Adapting to Delays and Data in Adversarial Multi-Armed Bandits

András György [1] Pooria Joulani [1]

Abstract

We consider the adversarial multi-armed bandit problem under delayed feedback. We analyze variants of the Exp3 algorithm that tune their step size using only information (about the losses and delays) available at the time of the decisions, and obtain regret guarantees that adapt to the observed (rather than the worst-case) sequences of delays and/or losses. First, through a remarkably simple proof technique, we show that with proper tuning of the step size, the algorithm achieves an optimal (up to logarithmic factors) regret of order $\sqrt{\log(K)(TK+D)}$ both in expectation and in high probability, where K is the number of arms, T is the time horizon, and D is the cumulative delay. The high-probability version of the bound, which is the first high-probability delay-adaptive bound in the literature, crucially depends on the use of implicit exploration in estimating the losses. Then, following Zimmert and Seldin (2019), we extend these results so that the algorithm can "skip" rounds with large delays, resulting in regret bounds of order $\sqrt{TK\log(K)}+|R|+\sqrt{D_{\bar{R}}\log(K)}$, where R is an arbitrary set of rounds (which are skipped) and $D_{\bar{R}}$ is the cumulative delay of the feedback for other rounds. Finally, we present another, data-adaptive (AdaGrad-style) version of the algorithm for which the regret adapts to the observed (delayed) losses instead of only adapting to the cumulative delay (this algorithm requires an a priori upper bound on the maximum delay, or the advance knowledge of the delay for each decision when it is made). The resulting bound can be orders of magnitude smaller on benign problems, and it can be shown that the delay only affects the regret through the loss of the best arm.

1. Introduction

The multi-armed bandit problem is a canonical model for sequential decision making with limited feedback. In this model a learner makes a sequence of actions. After every action, the learner immediately observes the loss corresponding to its action. On the other hand, in many practical applications of bandit algorithms, the loss feedback to the algorithm may be severely delayed. For example, in recommender systems, including display advertisement, content optimization for websites, or content recommendation for paid services, often several new recommendations need to be made before a user can even react to the recommendation they received, not to mention the time often needed to propagate the feedback (observation) to the decision making system (Li et al., 2010; Dudik et al., 2011; Chapelle, 2014), or the performance metrics may be delayed by design (e.g., user engagement in a week, see, e.g., Mann et al., 2019; Vernade et al., 2020). Other types of applications include interactive multi-agent learning systems (Cesa-Bianchi et al., 2019), where the feedback is not immediately available due to communication delays, or distributed optimization, where gradient computations are performed with various delays (Agarwal and Duchi, 2011), to mention a few.

In this paper we consider the adversarial version of the bandit problem with delayed feedback and an oblivious adversary. Given a set of K actions and a time horizon T, it is well known that the worst-case regret achievable by a learner in the non-delayed setting is of order $\sqrt{KT}$ (Audibert and Bubeck, 2009). The delayed setting was perhaps first considered by Neu et al. (2010; 2014), who showed that in case every feedback is delayed by a constant d, the Exp3 algorithm (Auer et al., 2002) achieves a regret of $O(\sqrt{dKT\log(K)})$. This result was extended to the general partial monitoring setting by Joulani et al. (2013). The next important step was made by Cesa-Bianchi et al. (2019), who showed that the effect of the delay and the number of arms is in fact not intertwined: they proved that the worst-case regret is at least $\Omega(\max\{\sqrt{KT}, \sqrt{dT\log(K)}\})$ (for $d \leq T/\log K$), and that the Exp3 algorithm achieves this (up to a logarithmic factor). These bounds show that, at least in the case of fixed delays, it is possible to achieve a regret that scales with the cumulative delay $D = dT$. In the full information case, Quanrud and Khashabi (2015) were the first to show that it is possible to achieve a regret that

[1] DeepMind, London, UK. Correspondence to: András György <agyorgy@deepmind.com>, Pooria Joulani <pjoulani@deepmind.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

scales with the cumulative delay (defined as the sum of the delays) in case of non-uniform delays, that is, when the delay for the different time steps can be different, and showed an optimal regret of order $O(\sqrt{(D+T)\log(K)})$, where D is the sum of the (arbitrary, not necessarily equal) delays. This result was strengthened by Joulani et al. (2016), who showed that this can be done in a fully adaptive way, without prior knowledge on the delays, and without resorting to the doubling trick.

Thinking along similar lines, Cesa-Bianchi et al. (2019) posed the question whether a regret growing with the cumulative delay is achievable for arbitrary delays in the bandit setting, more precisely, if a $\sqrt{KT}+\sqrt{D\log(K)}$ regret is achievable. Recently, Thune et al. (2019) gave a natural delayed variant of the Exp3 algorithm which achieves essentially the same bound but with an oracle tuning depending on the cumulative delay D, or with the advance knowledge of the delays at the time of the action and using a doubling trick. At the same time, Bistritz et al. (2019) claimed that a properly tuned version of the delayed Exp3 algorithm combined with a doubling trick can achieve the desired bound depending on the cumulative delay (albeit with a sub-optimal dependence on K, i.e., K rather than $\sqrt{K}$), but their proof is unfortunately incorrect.[1]

More recently, Zimmert and Seldin (2019) achieved an optimal $O(\sqrt{KT}+\sqrt{D\log(K)})$ bound (optimal in terms of the cumulative delay D) with an anytime algorithm that requires no advance knowledge about the delays. On the other hand, Thune et al. (2019) pointed out that the scaling of the regret with the cumulative delay D can be quite pessimistic in certain cases (e.g., if the feedback of the first round is missing until the very end but no other feedback is delayed, the resulting cumulative delay is $D=T$, which seems an unreasonably large price to pay in the regret bound in this case), and proposed to "skip" rounds with excessive delays. This leads to regret bounds where the $\sqrt{D\log(K)}$ term is replaced with $|R|+\sqrt{D_{\bar{R}}\log(K)}$, where R is an arbitrary set of rounds (which are skipped) and $D_{\bar{R}}$ is the cumulative delay of the feedback for other rounds. While they achieved this bound using the advance knowledge of the delay for every prediction made, the method of Zimmert and Seldin (2019) achieves this goal under the natural assumption that the delays become known when the feedback arrives.

The analysis of Zimmert and Seldin (2019) (like ours) uses a follow-the-regularized-leader (FTRL) approach, but like the other papers mentioned above, requires specializing the FTRL analysis (respectively, the analysis of Exp3 in the works of Cesa-Bianchi et al., 2019; Thune et al., 2019) to handle the effect of delays on the updates, and hence repeating the main analysis steps from scratch. In addition, their modified FTRL analysis is specialized to a relatively complicated regularizer to avoid the $\sqrt{\log(K)}$ term in the Exp3 bound (hence their update cannot be computed in closed form), leaving the simple Exp3 case unattended.

1.1. Contributions

In this paper we are concerned with similar, fully delay-adaptive methods, based on different versions of the Exp3 algorithm, and derive several novel results for the adversarial bandit problem with delayed feedback:

(i) Using a remarkably simple proof technique, we derive the first proper step-size tuning of the delayed Exp3 algorithm, called the Delay-Adaptive Exp3 (DAda-Exp3) algorithm, which only uses information available to the algorithm at the time of each decision, and achieves the optimal (up to a logarithmic factor) regret rate $\sqrt{\log(K)(KT+D)}$ (Section 3). Compared to the results of Zimmert and Seldin (2019), our bounds are a logarithmic factor worse, which is due to the fact that our method is based on the simpler Exp3 algorithm. In return, our analysis is much simpler.

(ii) Combined with the implicit exploration technique of Neu (2015b) in estimating the losses, we also derive a version of Exp3 that achieves the *first fully delay-adaptive high-probability regret bound* in the literature (Section 4). While the latter bound also depends on the maximum delay $d^\star = \max_{t\in[T]} d_t$, this can be avoided by using the skipping technique (Zimmert and Seldin, 2019): for any set of time steps R to be skipped, the resulting variants of the above Exp3 algorithms achieve the optimal (up to a logarithmic factor) regret of order $\sqrt{KT\log(K)}+|R|+\sqrt{D_{\bar{R}}\log(K)}$ in expectation and with high probability, respectively (Section 5).

(iii) The performance of a learning algorithm can be significantly better than the minimax regret for nice problem instances. To take advantage of such situations in the delayed case, we develop a new version of the DAda-Exp3 algorithm, called Delay- and Data-Adaptive Exp3 (DeDa-Exp3), which is the first algorithm for the delayed setting whose expected regret scales with the actual (rather than the worst-case possible) losses, also improving our bound for DAda-Exp3 (Section 6). DeDa-Exp3 is based on a combination of our analysis technique intro-

[1] From a technical perspective, the difficulties in the works of Cesa-Bianchi et al. (2019), Thune et al. (2019), and Bistritz et al. (2019) arise because the analysis technique they adopt requires bounding a hard-to-control "drift" term. Thune et al. (2019) control this through a bound that requires the step-size to be diminished using the knowledge of the total or upcoming delay (this is also used by Cesa-Bianchi et al., 2019 as they consider the case of fixed, known delays), while Bistritz et al. (2019) also need to bound a similar term (however, in their derivation, namely in their Eq. 36, they incorrectly drop a hard-to-bound term corresponding to the ratio of the true action selection probability and the action selection probability computed right before the arrival of the corresponding feedback, which may depend on additional, delayed, loss values).

duced for DAda-Exp3 and the data- and delay-adaptive full-information algorithm of Joulani et al. (2016). As a simple example of the resulting bounds, the algorithm achieves a regret of order $d^\star + \sqrt{\log(K)\left(d^\star L_{T,A^*} + \sum_{i=1}^K L_{T,i}\right)}$, where $L_{T,i}$ denotes the cumulative loss of action i and A^* denotes the optimal arm in T time steps. This bound is essentially the same as the best data-dependent bound for Exp3 (of order $\sqrt{\log(K)\sum_{i=1}^K L_{T,i}}$, as follows from Neu, 2015a) and some extra delay term, where the effect of the delay depends only on the loss of the best arm but not of the other arms.

