Delayed Bandits: When Do Intermediate Observations Help?

Emmanuel Esposito [*1 2] **Saeed Masoudian** [*3] **Hao Qiu** [1] **Dirk van der Hoeven** [4]
Nicolò Cesa-Bianchi [1 5] **Yevgeny Seldin** [3]

Abstract

We study a K-armed bandit with delayed feedback and intermediate observations. We consider a model where intermediate observations have a form of a finite state, which is observed immediately after taking an action, whereas the loss is observed after an adversarially chosen delay. We show that the regime of the mapping of states to losses determines the complexity of the problem, irrespective of whether the mapping of actions to states is stochastic or adversarial. If the mapping of states to losses is adversarial, then the regret rate is of order $\sqrt{(K+d)T}$ (within log factors), where T is the time horizon and d is a fixed delay. This matches the regret rate of a K-armed bandit with delayed feedback and without intermediate observations, implying that intermediate observations are not helpful. However, if the mapping of states to losses is stochastic, we show that the regret grows at a rate of $\sqrt{\big(K+\min\{|\mathcal{S}|,d\}\big)T}$ (within log factors), implying that if the number $|\mathcal{S}|$ of states is smaller than the delay, then intermediate observations help. We also provide refined high-probability regret upper bounds for non-uniform delays, together with experimental validation of our algorithms.

1. Introduction

Delay is an ubiquitous phenomenon that many sequential decision makers have to deal with. For example, outcomes of medical treatments are often observed with delay, purchase events happen with delay after advertisement impressions, and acceptance/rejection decisions for scientific papers are observed with delay after manuscript submissions. The impact of delay on the performance of sequential decision makers, measured by regret, has been extensively studied under full information and bandit feedback, and in stochastic and adversarial environments. Yet, in many situations in real life *intermediate observations* may be available to the learner. For example, a health check-up might give a preliminary indication on the effect of a treatment, an advertisement click might be a precursor for an upcoming purchase, and preliminary reviews might provide some information regarding an upcoming acceptance or rejection decision. In this work we study when, and how, intermediate observations can be used to reduce the impact of delay in observing the final outcome of an action in a multi-armed bandit setting.

Online learning with delayed feedback and intermediate observations was studied by Mann et al. (2019) in a full-information setting, and then by Vernade et al. (2020) in a nonstationary stochastic bandit setting. In the paper of Vernade et al. (2020), at each time step the learner chooses an action and immediately observes a signal (also called state) belonging to a finite set. The actual loss (i.e., feedback) incurred by the learner in that time step is only received with delay, which can be fixed or random. More formally, the observed state is drawn from a distribution that only depends on the chosen action, and the incurred loss is drawn from a distribution that only depends on the observed state (and not on the chosen action), forming a Markov chain.

$$\begin{array}{ccccc} \text{Action} & \xrightarrow{\text{no delay}} & \text{State} & \xrightarrow{\text{delay } d_t} & \text{Loss} \\ A_t & & S_t = s_t(A_t) & & \ell_t(S_t) \end{array}$$

The work of Vernade et al. (2020) studies a setting, where s_t are nonstationary and ℓ_t are i.i.d. stochastic.

In this work, we consider two possible regimes for the mappings s_t from actions to states (stochastic and adversarial) and two possible regimes for the mappings ℓ_t from states to losses (also stochastic and adversarial). Altogether, we study four different regimes, defined by the combination of the first and the second mapping type.

We characterize (within logarithmic factors) the minimax re-

[*]Equal contribution [1]Università degli Studi di Milano, Milan, Italy [2]Istituto Italiano di Tecnologia, Genoa, Italy [3]University of Copenhagen, Copenhagen, Denmark [4]Korteweg-de Vries Institute for Mathematics University of Amsterdam, Amsterdam, Netherlands [5]Politecnico di Milano, Milan, Italy. Correspondence to: Emmanuel Esposito <emmanuel@emmanuelesposito.it>, Saeed Masoudian <saeed.masoudian@di.ku.dk>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

gret rates for all of them, by giving upper and lower bounds. Similar to Vernade et al., we assume that the states are observed instantaneously, and we assume that the losses are observed with delay d. We show that the minimax regret rate is fully determined by the regime of the states to losses mapping, regardless of the regime of the actions to states mapping. The results are informally summarized in the following table, where K denotes the number of actions, S denotes the number of states, and T denotes the time horizon. It is assumed that the losses belong to the $[0, 1]$ interval.

