt{\log(K)}\sqrt{\textstyle\sum_{t=1}^T d_t + C d_{\max}\mu_{\max}} + \delta T$$

$$\leqslant 4\sqrt{K}\sqrt{T + \tfrac{(1+\alpha)\nu_T}{C}(D + T)} + 3\sqrt{D\log(K)}$$

$$+ 7\sqrt{C\mu_{\max}(K\mu_{\max} + \log(K)d_{\max})} + \delta T$$

where (a) applies Lemma 9 and Lemma 27 for our choice of the learning rates, (b) applies Jensen's inequality and substitutes $\mathbb{E}[z_t] \leqslant 1$ and $\mathbb{E}[z_t^2] \leqslant \mu_t$. Similarly, in the full-information regime, from Theorem 29, we have

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\textstyle\sum_{t=1}^T \left(\beta_t z_t^2 + \beta_t z_t d_t \right) + \log(K)\beta_T^{-1} \right] + \textstyle\sum_{t=1}^T \mathbb{P}(|S_t^0| = C)$$

$$\overset{\text{(c)}}{\leqslant} \mathbb{E}\left[3\sqrt{\log(K)}\sqrt{\textstyle\sum_{t=1}^T z_t(z_t + d_t) + C\mu_{\max}(\mu_{\max} + d_{\max})} \right] + \delta T$$

$$\overset{\text{(d)}}{\leqslant} 3\sqrt{\log(K)}\sqrt{\textstyle\sum_{t=1}^T (\mu_t + d_t) + C\mu_{\max}(\mu_{\max} + d_{\max})} + \delta T$$

$$\leqslant 3\sqrt{\log(K)}\sqrt{(D + T)(1 + \tfrac{(1+\alpha)\nu_T}{C})} + 3\sqrt{C\mu_{\max}\log(K)(\mu_{\max} + d_{\max})} + \delta T,$$

where (c) applies Lemma 9 and Lemma 27 for our choice of the learning rates, (d) applies Jensen's inequality and substitutes $\mathbb{E}[z_t] \leqslant 1$ and $\mathbb{E}[z_t^2] \leqslant \mu_t$. ∎

Appendix H. Non-Clairvoyant Non-Preemptive Delay Scheduling

For completeness, we also consider the Non-Clairvoyant and Non-Preemptive Delay Scheduling. The restrictions of this framework put the player at a great disadvantage. For instance, any unlucky scheduling of a round with an $\Omega(T)$-long delay effectively removes one unit of capacity from the

player for the rest of the game. Thus, in the absence of preemption, runtime information about delays in this framework is even more limited than in the Non-Clairvoyant Preemptive framework, for which we already require knowledge of $d_{\max}$.

Nonetheless, given prior knowledge of either T and D or $d_{\max}$ (which could be the vacuous upper bound $d_{\max} = T$), we can derive several upper bounds on expected regret using the Scheduling and Batching techniques from Sections 4 and 3, as stated in Corollaries 33 and 34, respectively.

We first prove Theorem 32, from which Corollary 33 directly follows.

Theorem 32 *Suppose that Algorithm 2 is run with Scheduler 3 with fixed probabilities $p_t = p$ and learning rates $\alpha_t = \alpha, \beta_t = \beta$. Then, in the bandit regime, we have:*

$$\mathfrak{R}_T \leqslant \sqrt{K}T\alpha p^{-1} + \beta D + 2\sqrt{K}\alpha^{-1} + \log(K)\beta^{-1} + \tfrac{pD}{C},$$

and in the full-information regime:

$$\mathfrak{R}_T \leqslant \beta T p^{-1} + \beta D + \log(K)\beta^{-1} + \tfrac{pD}{C}.$$

Proof First of all, note that constant learning rates are clearly $\mathcal{F}_t^{\mathcal{S}}$-measurable. The Bernoulli scheduler (Scheduler 3) is also clearly quantified by the sequence $p_t = p$. Therefore, we can apply Theorem 7. Also, since $\mathbb{E}[Z_t] \leqslant p$, applying Markov's inequality gives, for every $t \in [T]$,

$$\mathbb{P}(|S_t^0| = C) \leqslant \mathbb{P}\left(\sum_{s \in \mathcal{W}_t} Z_t \geqslant C\right) \leqslant \frac{p\sigma_t}{C}.$$