On the technical side, the novelty in our analysis can be summarized as follows:

(i) We provide a direct reduction from the regret of the delayed-feedback bandit problem to that of a non-delayed (full-information) problem. As such, in contrast to previous work, our analysis does not need to modify the proof of the basic non-delayed exponential-weights algorithm; instead, we only need to bound the "drift term" arising from the reduction. Such a reduction has proved beneficial in the full-information setting (Joulani et al., 2016), but so far has not been found in the bandit setting (Zimmert and Seldin, 2019, Section 1), partially because the delays change the order in which the losses are observed (Thune et al., 2019, Appendix A). In addition to considerably simplifying the analysis of DAda-Exp3 (e.g., compared to what could be obtained for Exp3 following the proof technique of Zimmert and Seldin, 2019), this reduction is crucial for adopting the technique of Joulani et al. (2016) to the bandit setting and obtaining the data-adaptive bound for DeDa-Exp3.

(ii) In addition, the drift term arising from this reduction is considerably easier to control than previous work, sidestepping the difficulties in the works of Cesa-Bianchi et al. (2019); Thune et al. (2019); Bistritz et al. (2019) as mentioned above and in Footnote 1. Interestingly, a very recent follow-up work of Bistritz et al. (2021) (published on arXiv after the ICML submission deadline) shows that using skipping directly (in a valid way, based on only the observed delays) in the analysis of the delayed Exp3 algorithm allows a simple control of the drift term, and hence leads to the desired adaptive regret bound of order $\sqrt{\log(K)(TK + D)}$ using a doubling trick in tuning the step size.

1.2. Notation

We denote the set $\{1, 2, \ldots, n\}$ of the first n natural numbers by $[n]$. The indicator of an event $\mathcal{E}$ is denoted by $\mathbb{I}[\mathcal{E}]$, taking the value 1 if the event $\mathcal{E}$ happens and 0 otherwise. For a sequence of functions, vectors, or scalars $a_s, a_{s+1}, \ldots, a_t$, we use $a_{s:t}$ to denote the sum $\sum_{n=s}^t a_n$, with $a_{s:t} = 0$ if $s > t$.

2. Problem formulation

The multi-armed bandit problem is a sequential decision problem. Given a finite set of K actions, denoted by $[K] = \{1, \ldots, K\}$, and the time horizon of the problem T, in every time step $t \in [T]$, the learner chooses an action $A_t \in [K]$ and suffers a loss ℓ_{t,A_t}, where $\ell_t \in [0,1]^K$ is a loss vector such that $\ell_{t,i}$ is the loss associated with choosing action i in time step t. We assume that the loss sequence $(\ell_t)_t$ is selected in advance and is not affected by the actions chosen by the learner (a.k.a. the oblivious setting). As usual, we allow the learner to randomize, that is, at time step t the learner determines a distribution p_t in the $K-1$ dimensional probability simplex, and samples action A_t from p_t (conditionally independently of previous random choices, given p_t). With a slight abuse of terminology, sometimes we will refer to both p_t and A_t as the *decision* of the learner at time t.

The learner's performance relative to any fixed action $A^\star$ is measured by the (expected) regret against $A^\star$, defined as

$$R_T(A^\star) = \sum_{t=1}^T \mathbb{E}\left[\ell_{t,A_t}\right] - \sum_{t=1}^T \ell_{t,A^*},$$

and the learner aims to minimize its regret $R_T = \min_{A^* \in [K]} R_T(A^\star)$ against the best action in hindsight.

In the standard multi-armed bandit setting, after taking an action A_t, the learner immediately observes ℓ_{t,A_t}, which can be used to improve its decisions in future time steps; however, the learner does not observe any loss $\ell_{t,i}$ for $i \neq A_t$. In the *delayed-feedback* setting we consider, the situation is somewhat different: after taking an action A_t in time step t, the learner observes the loss ℓ_{t,A_t} only after a delay of d_t time steps, after making a decision in time step $t + d_t$. This means that the decision A_t in time step t can only depend on the feedback which arrives before that time step, that is, on the losses $\{\ell_{s,A_s} : s + d_s < t\}$. Note that delay $d_t = 0$ means that the corresponding feedback becomes available immediately after a decision is made. Without loss of generality[2], we assume that all feedback arrives at the end of time step T, that is, $t + d_t \leq T$ for all $t \in [T]$. We assume that the sequence of delays $(d_t)_t$ is selected before the process starts, obliviously to the actions of the learner. Note, however, that the losses and delays can be selected jointly with an arbitrary dependence among them.

Definitions. The following definitions will be useful in analyzing the regret of delayed algorithms. We use $I_{t,i} = \mathbb{I}[A_t = i]$ to indicate whether action $i \in [K]$ is played at time t. The set of time steps with feedback missing when

[2] This is because the actions $A_1, \ldots, A_T$ that determine the regret R_T only depend on the feedback that arrives before time step T; any remaining feedback can thus be assumed to arrive at the end of time T without affecting R_T.

computing p_t is denoted by $O_t = \{s \in [t-1] : s + d_s \geq t\}$ (with $O_1 = \emptyset$). The number of missing feedbacks at time step t is $\tau_t = |O_t| = \sum_{s=1}^{t-1} \mathbb{I}[s + d_s \geq t]$. The set of time steps where the feedback for time step t is missing is denoted by $D_t = \{s : t \in O_s\} = \{s : t < s \leq t + d_t\}$. Note that the size of this set is $|D_t| = d_t$. We denote the maximum delay by $d^\star = \max_{t \in [T]} d_t$ and the cumulative delay by $D = \sum_{t=1}^T d_t$. Note that $D = \sum_{t=1}^T \tau_t$, since $D = \sum_{t=1}^T \sum_{s=1}^T \mathbb{I}[t \in O_s] = \sum_{s=1}^T \sum_{t=1}^T \mathbb{I}[t \in O_s] = \sum_{s=1}^T \tau_s$.

Loss estimates. Standard bandit algorithms form some estimate $\hat{\ell}_t$ of the loss vector ℓ_t when the feedback ℓ_{t,A_t} is received. A standard estimate is the importance-weighted estimator (Auer et al., 2002), defined as

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{I}[A_t = i]}{p_{t,i}} = \frac{\ell_{t,A_t} I_{t,i}}{p_{t,i}}, \qquad i \in [K] \tag{1}$$

Note that $\hat{\ell}_{t,i}$ is computable when feedback ℓ_{t,A_t} has arrived as it does not depend on any other component of ℓ_t. Let $\mathcal{H}_t = \{(s, A_s, \hat{\ell}_s) : s \in [t-1] \setminus O_t\}$ denote the history of the actual observations when computing p_t. The estimator (1) is unbiased as $\mathbb{E}\left[\hat{\ell}_t | p_t\right] = \ell_t$. We consider learning algorithms whose decision p_t depends on $\mathcal{H}_t$, that is, p_t is $\sigma(\mathcal{H}_t)$-measurable (where $\sigma(\mathcal{H}_t)$ denotes the σ-field generated by $\mathcal{H}_t$).[3] Therefore, for the estimate $\hat{\ell}_t$ in (1) we have $\mathbb{E}\left[\hat{\ell}_t | \mathcal{H}_t\right] = \ell_t$. In addition, we also consider loss estimates with the so-called implicit exploration (see, e.g., Neu, 2015b):

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i} I_{t,i}}{p_{t,i} + \gamma_t}\,, \tag{2}$$

where $\gamma_t \in \sigma(\mathcal{H}_t), t \in [T]$ is a non-negative sequence of reals. The $\sigma(\mathcal{H}_t)$-measurability of γ_t ensures that $\mathbb{E}\left[\hat{\ell}_t | \mathcal{H}_t\right] \leq \ell_t$, thus encouraging exploration by reducing the observed expected loss.

3. The Delay-Adaptive Exp3 Algorithm

Probably the simplest way to extend an algorithm designed for the non-delayed case to the delayed-feedback setting is to apply the same algorithm to the available losses. Such algorithms have been used extensively in the literature (see, e.g., Joulani et al., 2013; Cesa-Bianchi et al., 2019). In this section we analyze a similar extension of the Exp3 algorithm, which we call the delay-adaptive Exp3 (DAda-Exp3) algorithm. While at time t, in the non-delayed case, Exp3 selects action i with probability proportional to $e^{-\eta_t \hat{L}_{t,i}}$ for some step-size $\eta_t > 0$, where $\hat{L}_t = \sum_{s=1}^{t-1} \hat{\ell}_s$,[4] in the delayed case the set of available loss estimates is potentially smaller, and the decision is made based on $\hat{L}_t - \hat{\Delta}_t$, where $\hat{\Delta}_t = \sum_{s \in O_t} \hat{\ell}_s$ is the sum of the missing loss estimates, which have not arrived, but would have arrived in the non-delayed setting.