States to losses mapping	Regret (within log factors)
Adversarial	$\sqrt{(K+d)T}$
Stochastic	$\sqrt{\big(K+\min\{S,d\}\big)T}$

All of our upper bounds hold with high probability (with respect to the learner's internal randomization) irrespective of the regime of the action to states mapping.

We recall that (within logarithmic factors) the minimax regret rate in multi-armed bandits with delays without intermediate observations is of order $\sqrt{(K+d)T}$ (Cesa-Bianchi et al., 2019). Therefore, we conclude that if the mapping from states to actions is adversarial, then intermediate observations do not help (in the minimax sense), because the regret rates are the same irrespective of whether the intermediate observations are used or not, and irrespective of whether the mapping from actions to states is stochastic or adversarial. However, if the mapping from states to losses is stochastic, and the number S of states is smaller than the delay d, then intermediate observations are helpful, and we provide an algorithm, `MetaAdaBIO`, which is able to exploit them. Our result improves on the $\widetilde{\mathcal{O}}\big(\sqrt{KST}\big)$ regret bound obtained by Vernade et al. (2020) for the case of stochastic and stationary action to states mapping. Our algorithm also applies to a more general setting of non-uniform delays $(d_t)_{t\in[T]}$ where we achieve a high-probability regret bound of order $\sqrt{KT+\min\{ST,\mathcal{D}_T\}}$ (ignoring logarithmic factors). This improves upon the total delay term $\mathcal{D}_T = d_1 + \cdots + d_T$ similarly to the respective term in the fixed delay setting.

Related work Adaptive clinical trials have served an inspiration for the multi-armed bandit model (Thompson, 1933), and, interestingly, they have also pushed the field to study the effect of delayed feedback (Simon, 1977; Eick, 1988). In the bandit setting Joulani et al. (2013) have studied a stochastic setting with random delays, whereas Neu et al. (2010; 2014) have studied a nonstochastic setting with constant delays. Cesa-Bianchi et al. (2019) have shown an $\Omega(\max\{\sqrt{KT}, \sqrt{dT\ln K}\})$ lower bound for nonstochastic bandits with uniformly delayed feedback, and an upper bound matching the lower bound within logarithmic factors by using an EXP3-style algorithm (Auer et al., 2002b), whereas Zimmert & Seldin (2020) have reduced the gap to the lower bound down to constants by using a Tsallis-INF approach (Zimmert & Seldin, 2021). Follow up works have studied adversarial multi-armed bandits with non-uniform delays (Thune et al., 2019; Bistritz et al., 2019; 2022; Gyorgy & Joulani, 2021; Van der Hoeven & Cesa-Bianchi, 2022) with Zimmert & Seldin (2020) providing a minimax optimal algorithm and Masoudian et al. (2022) deriving a matching lower bound and a best-of-both-worlds extension. Two key techniques for handling non-uniform delays are skipping, introduced by Thune et al. (2019), and algorithm parametrization by the number of outstanding observations (an observed quantity at action time), as opposed to the delays (an unobserved quantity at action time), introduced by Zimmert & Seldin (2020).

Paper structure In Section 2 we provide a formal problem definition. In Section 3 we introduce two algorithms, `MetaBIO` and `MetaAdaBIO`, for the model of bandits with intermediate observations. In Section 4 we analyze both algorithms and prove high-probability regret bounds for the setting of adversarial action-state mappings and stochastic losses. In Section 5 we provide the lower bounds, and in Section 6 experimental evaluation, concluding with a discussion in Section 7.

2. Problem definition

We consider an online learning setting with a finite set $\mathcal{A} = [K]$ of $K \geq 2$ actions and a finite set $\mathcal{S} = [S]$ of $S \geq 2$ states. In each round $t = 1, 2, \ldots$ the learner picks an action $A_t \in \mathcal{A}$ and receives a state $S_t = s_t(A_t) \in \mathcal{S}$ as an intermediate observation according to some mapping $s_t \in \mathcal{S}^{\mathcal{A}}$. The learner also incurs a loss $\ell_t(S_t) \in [0, 1]$, which is only observed at the end of round $t + d_t$, where the delay $d_t \geq 0$ is revealed to the learner only when the observation is received.

The difficulty of this learning task depends on three elements all initially unknown to the learner:

- the sequence of action-state mappings $s_1, \ldots, s_T \in \mathcal{S}^{\mathcal{A}}$;
- the sequence of loss vectors $\ell_1, \ldots, \ell_T \in [0, 1]^S$;
- the sequence of delays $d_1, \ldots, d_T \in \mathbb{N}$, where $d_t \leq T - t$ for all $t \in [T]$ without loss of generality.