Then, applying Theorem 29 (modification of Theorem 7) in the bandit regime, grants us

$$\begin{aligned}
\mathfrak{R}_T &\leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\sqrt{K}\alpha \tfrac{Z_t}{p^2} + \beta \tfrac{Z_t}{p} d_t\right) + 2\sqrt{K}\alpha^{-1} + \log(K)\beta^{-1}\right] + \sum_{t=1}^T \mathbb{P}(|S_t^0| = C) \\
&\leqslant \sqrt{K}T\alpha/p + \beta \sum_{t=1}^T d_t + 2\sqrt{K}\alpha^{-1} + \log(K)\beta^{-1} + \sum_{t=1}^T \tfrac{p\sigma_t}{C} \\
&= \sqrt{K}T\alpha p^{-1} + \beta D + 2\sqrt{K}\alpha^{-1} + \log(K)\beta^{-1} + \tfrac{pD}{C}.
\end{aligned}$$

and in the full-information regime:

$$\begin{aligned}
\mathfrak{R}_T &\leqslant \mathbb{E}\left[\sum_{t=1}^T \left(\beta \tfrac{Z_t}{p^2} + \beta \tfrac{Z_t}{p} d_t\right) + \log(K)\beta^{-1}\right] + \sum_{t=1}^T \mathbb{P}(|S_t^0| = C) \\
&\leqslant \beta T/p + \beta \sum_{t=1}^T d_t + \log(K)\beta^{-1} + \tfrac{p\sigma_t}{C} \\
&\leqslant \beta T p^{-1} + \beta D + \log(K)\beta^{-1} + \tfrac{pD}{C}.
\end{aligned}$$

∎

Corollary 33 (Scheduling approach with known T, D) *In the bandit regime, suppose $C \leqslant \frac{D+T}{\sqrt{TK}}$. Setting parameters as $p = \sqrt[3]{\frac{C^2 TK}{(D+T)^2}}$, $\alpha = \sqrt[3]{\frac{C\sqrt{K}}{T(D+T)}}$, and $\beta = \sqrt{\frac{\log(K)}{D+T}}$, the algorithm in Theorem 32 achieves a regret bound of*

$$\mathfrak{R}_T \leqslant 4\sqrt[3]{\frac{T(D+T)K}{C}} + 2\sqrt{(D+T)\log(K)}.$$

In the full-information regime, suppose $C \leqslant \frac{T}{\sqrt{(D+T)\log(K)}}$. Setting parameters as $p = \sqrt[3]{\frac{C^2 T \log(K)}{(D+T)^2}}$ and $\beta = \sqrt[3]{\frac{C \log^2(K)}{T(D+T)}}$, the algorithm in Theorem 32 achieves a regret bound of:

$$\mathfrak{R}_T \leqslant 3\sqrt[3]{\frac{T(D+T)\log(K)}{C}} + \sqrt{(D+T)\log(K)}.$$

Proof Corollary 33 restricts the capacity in order to ensure that the chosen probability p remains within the interval $(0, 1]$. Nevertheless, an algorithm designed for a smaller capacity can be trivially simulated on a larger one. Moreover, in both the bandit and full-information regimes, as the capacity approaches its restriction, the stated regret bounds converge to those of Delayed Online Learning.

To establish these bounds on expected regret, substitute the chosen values of p, α, and β into the bounds from Theorem 32 for both the bandit and full-information regimes. ∎

Corollary 34 (Batching approach with known $d_{\max}$) *Suppose $C \geqslant 2$ and $d_{\max} > 0$. Algorithm 1 with batch size $b = \lceil \frac{d_{\max}}{C-1} \rceil$ and learning rates from Theorem 2 guarantees that*

$$\mathfrak{R}_T \leqslant 28\sqrt{\frac{T d_{\max} K}{C-1}} + 3\sqrt{D \log(K)}$$

in the bandit regime and

$$\mathfrak{R}_T \leqslant 24\sqrt{\frac{T d_{\max} \log(K)}{C-1}} + 3\sqrt{D \log(K)}$$

in the full-information regime.

Proof Follows directly from Theorem 2. ∎

Appendix I. Delay Scheduling under the Expectation-Capacity Constraint

In this section, we examine a variant of Delay Scheduling in which the expected size of the tracking set in each round is constrained by the expectation-capacity C_E. Provided prior knowledge of $\log(T)$, we derive the regret bounds presented in Table 8.