Algorithm 1: Delay-Adaptive Exp3 (DAda-Exp3).

```
Input: Number of actions $K$.
Initialization:
  $\tilde{L}_{1,i} \leftarrow 0$ for all $i \in [K], \tau_1 \leftarrow 0$.
for $t = 1, 2, \ldots, T$ do
    $\eta_t \leftarrow \sqrt{\frac{\log(K)}{tK + \sum_{s=1}^t \tau_s}}$.
    $p_{t,i} \leftarrow \frac{e^{-\eta_t \tilde{L}_{t,i}}}{\sum_{j=1}^K e^{-\eta_t \tilde{L}_{t,j}}}$ for all $i \in [K]$.
    Play action $A_t \in [K]$ selected randomly according
      to distribution $p_t$.
    Store $p_{t,A_t}$ and $A_t$ in the memory.
    for $s : s + d_s = t$ do
        Observe $\ell_{s,A_s}$ and retrieve $(s, A_s, p_{s,A_s})$ from
          the memory.
        Let $\hat{\ell}_{s,i} = \frac{\ell_{s,A_s}\mathbb{I}[A_s = i]}{p_{s,A_s}}$ for all $i \in [K]$.
    end
      $\tilde{L}_{t+1,i} \leftarrow \tilde{L}_{t,i} + \sum_{s : s + d_s = t} \hat{\ell}_{s,i}$, for all $i \in [K]$.
      $\tau_{t+1} \leftarrow \tau_t + 1 - |\{s : s + d_s = t\}|$.
end
```

Thus, at time t, DAda-Exp3 samples action i with probability

$$p_{t,i} = \frac{e^{-\eta_t(\hat{L}_{t,i} - \hat{\Delta}_{t,i})}}{\sum_{j \in [K]} e^{-\eta_t(\hat{L}_{t,j} - \hat{\Delta}_{t,i})}}\,, \tag{3}$$

where $\eta_t > 0$ is $\sigma(\mathcal{H}_t)$-measurable (i.e., η_t may depend on any feedback information available at the beginning of time step t). DAda-Exp3 adapts to the delays by properly tuning the step-size η_t. In fact, this step-size tuning is the key contribution in the algorithm design. The method, including the step-size tuning is presented in Algorithm 1. It uses the notation $\tilde{L}_{t,i} = \hat{L}_{t,i} - \Delta_{t,i}$ for the sum of the estimated losses for arm $i \in [K]$ available before time step t. Note that although the algorithm only stores p_{t,A_t} for all t (not the whole distribution p_t), the loss estimates $\hat{\ell}_{s,i}$ can be calculated for all $i \in [K]$, because by definition $\hat{\ell}_{s,i} = 0$ for all $i \neq A_s$ (because the indicator function is 0).

The next result gives an upper bound on the expected regret of DAda-Exp3.

[3] Note, however, that this is a restriction, as the computation of p_t could also depends on past decisions with missing feedback, that is, on $\{(p_s, A_s) : s \in O_t\}$.

[4] Note that, perhaps unusually, $\hat{L}_t$ is the sum of losses up to time step $t - 1$, not t.

Theorem 3.1. *Suppose that $\eta_1, \eta_2, \dots, \eta_T$ is a positive, non-increasing sequence of step sizes. Then, for all $A^\star \in [K]$, `DAda-Exp3` satisfies*

$$R_T(A^\star) \le \mathbb{E}\left[\eta_T^{-1}\right]\log(K) + \sum_{t=1}^{T} \min\{1, \mathbb{E}\left[\eta_t(\tau_t + K)\right]\}\,.$$

Proof. Let $p^\star$ be the probability distribution with all mass on $A^\star$. From the definition of regret and the loss-estimates, for any sequence of probability distributions $\tilde{p}_{t+1}, t \in [T]$, we have

$$\begin{aligned}
R_T(A^\star) &= \mathbb{E}\left[\sum_{t=1}^{T} \hat{\ell}_t^\top (p_t - p^\star)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \hat{\ell}_t^\top (\tilde{p}_{t+1} - p^\star)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \hat{\ell}_t^\top (p_t - \tilde{p}_{t+1})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \hat{\ell}_t^\top (\tilde{p}_{t+1} - p^\star)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i}\left(1 - \frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right)\right], \qquad (4)
\end{aligned}$$

where in the first equality we used the fact that $\ell_{t,A_t} = (\ell_{t,A_t}/p_{t,A_t})p_{t,A_t} = \hat{\ell}_t^\top p_t$, and $\ell_{t,A^\star} = \ell_t^\top p^\star = \mathbb{E}\left[\hat{\ell}_t | \mathcal{H}_t\right]^\top p^\star$. Now, let $\tilde{p}_t$ be the (full-information) adaptive exponential-weights updates for the sequence of linear losses $\hat{\ell}_t$, with non-increasing step-sizes η_{t-1}, that is, $\tilde{p}_1$ is the uniform distribution over $[K]$, and for all $t \in [T]$,

$$\tilde{p}_{t+1,i} = \frac{e^{-\eta_t \hat{L}_{t+1,i}}}{\sum_{j=1}^{K} e^{-\eta_t \hat{L}_{t+1,j}}}\,.$$

Note that the index of η is shifted by one, that is, $\tilde{p}_{t+1}$ uses the same step-size η_t that is used by p_t (rather than η_{t+1}, which is used by p_{t+1}). This is not problematic: we only assume the imaginary iterate $\tilde{p}_t$ uses slightly outdated information for tuning the step-size, and is still $\sigma(\mathcal{H}_t)$-measurable.

Thus, (4) decomposes the regret into two terms: the first one is the "cheating" (or look-ahead) regret for the ideal (imaginary) exponential-weights iterate $\tilde{p}_{t+1}$ (which depends on $\hat{\ell}_t$ at time step t), while the second term is a "drift" term which measures the effect of using p_t instead of $\tilde{p}_{t+1}$.

The first, cheating regret term can be directly bounded by Theorem 3 of Joulani et al. (2020)[5] as

$$\sum_{t=1}^{T} \hat{\ell}_t^\top (\tilde{p}_{t+1} - p^\star) \le \eta_T^{-1}\log(K)\,. \qquad (5)$$

To control the "drift term", that is, the second term on the right hand side of (4), we bound $\mathbb{E}\left[\frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right]$ from below. Observe that since the losses are non-negative, for all $t \in [T]$ and $i \in [K]$, $\hat{L}_{t,i}$, $\hat{\Delta}_{t,i}$ and η_t are positive. Hence, we have

$$e^{-\eta_t(\hat{L}_{t,i} - \hat{\Delta}_{t,i})} = e^{-\eta_t(\hat{L}_{t+1,i} - \hat{\ell}_{t,i} - \hat{\Delta}_{t,i})} \ge e^{-\eta_t \hat{L}_{t+1,i}}$$

for all $t \in [T]$ and $i \in [K]$, which implies

$$\begin{aligned}
\frac{\tilde{p}_{t+1,i}}{p_{t,i}} &= \frac{e^{-\eta_t \hat{L}_{t+1,i}}}{e^{-\eta_t(\hat{L}_{t,i} - \hat{\Delta}_{t,i})}} \cdot \frac{\sum_{j\in[K]} e^{-\eta_t(\hat{L}_{t,j} - \hat{\Delta}_{t,j})}}{\sum_{j\in[K]} e^{-\eta_t \hat{L}_{t+1,j}}} \\
&\ge e^{-\eta_t \hat{\Delta}_{t,i} - \eta_t \hat{\ell}_{t,i}} \ge 1 - \eta_t \hat{\Delta}_{t,i} - \eta_t \hat{\ell}_{t,i}\,, \qquad (6)
\end{aligned}$$

using in the last step the fact that $e^x \ge 1 + x$ for all $x \in \mathbb{R}$. Thus, we can use (6) to upper-bound the second expectation on the right-hand-side of (4) as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i}\left(1 - \frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right)\right] \\
&\le \mathbb{E}\left[\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i} \eta_t \hat{\Delta}_{t,i} + \eta_t \hat{\ell}_{t,i}^2 p_{t,i}\right] \qquad (7) \\
&= \mathbb{E}\left[\sum_{i=1}^{K} \ell_{t,i} I_{t,i} \eta_t \sum_{s\in O_t} \frac{\ell_{s,i} I_{s,i}}{p_{s,i}}\right] + \mathbb{E}\left[\sum_{i=1}^{K} \eta_t \ell_{t,i}^2 I_{t,i}/p_{t,i}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{s\in O_t} \ell_{t,i}\ell_{s,i} \mathbb{E}\left[\eta_t \frac{I_{t,i} I_{s,i}}{p_{s,i}} \middle| \mathcal{H}_t\right]\right] \\
&\quad + \mathbb{E}\left[\sum_{i=1}^{K} \ell_{t,i}^2 \mathbb{E}\left[\frac{\eta_t I_{t,i}}{p_{t,i}} \middle| \mathcal{H}_t\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{s\in O_t} \ell_{t,i} \eta_t \ell_{s,i} p_{t,i}\right] + \mathbb{E}\left[\sum_{i=1}^{K} \eta_t \ell_{t,i}^2\right] \\
&\le \mathbb{E}\left[\sum_{i=1}^{K} p_{t,i} \eta_t \tau_t\right] + \mathbb{E}\left[\eta_t K\right] = \mathbb{E}\left[\eta_t(\tau_t + K)\right]\,, \qquad (8)
\end{aligned}$$

where in the second line we have used the definitions of $\hat{\Delta}_{t,i}$ and $\hat{\ell}_{t,i}$ for $t = 1, 2, \dots, T$, and in the third line we have used the tower rule. The fourth step follows since $p_{s,i}$ and $p_{t,i}$ are determined by the feedback that is received by time t, and since the feedback for $I_{s,i}$ is missing, $I_{s,i}$ and $I_{t,i}$ are conditionally independent given $\mathcal{H}_t$ with distributions $p_{s,i}$ and $p_{t,i}$, respectively. The last inequality uses the assumption that the losses are upper-bounded by 1. Combining (8) with

$$\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i}\left(1 - \frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right) \le \sum_{i=1}^{K} \ell_{t,i} \mathbb{I}\left[A_t = i\right] = \ell_{t,A_t} \le 1,$$

summing up for all t and putting back the resulting bound into (4), together with (5), gives the desired bound on the regret. □

[5] We have invoked Theorem 3 of Joulani et al. (2020) with $p_t \equiv 0, t \in [T]$, $r_0 = (1/\eta_0)\sum_i p_i \log(p_i)$ and $r_t(p) = (1/\eta_t - 1/\eta_{t-1})\sum_i p_i \log(p_i)$, $t \in [T]$, and dropped the Bregman-divergence terms due to the convexity of r_t.