Note that unlike standard bandits, here the losses are functions of the states instead of the actions. However, since actions are chosen without a-priori information on the action-state mappings, learners have no direct control on the losses they will incur and, because of the delays, they also have no immediate feedback on the loss associated with the observed states. Note also that, for all $t \geq 1$, the states $s_t(a)$ for $a \neq A_t$ and the losses $\ell_t(s)$ for $s \neq S_t$ are never re-

vealed to the algorithm. For brevity, we refer to this setting as (delayed) Bandits with Intermediate Observations (BIO).

In the setting of stochastic losses, we assume the loss vectors $\ell_t \in [0,1]^S$ are sampled i.i.d. from some fixed but unknown distribution Q, and let $\theta \in [0,1]^{\mathcal{S}}$ be the unknown vector of expected losses for the states. That is, $\ell_t(s) \sim Q(\cdot \mid s)$ has mean $\theta(s)$ for each $t \in [T]$ and $s \in \mathcal{S}$. Note that we allow dependencies between the stochastic losses of distinct states in the same round, but require losses to be independent across rounds. In the setting of stochastic action-state mappings, we assume that each observed state S_t is independently drawn from a fixed but unknown distribution $P(\cdot \mid A_t)$. If both losses and action-state mappings are stochastic, then $\ell_t(S_t)$ is independent of A_t given S_t. When losses or action-state mappings are adversarial, we always assume oblivious adversaries.

Our main quantity of interest is the regret measured via the learner's cumulative loss $\sum_t \ell_t(S_t)$, where $S_t = s_t(A_t)$ and $(A_t)_{t\geq 1}$ is the sequence of learner's actions. In case of stochastic losses, we define the learner's performance by $\sum_t \theta(S_t)$. In case of stochastic action-state mappings, we average each instantaneous loss over the random choice of the state: $\sum_s \ell_t(s)P(s \mid A_t)$ for adversarial losses and $\sum_s \theta(s)P(s \mid A_t)$ for stochastic losses. Regret is always computed according to the best action with respect to appropriate notion of cumulative loss. In particular, for stochastic state-action mappings, the cumulative losses of the best action are

$$\min_{a\in\mathcal{A}} \sum_{t=1}^{T} \sum_{s\in\mathcal{S}} \ell_t(s)P(s \mid a)$$

and

$$\min_{a\in\mathcal{A}} \sum_{t=1}^{T} \sum_{s\in\mathcal{S}} \theta(s)P(s \mid a) \ .$$

3. Algorithm

In this section we introduce `MetaBIO` (Algorithm 1) that transforms any algorithm $\mathcal{B}$ tailored for the delayed setting *without* intermediate observations into an algorithm for our setting. We then propose `MetaAdaBIO`, a modification of `MetaBIO` that delivers an improved regret bound for our setting.

The idea of `MetaBIO` is to reduce the impact of delays using the information we get from intermediate observations. More precisely, if we have *enough* observations for the current state S_t at time t, we immediately feed to $\mathcal{B}$ the *estimate* of the mean loss of this state as if it were the actual loss at time t; otherwise, we wait for d_t time steps and refine our estimate using the additional loss observations.

The are two key steps in the design of our algorithm: *how* we construct the mean estimate and *when* we use it instead of waiting for the actual loss. They are the steps highlighted in green in Algorithm 1. For all $t \in [T]$ and $s \in \mathcal{S}$, we use $\widetilde{\theta}_t(s)$ to denote the mean estimate of $\theta(s)$ at round t and $n_t(s)$ to denote the number of observations for state s that we want to observe before using $\widetilde{\theta}_t(s)$. We add a subscript t to $\mathcal{L}(s)$ in Algorithm 1 to denote the set of observations we have collected at the end of round t. Thus, $\widetilde{\theta}_t(s)$ uses $N_t(s) = |\mathcal{L}_t(s)|$ observations.

Algorithm 1: `MetaBIO`

Input: Algorithm $\mathcal{B}$ for standard delayed bandits, confidence parameter $\delta \in (0,1)$
Initialize $\mathcal{L}(s) = \emptyset$ *for all* $s \in \mathcal{S}$
for $t = 1, \dots, T$ **do**
  Get A_t from $\mathcal{B}$
  Observe $S_t = s_t(A_t)$
  for $j : j + d_j = t$ **do**
    Receive $(j, \ell_j(S_j))$
    Update $\mathcal{L}(S_j) = \mathcal{L}(S_j) \cup \{(j, \ell_j(S_j))\}$
  Initialize feedback set $\mathcal{M} = \emptyset$
  Compute $n_t(S_t)$
  if $|\mathcal{L}(S_t)| \geq n_t(S_t)$ **then**
    Add t to $\mathcal{M}$
  for $j : j + d_j = t \wedge |\mathcal{L}(S_j)| < n_j(S_j)$ **do**
    Add j to $\mathcal{M}$
  for $j \in \mathcal{M}$ **do**
    Compute $\widetilde{\theta}_t(S_j)$ from $\mathcal{L}(S_j)$ `// using` δ
    Feed $(j, A_j, \widetilde{\theta}_t(S_j))$ to $\mathcal{B}$