Delay Scheduling under the expectation-capacity constraint for $C_E > 0$		
Framework	Regime	Regret Bounds
Clairvoyant Non-Preemptive	Bandit	$O\left(\sqrt{TK + \frac{\log(T)}{C_E}(D+T)K} + \sqrt{D\log(K)} \right)$
	Full-info	$O\left(\sqrt{(1 + \frac{\log(T)}{C_E})(D+T)\log(K)} \right)$
Non-Clairvoyant Preemptive	Bandit	$O\left(\sqrt{TK + \frac{\log(T)}{C_E}(D+T)K} + \sqrt{D\log(K)} \right) + \tilde{O}\left(d_{\max}(1 + \frac{K}{C_E}) \right)$
	Full-info	$O\left(\sqrt{(1 + \frac{\log(T)}{C_E})(D+T)\log(K)} \right) + \tilde{O}\left(d_{\max}(1 + \frac{1}{C_E}) \right)$

Table 8   Regret upper bounds for Delay Scheduling under the expectation-capacity constraint.

Proof: For Theorems 10 and 11 to hold, it suffices for the normalization sequence (ν_t) to be non-decreasing and satisfy $\nu_t \geqslant 2H_t$ for all $t \in [T]$. Then, the results in Table 8 follow directly from these corollaries if we were to run their corresponding algorithms with capacity $C = \lceil \max\{3, K\} \log(T) \rceil$ in the bandit regime or capacity $C = \lceil \max\{3, \log(K)\} \log(T) \rceil$ in the full-information regime, Chernoff parameter $\alpha = 1$, and sequence $\nu_t = 2H_t \max\{1, C/C_E\}$, while considering $\delta = T^{-0.5}$, assuming that the expectation capacity-constraint is satisfied for this choice of parameters.

It remains to verify this constraint. Fix arbitrary $t \in [T]$. Following a similar argument as in the proof of Lemma 9 and applying Fact 31, we obtain

$$\mathbb{E}[|S_t^1|] \leqslant \sum_{s=1}^{t} \mathbb{P}(\tilde{d}_s \geqslant t-s) = \sum_{s=1}^{t} \min\left\{1, \frac{C}{(1+\alpha)\nu_s(t-s+1)}\right\} \leqslant \frac{C}{(1+\alpha)\max\{1, C/C_E\}} < C_E.$$

Thus, the expectation-capacity constraint holds for every round. ∎

Additionally, we derive matching lower bounds, up to logarithmic factors, by analyzing the fixed delays scenario and applying reduction techniques analogous to those used in Theorem 28.

Theorem 35 *Suppose $C_E \geqslant \frac{(d+1)K}{T}$. Then, in the bandit regime, the minimax regret of Delay Scheduling with fixed delays under the expectation-capacity constraint is of the order*

$$\Omega\left(\sqrt{TK\left(1 + \frac{d+1}{C_E}\right)} + \sqrt{Td \log(K)}\right).$$

And in the full-information regime, regret is of the order

$$\Omega\left(\sqrt{\left(1 + \frac{1}{C_E}\right)T(d+1)\log(K)}\right).$$

Proof By closely examining the proof of Theorem 30 in Audibert and Bubeck (2010) (see Theorem 13 here), we note that the lower bound for label-efficient settings remains valid even when the expected number of queries is at most M. In particular, equation (30) of their proof bounds $\mathbb{E}_0[\sum_{t=1}^{T} \mathbb{I}(Z_t = 1)]$ by M.

In the Delay Scheduling game, for feedback from round t to be observed, it must satisfy $t \in S_\tau^1$ for all $\tau \in \{t, t+1, \ldots, t+d\}$. Consequently, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} Z_t(d+1)\right] \leqslant \sum_{t=1}^{T} \mathbb{E}\left[|S_t^1|\right] \leqslant C_E T.$$

Therefore, in expectation, the player observes losses from no more than $M = \frac{C_E T}{d+1}$ different rounds. Note that $K \leqslant M$ by assumption.

As in the proof of Theorem 28, we use reductions from both Delayed Online Learning with with fixed delays and Label-Efficient learning, in order to derive lower bounds on the regret for both regimes. From Theorems 13 and 12 for the bandit regime, we have

$$\mathfrak{R}_T = \Omega\left(\max\left\{\sqrt{\frac{T^2 K}{M}}, \sqrt{TK} + \sqrt{Td \log(K)}\right\}\right) = \Omega\left(\sqrt{TK\left(1 + \frac{d+1}{C_E}\right)} + \sqrt{Td \log(K)}\right).$$

And, from Theorem 13 and Weinberger and Ordentlich (2002) for the full-information regime, we have

$$\mathfrak{R}_T = \Omega \left(\max \left\{ \sqrt{\frac{T^2 \log(K)}{M}}, \sqrt{T(d+1)\log(K)} \right\} \right) = \Omega \left(\sqrt{(1 + \frac{1}{C_E})T(d+1)\log(K)} \right).$$

∎