Using the non-increasing step-size sequence $\eta_t = \sqrt{\frac{\log(K)}{tK+\sum_{s=1}^t \tau_s}}$, we obtain the first fully delay-adaptive bound for the Exp3-family of algorithms:

Corollary 3.2. *With $\eta_t = \sqrt{\frac{\log(K)}{tK+\sum_{s=1}^t \tau_s}}$, the regret of the* DAda-Exp3 *algorithm can be bounded as*

$$R_T \leq 3\sqrt{\log(K)\,(TK+D)}.$$

Proof. The result is a direct corollary of Theorem 3.1: We can bound the second term in the regret bound in the theorem by the standard inequality that for any $a_t > 0$, $\sum_{t=1}^T a_t / \sqrt{\sum_{s=1}^t a_s} \leq 2\sqrt{\sum_{t=1}^T a_t}$ (see, e.g., Lemma 4 of McMahan, 2017). Applying this inequality for $a_t = K + \tau_t$, we obtain

$$\sum_{t=1}^T \left(\mathbb{E}\left[\ell_{t,A_t}\right] - \ell_{t,A^\star}\right) \leq 3\sqrt{\log(K)\left(TK + \sum_{t=1}^T \tau_t\right)}.$$

The statement of the corollary then follows by the fact that $D = \sum_{t=1}^T \tau_t$. □

Remark 3.3. The decomposition in (4) in the proof of Theorem 3.1 involves the cheating regret, which is a well-known technique in online learning (see, e.g., Joulani et al., 2020). In fact, the non-delayed case can be thought of as the cheating case with a delay of 1, where $\hat{\ell}_t$ is not available at the time of computing $\tilde{p}_t$, but is available at time $t+1$. In this sense, our decomposition follows naturally by collecting all such delayed losses only in the drift term; the cheating regret term can then be bounded in a black-box manner. This also has a further benefit in the bandit setting: had we used the standard regret (with $\tilde{p}_t$ in place of $\tilde{p}_{t+1}$), the products $\hat{\ell}_t^\top \tilde{p}_t$ that would show up in the standard regret decomposition would include a ratio $\tilde{p}_{t,i}/p_{t,i}$ (due to the importance weight $p_{t,i}$ used in $\hat{\ell}_{t,i}$), and bounding this ratio from above has been the source of much difficulty in previous work, preventing the application of the simple analysis techniques we use (Cesa-Bianchi et al., 2019; Thune et al., 2019; Bistritz et al., 2019).

4. High-probability bounds

In this section, we show that using the loss estimate with implicit exploration (cf. equation 2) enables us to prove a regret bound that hold with high probability instead of holding only in expectation. We present a bound for DAda-Exp3. The derivation of a bound for DeDa-Exp3 is more involved, and we leave it for an extended version of the paper.

Theorem 4.1. *Suppose* DAda-Exp3 *is run with a non-increasing step-size sequence (η_t) and using the loss estimate (2) with $\gamma_t = \eta_t$. Let $\delta \in (0,1)$ and $A^\star \in [K]$ be an arbitrary action. Then, with probability at least $1-\delta$, the regret of the algorithm against $A^\star$ can be bounded as*

$$\sum_{t=1}^T (\ell_{t,A_t} - \ell_{t,A^\star}) \leq \frac{3\log(K)}{2\eta_T} + \sum_{t=1}^T \eta_t\,(\tau_t + 2K) + \left(\frac{\eta_T^{-1} + d^\star + 2}{2}\right)\log\left(\frac{2}{\delta}\right). \quad (9)$$

This implies the following corollary, which is proved exactly as Corollary 3.2.

Corollary 4.2. *With $\eta_t = \frac{1}{2}\sqrt{\frac{3\log(K)}{2tK+\sum_{s=1}^t \tau_s}}$, under the conditions of Theorem 4.1,*

$$\sum_{t=1}^T (\ell_{t,A_t} - \ell_{t,A^\star}) \leq 2\sqrt{3\log(K)\,(2KT+D)} + \left(\sqrt{\frac{2TK+D}{3\log(K)}} + \frac{d^\star}{2} + 1\right)\log\left(\frac{2}{\delta}\right).$$

Note that the dependence on δ can be improved if the step size η_t also depends on δ. In the corollary we chose to tune the algorithm to be oblivious to the error parameter δ.

In the proof of the theorem (given in Appendix A),we apply similar ideas as in the proof of Theorem 3.1, but instead of taking expectations to control the loss estimate terms, we use a lemma of Neu (2015b) to replace the loss estimates by the true losses at the expense of an additive logarithmic penalty, with high probability.

5. Skipping time steps

Looking at the form of the bound in Theorem 3.1, one can observe that the terms in the second summation are a minimum of 1 and $\eta_t(\tau_t + K)$, where the latter comes from bounding the drift terms. As such, whenever 1 is smaller than $\eta_t(\tau_t + K)$, our analysis in (8) is too pessimistic. The effect of this could be avoided by keeping the minimum term when we define the step size η_t, but this is not straightforward if we want to keep the simple sum structure of $1/\eta_t^2$, which is used in the proof of Corollary 3.2. A simpler approach is to ensure that τ_t never becomes too large (compared to η_t), which can be done by limiting individual delays d_t by pretending that their corresponding loss arrives (and has value 0) when d_t is too large, and bound the regret in the corresponding time step separately by 1. This essentially means that the algorithm eventually skips time steps with excessive delays. This also addresses another problem: namely that the cumulative delay D can be dominated by a few large delay values, which–intuitively–should not cause such a large penalty in the regret.

The idea of skipping, originally coined by Thune et al. (2019), was perfected by Zimmert and Seldin (2019), who provided a proper way to skip some time steps and tune the step size accordingly. In what follows, we adopt their way of tuning the step size, and closely follow their analysis (which they do in the context of their more complicated follow-the-regularized-leader algorithm).

Next we describe how the tuning method of Zimmert and Seldin (2019) can be adapted to our `DAda-Exp3` algorithm: The advanced tuning method of skipping procedure works as follows: For every time step t and time steps $s \le t$, we keep a binary indicator $a_s^t \in \{0,1\}$ such that for any round t we include the loss from time s to the set of missing losses if $a_s^t = 1$, and not if $a_s^t = 0$. That is, the number of *counted* missing losses is

$$\tilde{\tau}_t = \sum_{s=1}^{t} a_s^t \mathbb{I}\left[s + d_s \ge t\right] .$$

Let $\tilde{D}_t = \sum_{s=1}^t \tilde{\tau}_t$ denote the cumulative number of counted missing feedbacks. a_s^t is originally set to 1 for all $t \ge s$, but if $s \in O_t$ and $\min\{d_s, t-s\} > \sqrt{\tilde{D}_t / \log(K)}$, we set $a_s^{t'} = 0$ for all $t' > t$ (by Lemma 7 of Zimmert and Seldin, 2019, this happens for at most one s value in any time step t).

Tuning the step size of `DAda-Exp3` with $\tilde{\tau}$ instead of τ, the regret of the `DeDa-Exp3` algorithm can be bounded as follows (the proof, given in Appendix B, combines our earlier bounds with some simple results from Zimmert and Seldin, 2019):

Theorem 5.1. *(i) Bound in expectation: The expected regret of the `DAda-Exp3` algorithm with loss estimates (1) and step sizes $\eta_t = \sqrt{\frac{\log(K)}{tK + \sum_{s=1}^t \tilde{\tau}_t}}$ can be bounded as*

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\ell_{t,A_t}\right] - \ell_{t,A^\star}\right) \le 3\sqrt{TK\log(K)} + 10 \max\left\{2\log K, \min_{R \subset [T]} \left(|R| + \sqrt{D_{\bar{R}} \log(K)}\right)\right\},$$

where for any $R \subset [T]$, $\bar{R} = [T] \setminus R$, and $D_{\bar{R}} = \sum_{t \in \bar{R}} d_t$.

(ii) High-probability bound: Let $\delta \in (0,1)$. The regret of the `DAda-Exp3` algorithm with loss estimates (2) and step sizes $\eta_t = \gamma_t = \frac{1}{2}\sqrt{\frac{3\log(K)}{2tK + \sum_{s=1}^t \tau_s}}$ can be bounded, with probability at least $1-\delta$, as

$$\sum_{t=1}^{T} \left(\ell_{t,A_t} - \ell_{t,A^\star}\right) \le C_{1,\delta}\sqrt{KT\log(K)} + C_{2,\delta} \max\left\{2\log K, \min_{R \subset [T]} \left(|R| + \sqrt{D_{\bar{R}} \log(K)}\right)\right\}$$

where $C_{1,\delta} = 2\sqrt{6} + \sqrt{\frac{2}{3}\frac{\log(2/\delta)}{\log(K)}}$ and $C_{2,\delta} = 4(\sqrt{3}+1) + \left(1 + \frac{2}{\sqrt{3}}\right)\frac{\log(2/\delta)}{\log(K)}$.

The above theorem shows that the regret of the algorithm is essentially of the same order as if a set of time steps R was to be skipped, and the algorithm was only run on its complement $\bar{R}$. Note that while our original high-probability regret bound (cf. Corollary 4.2) depended on the maximum delay $d^\star$, this dependence is eliminated from the high-probability bound of the theorem, as the maximum delay is effectively bounded by $\sqrt{\tilde{D}_T / \log(K)}$.