Fixed delay setting. When all rounds have delay d, we simply choose $n_t(s) = d$ for all $s \in \mathcal{S}, t \in [T]$. In other words, if we have at least d observations for some state, then we can compensate for the effect of delays and construct a well concentrated mean estimate around the actual mean. Let $\widehat{\theta}_t(s) = \sum_{j\in\mathcal{L}_t(s)} \ell_j(s)/N_t(s)$. Then our mean loss estimate is a lower confidence bound for $\theta(s)$ defined by

$$\widetilde{\theta}_t(s) = \max\left\{0, \widehat{\theta}_t(s) - \frac{1}{2}\varepsilon_t(s)\right\} \tag{1}$$

for $\varepsilon_t(s) = \sqrt{\frac{2}{N_t(s)} \ln \frac{4ST}{\delta}}$.

Arbitrary delay setting. In the arbitrary delay setting, where we do not have preliminary knowledge of delays, we can not use the delays to set $n_t(s)$. Instead, at the *end* of time t, we have access to the number of outstanding observations $\sigma_t = \left|\{j \in [t] \ : \ j + d_j > t\}\right|$, which is different from prior works that consider outstanding observation at the *beginning* of the round. Then, for any $s \in \mathcal{S}$, we may set $n_t(s) = \sigma_t$. With this choice, incurring zero delay at some round implies that we received at least half of all the observations we could

have received in the no-delay setting (see Appendix B.4). In Section 4 we see that this ensures our mean estimate is well concentrated around its mean.

Since Algorithm 1 waits for the actual loss at time t only if $N_t(S_t) < \sigma_t$, then $\widetilde{d}_t = d_t \, \mathbb{1}[N_t(S_t) < \sigma_t]$ is the actual delay incurred by the algorithm, and $\mathcal{L}_{t+\widetilde{d}_t}(s)$ is the set of observations used to compute the estimate of the mean loss at time t. Because some observations may arrive at the same time, the high-probability analysis of `MetaBIO` requires these observations to be ordered. More precisely, we construct our mean estimate at time $t + \widetilde{d}_t$ for the feedback of round t using the set

$$\mathcal{L}'_t(s) = \left\{ (j, \ell_j(s)) \in \mathcal{L}_{t+\widetilde{d}_t}(s) \,\middle|\, j + \widetilde{d}_j = t + \widetilde{d}_t \Rightarrow j < t \right\}.$$

Letting $N'_t(s) = |\mathcal{L}'_t(s)|$, we define the empirical mean

$$\widehat{\theta}_t(s) = \sum_{j \in \mathcal{L}'_t(s)} \frac{\ell_j(s)}{N'_t(s)} \quad . \tag{2}$$

Then, we set $\varepsilon_t(s) = \sqrt{\frac{2}{N'_t(s)} \ln \frac{4ST}{\delta}}$ and define the mean loss estimator similarly to Equation (1).

The `MetaAdaBIO` algorithm. As we said already, the goal of intermediate observations is to reduce the impact of delays. However, if the number of states is too large compared to the average delay, then the information we get from intermediate observations could be misleading. To address this issue, we introduce `MetaAdaBIO` (Algorithm 2). Given a horizon T,[1] this algorithm runs $\mathcal{B}$ (which is tailored for the setting *without* intermediate observations) until the total incurred delay exceeds ST, and then switches to `MetaBIO`. We precise that `MetaAdaBIO` computes $\mathfrak{D}_t$ as the sum of outstanding observation counts up to round t, which is then used in the switching condition.