6. Adapting to delay and data at the same time

In this section, we consider a different step-size sequence that yields AdaGrad-style bounds. Recall that from (7) in the proof of Theorem 3.1, we have

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i}\left(1 - \frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right) \le \sum_{t=1}^{T} \eta_t \sum_{i=1}^{K} \ell_{t,i}^{\text{fwd}},$$

where $\ell_{t,i}^{\text{fwd}} = \hat{\ell}_{t,i} p_{t,i} \hat{\Delta}_{t,i} + \hat{\ell}_{t,i}^2 p_{t,i}$. Therefore, ideally, we want to set the step-size η_t as

$$\eta_t = \sqrt{\frac{\log(K)}{\sum_{i=1}^K \ell_{1:t,i}^{\text{fwd}}}}, \tag{10}$$

to optimize the regret bound and obtain a data-adaptive bound of the form

$$R_T \le 3\mathbb{E}\left[\sqrt{\log(K) \sum_{i=1}^K \ell_{1:T,i}^{\text{fwd}}}\right] . \tag{11}$$

However, we do not have access to the missing observations $\hat{\ell}_s, s \in O_t \cup \{t\}$, when calculating η_t. Therefore, we approximate the step-size of (10) with another sequence that can be computed at each time step. Algorithm 2 provides the details of this approximation.

The algorithm keeps track of the largest delay ($d_t^\star$) and, similarly to Thune et al. (2019), to do so it needs to observe the delay for action A_t in advance. This can be avoided if the algorithm has access to an a priori upper bound d^B on the maximum delay: setting $d_0^\star$ to this upper bound results in $d_t^\star = d^B$ for all $t \in [T]$. As usual in AdaGrad-style algorithms, the algorithm maintains additional vectors m_t, z_t and L_t^{bck} to compute the step size. In addition, the algorithm uses a memory to store values of m_s, z_s, and p_s for past steps s with missing feedback. In particular, after coming up with the action distribution p_s for time step s, we store the current values of m_s and z_s, as well as the action distribution p_s and the action taken, A_s. When the feedback for time step s arrives at the end of time step $t = s + d_s$, we retrieve these values, and use them to compute $\tilde{L}_{t+1}$, m_{t+1}, and L_{t+1}^{bck}.

Memory requirement. Clearly, Algorithm 2 requires $\Theta\left(d_T^\star\right)$ memory, which can be explicitly implemented using

Algorithm 2: Delay- and Data-Adaptive `Exp3` (`DeDa-Exp3`).

```
Input: Number of actions $K$.
Initialization:
  $\tilde{L}_{1,i} \leftarrow 0, m_{1,i} \leftarrow 0$ for all $i \in [K]$.
  $d_0^\star \leftarrow 0, L_1^{\text{bck}} \leftarrow 0$.
for $t = 1, 2, \dots, T$ do
    $d_t^\star \leftarrow \max\{d_t, d_{t-1}^\star\}$ .
    $\eta_t = \gamma_t \leftarrow \left[\frac{4(d_t^\star)^2+6d_t^\star+2}{\log(K)} + \sqrt{\frac{L_t^{\text{bck}}}{\log(K)}}\right]^{-1}$ .
    $p_{t,i} \leftarrow \frac{e^{-\eta_t \tilde{L}_{t,i}}}{\sum_{j=1}^K e^{-\eta_t \tilde{L}_{t,j}}}$ .
    Play action $A_t \in [K]$ selected randomly according
      to distribution $p_t$.
    Store $\gamma_t, m_t, \tilde{L}_t, p_t$, and $A_t$ in memory.
    for $s : s + d_s = t$ do
        Observe $\ell_{s,A_s}$ and retrieve $A_s, p_s, m_s, z_s$ and
          $\gamma_s$ from memory.
        Let $\hat{\ell}_{s,i} = \frac{\ell_{s,A_s} I_{s,i}}{p_{s,i}+\gamma_s}$ for all $i \in [K]$.
    end
    $\tilde{L}_{t+1,i} \leftarrow \tilde{L}_{t,i} + \sum_{s:s+d_s=t} \hat{\ell}_{s,i}$, for all $i \in [K]$.
    $m_{t+1,i} \leftarrow m_{t,i} + \sum_{s:s+d_s=t} \hat{\ell}_{s,i} p_{s,i}$, for all $i \in [K]$.
    $L_{t+1}^{\text{bck}} \leftarrow L_t^{\text{bck}} + \sum_{i=1}^K \sum_{s:s+d_s=t} \Big(\hat{\ell}_{s,i}(m_{t+1,i} - m_{s,i})$
        $+ \hat{\ell}_{s,i} p_{s,i} \left(\tilde{L}_{t+1,i} - \tilde{L}_{s,i}\right)\Big)$.
end
```

a hash table with an amortized computation cost of $\Theta(1)$ per storage and retrieval.

Step size. It is easy to verify that for all $t \in [T+1]$ and $i \in [K]$, $\tilde{L}_{t,i} = \sum_{j:j+d_j<t} \hat{\ell}_{j,i}$ and $m_{t,i} = \sum_{j:j+d_j<t} \hat{\ell}_{j,i} p_{j,i}$. In addition, if we define

$$\ell_{s,i}^{\text{bck}} = \sum_{j:s \le j+d_j \le s+d_s} \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i})\,,$$

then it is easy to see that $\ell_{s,i}^{\text{bck}} = \hat{\ell}_{s,i}\left(m_{s+d_s+1,i} - m_{s,i}\right) + \hat{\ell}_{s,i} p_{s,i}\left(\tilde{L}_{s+d_s+1,i} - \tilde{L}_{s,i}\right)$. Therefore, $L_t^{\text{bck}} = \sum_{i=1}^K \sum_{s:s+d_s<t} \ell_{s,i}^{\text{bck}}$, and the algorithm uses the step-size schedule

$$\eta_t^{-1} = \frac{4(d_t^\star)^2 + 6d_t^\star + 2}{\log(K)} + \sqrt{\frac{\sum_{i=1}^K \sum_{s:s+d_s<t} \ell_{s,i}^{\text{bck}}}{\log(K)}}\,. \tag{12}$$

We start the analysis by showing that the backward loss estimates $\ell_{t,i}^{\text{bck}}$ are not far away from the forward losses $\ell_{t,i}^{\text{fwd}}$. Hence, the step-size sequence given by (12) will result in a regret not far away from what could be achieved by the ideal step-size (10). This is captured by the following lemma, proved in Appendix C.1:

Lemma 6.1 (Step-size control)**.** *For all $t \in [T]$,*

$$\sum_{s:s+d_s<t} \ell_{s,i}^{\text{bck}} \le 2 \sum_{j=1}^{t} \sum_{s \in O_j \cup \{j\} \cup D_j} \hat{\ell}_{s,i}\hat{\ell}_{j,i} p_{j,i}\,, \tag{13}$$

where for every $t \in [T]$, $D_t = \{s : t \in O_s\}$ is the set of time steps at which the feedback ℓ_{t,A_t} is delayed. In addition, if $\hat{\ell}_t, t \in [T]$, is given by Algorithm 2 using a non-increasing sequence of γ_t values, then for all $t \in [T]$,

$$\sum_{i=1}^K \ell_{1:t,i}^{\text{fwd}} \le \sum_{i=1}^K \sum_{s:s+d_s<t} \ell_{s,i}^{\text{bck}} + \frac{4d_t^{\star 2} + 6d_t^\star + 2}{\gamma_t}\,. \tag{14}$$

Based on this lemma, we can show that using the step size in (12) results in at most a lower-order penalty in the regret compared to regret bound (11) for the ideal step size. This is captured by the next theorem, proved in Appendix C.

Theorem 6.2 (Adapting to delay and data)**.** *The expected regret of `DeDa-Exp3` can be bounded as*

$$\sum_{t=1}^T \left(\mathbb{E}\left[\ell_{t,A_t}\right] - \ell_{t,A^\star}\right)$$

$$\le C_T + c\sqrt{\log(K) \sum_{t=1}^T \Bigg(\sum_{s \in O_t \cup D_t} \mathbb{E}\left[\ell_{s,A_s}\ell_{t,A_t}\right] + \sum_{i=1}^K \ell_{t,i}\Bigg)},$$

where $C_T = 4(d_T^\star)^2 + 6d_T^\star + 2$, $c = 2 + \sqrt{2}$, and for all $t \in [T]$, D_t is defined in Lemma 6.1.

Remark 6.3. A bound that depends on L_{T,A^*} instead of $\sum_i L_{T,i}$ would be preferable, but to our knowledge such bounds are not available for `Exp3` even in the non-delayed case and require other techniques (such as using the log-barrier regularizer, Neu, 2015a). Nevertheless, the above bound still preserves the separation of the cumulative delay and K. Using that $\sum_{t=1}^T |D_t| = \sum_{t=1}^T |O_t| = D$ and that the losses are in $[0, 1]$, we obtain

$$R_T \le C_T + c\sqrt{\log(K)\,(KT + 2D)}.$$

On the other hand, $\sum_i L_{T,i}$ can be much smaller than KT, for example, when most arms have a small loss or when the actual loss range is $[0, B]$ for some unknown $B \ll 1$ (i.e., the algorithm adapts to the unknown B, and the final bound depends only on B and d_t, not on the algorithms' choices). Finally, defining $c' = c\sqrt{\log(K)}$ and $C'_T = C_T + c'^2 d_T^\star$, easy calculations (presented in Appendix D) give

$$R_T \le 2C'_T + 2c'\sqrt{2d_T^\star L_{T,A^*}} + 2c'\sqrt{\textstyle\sum_{i=1}^K L_{T,i}}\,, \tag{15}$$

showing that *the effect of the delay on the regret scales only with the loss of the optimal arm*. Finally, it is worth mentioning that the constant term C_T in the theorem and C'_T above are quadratic in the maximum delay $d^\star_T$, and hence the bound is only meaningful when $d^\star_T$ is sufficiently small (i.e., $o(\sqrt{T})$), which includes, for example, the important case of large but bounded delays. For delays of order $\Omega(\sqrt{T})$, even the skipping technique of Section 5 (or the simpler skipping method of Thune et al., 2019 which utilizes the a priori knowledge of d_t) may not completely resolve the issue, as skipping ensures the maximum delay faced by the algorithm grows with the same rate as the regret, which is not enough to keep the term C_T sublinear. Extending `DeDa-Exp3` to arbitrarily large delays in a meaningful way remains an open problem.

7. Conclusions

In this paper we presented delay- and data-adaptive algorithms for the multi-armed bandit problem with delayed feedback. First, through a remarkably simple proof technique, we showed that the expected regret of our simpler algorithm, `DAda-Exp3`, scales optimally with the sum of the delays, up to logarithmic factors (without any advance knowledge of the delays). We also showed that using the implicit-exploration loss estimate of Neu (2015b), `DAda-Exp3` achieves the same near-optimal regret guarantees with high probability, providing the first high-probability regret bound in the literature for a fully delay-adaptive bandit algorithm.

One problem with the regret bounds that scale with the sum of the delays is that they become too large when individual delays are large, for example, a single delay of T has a significant impact on the regret bound. Recently, Thune et al. (2019) addressed this question by "skipping" rounds with large delays, significantly reducing the regret. However, to achieve this, they needed to know the delays at action-time. Zimmert and Seldin (2019) provided a delay-adaptive solution for this problem. Based on the latter result, we proved similar "skipping" regret bounds for modified versions of `DAda-Exp3`, both in expectation and with high probability.

Finally, we presented the `DeDa-Exp3` algorithm, the first method for delayed bandits that, besides the delays, also adapts to the losses, achieving a potentially large improvement on easy problems. While for `DAda-Exp3`, a bound on the expected regret was possible with the standard importance-weighting loss estimator, and the estimator based on implicit exploration was only needed for the high probability bound, employing the latter in `DeDa-Exp3` is crucial for being able to to control η_t properly, and hence for obtaining a meaningful regret bound even in expectation. Deriving high-probability regret bounds and extending the skipping technique to `DeDa-Exp3` is left for future work. Solving these problems require some innovations: For the first one, new results concerning the concentration of products of certain loss estimates are needed. The issue with the second problem is that the natural data-dependent variant of the skipping decision (rather than the version used together with `DAda-Exp3`, which only depends on the delays, but not on the observed losses) induces a complicated dependence on past actions, significantly complicating the simple deterministic skipping mechanism which we used and analyzed for `DAda-Exp3`.

Acknowledgements

The authors would like to thank Tor Lattimore for his insightful comments.

References

Alekh Agarwal and John Duchi. Distributed delayed stochastic optimization. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 873–881, 2011.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *In Proceedings of the 22nd Conference on Learning Theory*, 2009.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2002.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems 32*, pages 11349–11358. 2019.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. No discounted-regret learning in adversarial bandits with delays. *arXiv preprint:2103.04550*, March 2021.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20(17):1–38, 2019.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1097–1105, 2014.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang.

Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 169–178, Corvallis, Oregon, 2011.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. (extended arXiv version : http://arxiv.org/abs/1306.0686).

Pooria Joulani, András György, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1744–1750, 2016.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808: 108–138, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670, 2010.

Timothy A. Mann, Sven Gowal, András György, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4324–4332, 2019.

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

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 1360–1375, 2015a.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 3168–3176, 2015b.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23*, pages 1804–1812, 2010.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, March 2014.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems 28*, pages 1270–1278. 2015.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems 32*, pages 6541–6550. 2019.

Claire Vernade, András György, and Timothy Mann. Nonstationary delayed bandits with intermediate observations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9722–9732, 2020.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. *arXiv preprint:1910.06054*, 2019.

A. Proof of Theorem 4.1

To prove the theorem, we use the following lemma of Neu (2015b).[6]

Lemma A.1 (Lemma 1 of Neu, 2015b)**.** *For $t \in [T], i \in [K]$, let $\gamma_t, \alpha_{t,i}$ be non-negative $\mathcal{H}_t$-measurable random variables satisfying $\alpha_{t,i} \le 2\gamma_t$, and let $\hat{\ell}_{t,i}$ be given by (2). Then, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \alpha_{t,i}\left(\hat{\ell}_{t,i} - \ell_{t,i}\right) \le \log(1/\delta)\,.$$

Proof of Theorem 4.1. We apply similar ideas as in the proof of Theorem 3.1, but instead of taking expectations to control the loss estimate terms, we use Lemma A.1 to replace the loss estimates by the true losses at the expense of an additive logarithmic penalty, with high probability.

To start, we define $\tilde{p}_t$ as in the proof of Theorem 3.1, which implies that we have (5) and (6). Therefore, on the one hand,

$$\begin{aligned}
\sum_{t=1}^{T} \hat{\ell}_t^\top (p_t - \tilde{p}_{t+1}) &= \sum_{t=1}^{T}\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i}\left(1 - \frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right) \\
&\le \sum_{t=1}^{T}\sum_{i=1}^{K} \hat{\ell}_{t,i} p_{t,i} \eta_t \hat{\Delta}_{t,i} + \sum_{t=1}^{T} \eta_t \hat{\ell}_{t,i}^2 p_{t,i} \\
&\le \sum_{t=1}^{T}\sum_{i=1}^{K} \ell_{t,i} I_{t,i} \eta_t \sum_{s \in O_t} \hat{\ell}_{s,i} + \sum_{t=1}^{T}\sum_{i=1}^{K} \eta_t \hat{\ell}_{t,i} \\
&= \sum_{s=1}^{T}\sum_{i=1}^{K} \hat{\ell}_{s,i}\left(\sum_{t:s\in O_t} \ell_{t,i} I_{t,i} \eta_t\right) + \sum_{t=1}^{T}\sum_{i=1}^{K} \eta_t \hat{\ell}_{t,i} \\
&\le \sum_{s=1}^{T}\sum_{i=1}^{K} \hat{\ell}_{s,i}\left(\sum_{t:s\in O_t} \eta_t I_{t,i}\right) + \sum_{t=1}^{T}\sum_{i=1}^{K} \eta_t \hat{\ell}_{t,i}\,, \qquad (16)
\end{aligned}$$

where the second line follows from (6), the third line follows by the fact that $\hat{\ell}_{t,i} p_{t,i} \le \ell_{t,i} I_{t,i} \le 1$, and the last line follows since $\ell_{t,i} \in [0,1]$.

On the other hand, we can derive a deterministic counterpart of (4) as follows:

$$\sum_{t=1}^{T} (\ell_{t,A_t} - \ell_{t,A^\star}) = \sum_{t=1}^{T} \hat{\ell}_t^\top (\tilde{p}_{t+1} - p^\star) + \sum_{t=1}^{T} \hat{\ell}_t^\top (p_t - \tilde{p}_{t+1}) + \epsilon^\star_{1:T} + \sum_{t=1}^{T}\left(\ell_{t,A_t} - \hat{\ell}_t^\top p_t\right)\,, \qquad (17)$$

where $\epsilon^\star_t = \hat{\ell}_{t,A^\star} - \ell_{t,A^\star}$. Following the proof of Theorem 1 of Neu (2015b), it is easy to show that

$$\sum_{t=1}^{T}\left(\ell_{t,A_t} - \hat{\ell}_t^\top p_t\right) = \sum_{t=1}^{T} \gamma_t \sum_{i=1}^{K} \hat{\ell}_{t,i}\,.$$

Then, combining with (16), we have

$$\begin{aligned}
\sum_{t=1}^{T} \hat{\ell}_t^\top (p_t - \tilde{p}_{t+1}) + \sum_{t=1}^{T}\left(\ell_{t,A_t} - \hat{\ell}_t^\top p_t\right) &\le \sum_{t=1}^{T}\sum_{i=1}^{K}\left(\eta_t + \gamma_t + \sum_{s:t\in O_s} \eta_s I_{s,i}\right)\hat{\ell}_{t,i} \\
&= \sum_{t=1}^{T}\sum_{i=1}^{K}\left(\eta_t + \gamma_t + \sum_{s:t\in O_s} \eta_s I_{s,i}\right)\ell_{t,i} + \frac{d^\star + 2}{2}\hat{\epsilon}_{1:T}\,,
\end{aligned}$$

[6] Neu (2015b) states the lemma for "a fixed sequence" of γ_t, but this is not used anywhere in their proof; their proof goes through without change, as long as γ_t is determined by the history $\mathcal{H}_t$.

where $\hat{\epsilon}_t = \sum_{i=1}^K 2\frac{\eta_t+\gamma_t+\sum_{s:t\in O_s}\eta_s I_{s,i}}{d^\star+2}(\hat{\ell}_{t,i}-\ell_{t,i})$. Note that in the latter definition, the coefficient of $(\hat{\ell}_{t,i}-\ell_{t,i})$ is bounded by $2\gamma_t$ since $\eta_t=\gamma_t$, $\eta_t \geq \eta_s$ for all s such that $t \in O_s$ (since $s > t$ in this case), and $d^\star \geq |O_s|$. Hence, with probability at least $1-\delta'$, $\hat{\epsilon}_{1:T} \leq \log(1/\delta')$ by Lemma A.1 for any $\delta' \in (0,1)$. Furthermore,

$$\sum_{t=1}^T\sum_{i=1}^K\sum_{s:t\in O_s}\eta_s I_{s,i} = \sum_{s=1}^T\sum_{t:t\in O_s}\eta_s = \sum_{s=1}^T \eta_s\tau_s\ .$$

Therefore, using that $\ell_{t,i}\in[0,1]$ and $\eta_t=\gamma_t$, the first term on the right hand side above can be bounded as

$$\sum_{t=1}^T\sum_{i=1}^K\left(\eta_t+\gamma_t+\sum_{s:t\in O_s}\eta_s I_{s,i}\right)\ell_{t,i} \leq \sum_{t=1}^T 2K\eta_t + \sum_{s=1}^T\eta_s\tau_s = \sum_{t=1}^T\eta_t(\tau_t+2K)\,.$$

Putting these back into (17), combining with (5), and letting $\gamma_t=\eta_t$, for any $\delta'\in[0,1]$ we have

$$\begin{aligned}\sum_{t=1}^T(\ell_{t,A_t}-\ell_{t,A^\star}) &\leq \frac{\log(K)}{\eta_T}+\sum_{t=1}^T\eta_t(\tau_t+2K)+\frac{d^\star+2}{2}\hat{\epsilon}_{1:T}+\epsilon^\star_{1:T}\\ &\leq \frac{\log(K)}{\eta_T}+\sum_{t=1}^T\eta_t(\tau_t+2K)+\frac{\log(K/\delta')}{2\eta_T}+\frac{d^\star+2}{2}\log(1/\delta')\\ &= \frac{3\log(K)}{2\eta_T}+\sum_{t=1}^T\eta_t(\tau_t+2K)+\frac{\eta_T^{-1}+d^\star+2}{2}\log(1/\delta')\,,\end{aligned}$$

with probability at least $1-2\delta'$, where we also used that

$$\epsilon^\star_{1:T} \leq \frac{1}{2\gamma_T}\sum_{t=1}^T 2\gamma_t\left(\hat{\ell}_{t,A^*}-\ell_{t,A^\star}\right) = \frac{1}{2\gamma_T}\sum_{t=1}^T\sum_{i=1}^K 2\gamma_t\mathbb{I}\left[A^*=i\right]\left(\hat{\ell}_{t,i}-\ell_{t,i}\right) \leq \frac{1}{2\gamma_T}\log(K/\delta')$$

with probability at least $1-\delta'$ simultaneously for all $A^\star$ by Lemma A.1 and the union bound.

Letting $\delta'=\delta/2$ and using the assumption that $\ell_{t,i}\leq 1$ completes the proof. □

B. Proof of Theorem 5.1

Proof. First define the *effective* delay for time step s as $\tilde{d}_s=\sum_{t=s+1}^{s+d_s}a_s^t$; that is, if time step s is not "skipped" (i.e., $a_s^t=1$ for all $t\in[T]$), $\tilde{d}_s=d_s$, and $\tilde{d}_s=t-s$ if s is skipped at the end of time step t, that is, $a_s^{t+1}=0$ and $a_s^t=1$. Let $S=\{t\in[T]: a_t^T=0\}$ denote the set of skipped time steps, and define a new loss sequence $(\tilde{\ell}_t)$ such that the loss is zeroed out if the corresponding time step is ever skipped by the algorithm, that is, $\tilde{\ell}_t=\ell_t$ if $t\notin S$ and $\tilde{\ell}_t=0$ if $t\in S$. Note that this loss sequence can be constructed deterministically from the loss sequence (ℓ_t) and the delay sequence (d_t).

It is easy to see that the τ_t-dependent tuning of DAda-Exp3, considered in this theorem, results in exactly the same sequence of predictions (p_t) as the original DAda-Exp3 algorithm for the losses $(\tilde{\ell}_t)$ and delays $(\tilde{d}_t)$.

To prove the upper bound on the expected regret, we start by applying Corollary 3.2 for the latter case:

$$\sum_{t=1}^T\left(\mathbb{E}\left[\ell_{t,A_t}\right]-\ell_{t,A^\star}\right) \leq |S|+\sum_{t=1}^T\left(\mathbb{E}\left[\tilde{\ell}_{t,A_t}\right]-\tilde{\ell}_{t,A^\star}\right) \leq |S|+3\sqrt{\log(K)\left(TK+\sum_{t=1}^T\tilde{d}_t\right)}\,. \tag{18}$$

To finish, we need to bound $|S|$, and relate the resulting bound to the bound in the theorem for an arbitrary $R\subset[T]$. To do so, we recycle a few results from Zimmert and Seldin (2019): In their Lemma 5, they show that $|S|\leq 2\sqrt{\log(K)\sum_{t=1}^T\tilde{d}_t}$. Furthermore, in the proof of their Theorem 2, they show that if $\sum_{t=1}^T\tilde{d}_t\geq 16\log(K)$, then

$$\sqrt{\sum_{t=1}^T\tilde{d}_t\log(K)} \leq 2\min_{R\subset[T]}\left(|R|+\sqrt{\sum_{t\in\bar{R}}d_t\log(K)}\right)\,.$$

Combining these results with (18) (and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for any $a, b > 0$), we obtain that the expected regret can be bounded as

$$\begin{aligned}\sum_{t=1}^{T} \left(\mathbb{E}\left[\ell_{t,A_t}\right] - \ell_{t,A^\star}\right) &\le 3\sqrt{TK\log(K)} + 5\sqrt{\log(K)\sum_{t=1}^{T}\tilde{d}_t} \\ &\le 3\sqrt{TK\log(K)} + 10\max\left\{2\log K, \min_{R\subset[T]}\left(|R| + \sqrt{\sum_{t\in\bar{R}} d_t\log(K)}\right)\right\},\end{aligned}$$

proving the bound on the expected regret.

To get the high-probability bound, we use Corollary 4.2 to get a high-probability version of (18): With $\eta_t = \gamma_t = \frac{1}{2}\sqrt{\frac{3\log(K)}{2tK+\sum_{s=1}^{t}\tau_s}}$, and defining $\tilde{d}^\star = \max_{t\in[T]}\tilde{d}_t$, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{t=1}^{T}\left(\ell_{t,A_t} - \ell_{t,A^\star}\right) \\ &\le |S| + 2\sqrt{3\log(K)\left(2KT + \sum_{t=1}^{T}\tilde{d}_t\right)} + \left(2\sqrt{\frac{2TK + \sum_{t=1}^{T}\tilde{d}_t}{3\log(K)}} + \tilde{d}^\star + 2\right)\frac{\log(2/\delta)}{2}\,. \end{aligned} \tag{19}$$

By construction, $\tilde{d}^\star \le \sqrt{\tilde{D}_T/\log(K)} = \sqrt{\sum_{t=1}^{T}\tilde{d}_t/\log(K)}$. Using this and the same steps as for the regret bound in expectation proves the high-probability bound of the theorem. □

C. Proof of Theorem 6.2

Proof. The proof follows the same lines as the proof of Theorem 3.1, but instead of bounding the losses by their maximum value, we approximate the adaptive data-dependent step-size that needs to be used to obtain a data-adaptive bound.

First, it can be easily seen that with γ_t and η_t as defined, we have

$$\eta_t \le \sqrt{\log(K)\left(\frac{4(d_t^\star)^2 + 6d_t^\star + 2}{\gamma_t} + \sum_{i=1}^{K}\sum_{s+d_s<t}\ell_{s,i}^{\text{bck}}\right)^{-1}} \le \sqrt{\log(K)\left(\sum_{i=1}^{K}\ell_{1:t,i}^{\text{fwd}}\right)^{-1}},$$

where the first inequality follows from the definition of $|eta_t$ and γ_t and the second step follows from Lemma 6.1.[7] Combining this with (7) and using the well-know AdaGrad lemma (see, e.g., Lemma 4 of McMahan, 2017, and also the proof of Corollary 3.2), we obtain

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{K}\hat{\ell}_{t,i}p_{t,i}\left(1 - \frac{\tilde{p}_{t+1,i}}{p_{t,i}}\right) &\le \sum_{t=1}^{T}\sum_{i=1}^{K}\eta_t\ell_{t,i}^{\text{fwd}} \\ &\le 2\sqrt{\log(K)\sum_{i=1}^{K}\ell_{1:T,i}^{\text{fwd}}} = 2\sqrt{\log(K)\sum_{i=1}^{K}\sum_{t=1}^{T}\sum_{s\in O_t\cup\{t\}}\hat{\ell}_{s,i}\hat{\ell}_{t,i}p_{t,i}} \\ &\le 2\sqrt{\log(K)\sum_{i=1}^{K}\sum_{t=1}^{T}\sum_{s\in O_t\cup\{t\}\cup D_t}\hat{\ell}_{s,i}\hat{\ell}_{t,i}p_{t,i}}\,.\end{aligned}$$

[7] Selecting η_t and γ_t such that the first inequality becomes an equality is possible by setting $\eta_t = \gamma_t = \left[\frac{4(d_t^\star)^2+6d_t^\star+2}{2\log(K)} + \sqrt{\frac{(4(d_t^\star)^2+6d_t^\star+2)^2}{4\log^2(K)} + \frac{\sum_{i=1}^{K}\sum_{s+d_s<t}\ell_{s,i}^{\text{bck}}}{\log(K)}}\right]^{-1}$. This would yield a slightly better but slightly uglier bound in the theorem. Our current choice of η_t is a little smaller than this, implying the inequality.

Next, from (5), we have

$$\begin{aligned}
\sum_{t=1}^{T} \hat{\ell}_t^\top (\tilde{p}_{t+1} - p^\star) &\le \log(K) \left(\frac{4(d_T^\star)^2 + 6d_T^\star + 2}{\log(K)} + \sqrt{\frac{\sum_{i=1}^K \sum_{s+d_s<T} \ell^{\text{bck}}_{s,i}}{\log(K)}} \right) \\
&= C_T + \sqrt{\log(K) \sum_{i=1}^{K} \sum_{s+d_s<T} \ell^{\text{bck}}_{s,i}} \\
&\le C_T + \sqrt{2\log(K) \sum_{i=1}^{K} \sum_{t=1}^{T} \sum_{s \in O_t \cup \{t\} \cup D_t} \hat{\ell}_{s,i} \hat{\ell}_{t,i} p_{t,i}} \,,
\end{aligned}$$

where the third step follows by Lemma 6.1. Putting everything together, taking expectation and moving it inside the square root by Jensen's inequality, we obtain

$$\sum_{t=1}^{T} \left(\mathbb{E}\left[\ell_{t,A_t}\right] - \ell_{t,A^\star} \right) \le C_T + c\sqrt{\log(K) \sum_{t=1}^{T} \sum_{s \in O_t \cup \{t\} \cup D_t} \mathbb{E}\left[\sum_{i=1}^{K} \hat{\ell}_{s,i} \hat{\ell}_{t,i} p_{t,i} \right]} \,. \tag{20}$$

What remains is to work out the expectations. To do this, notice that in our algorithm any action only affects another future action if the corresponding feedback arrives before the second action is taken. Therefore, whenever $s \in O_t \cup D_t$, the indicators $I_{s,i}$ and $I_{t,i}$ are independent given $\mathcal{H}_{\max\{s,t\}}$, and for such t and s,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{K} \hat{\ell}_{s,i} \hat{\ell}_{t,i} p_{t,i} \right] &\le \mathbb{E}\left[\sum_{i=1}^{K} (\ell_{s,i} \ell_{t,i} I_{t,i}) \frac{I_{s,i}}{p_{s,i}} \right] = \mathbb{E}\left[\sum_{i=1}^{K} \ell_{s,i} \ell_{t,i} \mathbb{E}\left[I_{t,i} \frac{I_{s,i}}{p_{s,i}} \middle| \mathcal{H}_{\max\{s,t\}} \right] \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K} \ell_{s,i} \ell_{t,i} p_{t,i} \right] = \mathbb{E}\left[\ell_{s,A_t} \ell_{t,A_t} \right] \,,
\end{aligned}$$

where the first step follows by the definition of $\hat{\ell}_j$, the second by the definition of $I_{t,i}$ and the tower rule, and the third by the fact that $\mathbb{E}\left[I_{t,i} I_{s,i}/p_{s,i} \middle| \mathcal{H}_{\max\{s,t\}} \right] = \mathbb{E}\left[I_{t,i} \middle| \mathcal{H}_{\max\{s,t\}} \right] \mathbb{E}\left[I_{s,i} \middle| \mathcal{H}_{\max\{s,t\}} \right] / p_{s,i} = p_{t,i}$ whenever $s \in O_t \cup D_t$. Furthermore, when $s = t$, the corresponding term is $\mathbb{E}\left[\hat{\ell}^2_{t,i} p_{t,i} \right] = \mathbb{E}\left[\ell^2_{t,i} I_{t,i}/p_{t,i} \right] \le \ell_{t,i}$. Substituting these in (20) completes the proof. □

C.1. Proof of Lemma 6.1

Proof. We start by fixing $t \in [T]$ and expanding the sum $\ell^{\text{fwd}}_{1:t,i}$:

$$\begin{aligned}
\ell^{\text{fwd}}_{1:t,i} &= \sum_{j=1}^{t} \sum_{s=1}^{t} \mathbb{I}\left[s \le j \le s + d_s \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} \\
&= \sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[s \le j \le s + d_s \le j + d_j \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} + \sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[s \le j \le j + d_j < s + d_s \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} \\
&= \sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[j \le s + d_s \le j + d_j \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} + \sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} - S_{t,i} \,,
\end{aligned}$$

where

$$\begin{aligned}
S_{t,i} = &\sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[s > j, j \le s + d_s \le j + d_j \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} + \sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[s > j, s \le j + d_j < s + d_s \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} \\
&+ \sum_{s=1}^{t} \sum_{j=1}^{t} \mathbb{I}\left[j + d_j = s + d_s \right] \hat{\ell}_{s,i} \hat{\ell}_{j,i} p_{j,i} \,.
\end{aligned}$$

Moving $S_{t,i}$ to the left,

$$\begin{aligned}\ell^{\text{fwd}}_{1:t,i} + S_{t,i} &= \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[j \le s + d_s \le j + d_j\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} + \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} \\ &= \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i}) \\ &= \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s < t\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i}) \\ &\quad + \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s, s + d_s \ge t\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i}) \\ &= \sum_{s:s+d_s<t}\ \sum_{j:s\le j+d_j \le s+d_s} \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i}) \\ &\quad + \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s, s + d_s \ge t\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i})\,,\end{aligned}$$

where the second step follows by swapping the names of s and j in the first sum on the r.h.s. Therefore,

$$\ell^{\text{fwd}}_{1:t,i} + S_{t,i} = \sum_{s:s+d_s<t} \ell^{\text{bck}}_{s,i} + M_{t,i}\,, \tag{21}$$

where $M_{t,i} = \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s \le j + d_j \le s + d_s, s + d_s \ge t\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}(p_{j,i} + p_{s,i})$.

To continue, note that $M_{t,i}$ and $S_{t,i}$ are non-negative. Hence, to get the results of the lemma it remains to bound the terms $M_{t,i}$ and $S_{t,i}$ from above. To that end, note

$$\begin{aligned}S_{t,i} &= \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[s > j, s \le j + d_j\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} + \sum_{s=1}^{t}\sum_{j=1}^{t} \mathbb{I}\left[j + d_j = s + d_s\right] \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} \\ &\le \sum_{j=1}^{t}\sum_{s:j\in O_s} \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} + \sum_{j=1}^{t}\sum_{s\in O_j\cup\{j\}\cup D_j} \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i}\,,\end{aligned}$$

where in the first step we have merged the first two summations, and split the last one, in the definition of $S_{t,i}$. The second step then follows from using the definition of O_s for the first term, and noting for the second term that either $j = s$, or $j < s$ (which, together with $j + d_j = s + d_s \ge s > j$, implies $j \in O_s$) or $j > s$ (which, together with $s + d_s = j + d_j \ge j > s$ implies $s \in O_j$).

Combining with the definition of $\ell^{\text{fwd}}_{1:t,i}$ and recalling the definition of D_j, we obtain

$$\begin{aligned}\sum_{s:s+d_s<t} \ell^{\text{bck}}_{s,i} &\le \sum_{s:s+d_s<t} \ell^{\text{bck}}_{s,i} + M_{t,i} \\ &= \ell^{\text{fwd}}_{1:t,i} + S_{t,i} \\ &\le \ell^{\text{fwd}}_{1:t,i} + \sum_{j=1}^{t}\sum_{s\in D_j} \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} + \sum_{j=1}^{t}\sum_{s\in O_j\cup\{j\}\cup D_j} \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i} \\ &= 2\sum_{j=1}^{t}\sum_{s\in O_j\cup\{j\}\cup D_j} \hat{\ell}_{s,i}\hat{\ell}_{j,i}p_{j,i}\,.\end{aligned}$$

This concludes the proof of the first inequality (13).

To bound $M_{t,i}$, note that the summations run up to t, and the conditions $\mathbb{I}\left[s + d_s \geq t\right]$ and $\mathbb{I}\left[s \leq j + d_j\right]$ imply, respectively, that the value of the sum is zero for $s < t - d_t^\star$ and $j < t - 2d_t^\star$ (since d_j and d_s are at most $d_t^\star$). Hence, we have

$$\begin{aligned}\sum_{i=1}^{K} M_{t,i} \;&\leq \sum_{s=\max\{1,t-d_t^\star\}}^{t} \;\sum_{j=\max\{1,t-2d_t^\star\}}^{t} \;\sum_{i=1}^{K} \left(\frac{\ell_{s,i} I_{s,i}}{p_{s,i} + \gamma_s} \ell_{j,i} I_{j,i} + \ell_{s,i} I_{s,i} \frac{\ell_{j,i} I_{j,i}}{p_{j,i} + \gamma_j} \right) \\ &\leq \sum_{s=\max\{1,t-d_t^\star\}}^{t} \;\sum_{j=\max\{1,t-2d_t^\star\}}^{t} \;\sum_{i=1}^{K} \left(\frac{I_{j,i} I_{s,i}}{\gamma_t} + \frac{I_{s,i} I_{j,i}}{\gamma_t} \right),\end{aligned}$$

using the fact that the losses are non-negative and upper-bounded by 1, the definition of $\hat{\ell}_k, k \in [T]$, and the fact that γ_t is a non-increasing sequence in t. Hence, $\sum_{i=1}^{K} M_{t,i} \leq (4{d_t^\star}^2 + 6d_t^\star + 2)/\gamma_t$. Putting back in (21) and summing over i completes the proof of the second inequality (14). □

D. Proof of (15)

Proof. By definition, if $s \in O_t$, then $s + 1 \leq t \leq s + d_T^\star$, and if $s \in D_t$ then $s - d_T^\star \leq t \leq s - 1$. Therefore, since the losses are $[0, 1]$-valued, we have

$$\sum_{t=1}^{T} \sum_{s \in O_t \cup D_t} \mathbb{E}\left[\ell_{s,A_s} \ell_{t,A_s}\right] = \sum_{s=1}^{T} \sum_{t : s \in O_t \cup D_t} \mathbb{E}\left[\ell_{s,A_s} \ell_{t,A_s}\right] \leq \sum_{s=1}^{T} \sum_{t : s \in O_t \cup D_t} \mathbb{E}\left[\ell_{s,A_s}\right] \leq 2 \sum_{s=1}^{T} d_T^\star \, \mathbb{E}\left[\ell_{s,A_s}\right].$$

Therefore,

$$\begin{aligned}R_T &\leq C_T + c' \sqrt{\sum_{t=1}^{T} \left(2d_T^\star \mathbb{E}\left[\ell_{t,A_t}\right] + \sum_{i=1}^{K} \ell_{t,i} \right)} \\ &\leq C_T + c' \sqrt{2d_T^\star (R_T + L_{T,A^*})} + c' \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{K} \ell_{t,i}} \\ &\leq C_T + 2c' \sqrt{\frac{d_T^\star}{2} R_T} + c' \sqrt{2d_T^\star L_{T,A^*}} + c' \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{K} \ell_{t,i}} \\ &\leq C_T' + \frac{R_T}{2} + c' \sqrt{2d_T^\star L_{T,A^*}} + c' \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{K} \ell_{t,i}}\,,\end{aligned}$$

where $C_T' = C_T + c'^2 d_T^\star$, and the last step uses $2\sqrt{ab} \leq a + b$. Hence,

$$R_T \leq 2C_T' + 2c' \sqrt{2d_T^\star L_{T,A^*}} + 2c' \sqrt{\sum_{i=1}^{K} L_{T,i}}\,.$$

□