Algorithm 2: `MetaAdaBIO`

Input: Algorithm $\mathcal{B}$ for standard delayed bandits, confidence parameter $\delta \in (0, 1)$
Initialize $\mathfrak{D}_0 = 0$
for $t = 1, \ldots, T$ **do**
  Get A_t from $\mathcal{B}$
  for $j : j + d_j = t$ **do**
    Receive $(j, \ell_j(S_j))$
    Feed $(j, A_j, \ell_j(S_j))$ to $\mathcal{B}$
  Set $\sigma_t = \sum_{j=1}^{t-1} \mathbb{1}[j + d_j > t]$
  Update $\mathfrak{D}_t = \mathfrak{D}_{t-1} + \sigma_t$
  if $\mathfrak{D}_t(3 \ln K + \ln(6/\delta)) > 49ST \ln \frac{8ST}{\delta}$ **then**
    break
if $t < T$ **then**
  Run `MetaBIO`$(\mathcal{B}, \delta/2)$ for the remaining rounds

4. Regret Analysis

We analyze `MetaBIO` and `MetaAdaBIO` in the setting of adversarial action-state mappings and stochastic losses where the regret is defined by

$$R_T = \sum_{t=1}^{T} \theta(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a)) \quad .$$

Our analysis guarantees a bound on R_T that holds with high probability (and not just in expectation). A related notion of regret is

$$\mathcal{R}_T = \sum_{t=1}^{T} \ell_t(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(s_t(a))$$

which considers the realized losses instead of their means. The two quantities are close with high probability: each inequality

$$-\sqrt{2T \ln(2K/\delta)} \le R_T - \mathcal{R}_T \le \sqrt{2T \ln(2/\delta)} \tag{3}$$

individually holds with probability at least $1 - \delta$ for any given $\delta \in (0, 1)$ (see Lemma A.1).

Let $\mathcal{D}_T = \sum_{t=1}^{T} d_t$ be the total delay. We start by showing an upper bound on the total actual delay $\widetilde{\mathcal{D}}_T = \sum_{t=1}^{T} d_t \mathbb{1}[N_t(S_t) < \sigma_t] \le \mathcal{D}_T$ incurred by `MetaBIO`. Then, we provide a high-probability regret analysis of both `MetaBIO` and `MetaAdaBIO`.

More precisely, we can show that `MetaBIO` incurs the delays of no more than $\min\{2S\sigma_{\max}, T\}$ rounds, where $\sigma_{\max} = \max_{t \in [T]} \sigma_t$. In the worst case, these rounds correspond with those from the set

$$\Phi \in \arg\max_{\mathcal{J} \subseteq [T]} \left\{ \mathcal{D}_{\mathcal{J}} : |\mathcal{J}| = \min\{2S\sigma_{\max}, T\} \right\} \quad . \tag{4}$$

where we denote $\mathcal{D}_{\mathcal{J}} = \sum_{t \in \mathcal{J}} d_t$ for any $\mathcal{J} \subseteq [T]$. Note that the set Φ is fully determined by the delay sequence $d_1, \ldots, d_T$. Moreover, the total delay incurred by `MetaBIO` cannot be worse than the sum of delays corresponding to the rounds in Φ, as stated in the lemma below.

Lemma 4.1 (Total actual delay)**.** *If* `MetaBIO` *is run with any algorithm $\mathcal{B}$ on delays $(d_t)_{t \in [T]}$, then $\widetilde{\mathcal{D}}_T \le \mathcal{D}_\Phi$.*

Lemma 4.1 (proof in Appendix B.1) implies that, if all delays are bounded by $d_{\max}$, then $\widetilde{\mathcal{D}}_T \le 2S\sigma_{\max} d_{\max}$, which does not depend on T. In the fixed-delay setting with delay d, for example, we get a total effective delay of at most $2Sd^2$, rather than the total delay dT we would incur

[1] Note that we may remove the a-priori knowledge of T by using a doubling trick at the cost of a polylog factor in the regret. See Remark 4.7 for further details.

without access to intermediate observations (when T is large enough).

We now turn `MetaBIO` into a concrete algorithm by instantiating $\mathcal{B}$. Specifically, we use `DAda-Exp3` (Gyorgy & Joulani, 2021), a variant of `Exp3` which does not use intermediate observations and is robust to delays. `DAda-Exp3` has the following regret bound.

Theorem 4.2 (Gyorgy & Joulani (2021, Corollary 4.2))**.** *For any $\delta \in (0,1)$, the regret with respect to realized losses of* `DAda-Exp3` *in the adversarial bandits with arbitrary delays with probability at least $1-\delta$ satisfies*

$$\mathcal{R}_T \leq 2\sqrt{3(2KT+\mathcal{D}_T)\ln K} + \left(\sqrt{\frac{2KT+\mathcal{D}_T}{3\ln K}} + \frac{\sigma_{\max}}{2} + 1\right)\ln\frac{2}{\delta} \ .$$

While Theorem 4.2 shows a high-probability bound on $\mathcal{R}_T$, Equation (3) shows that a high-probability bound for one notion of regret ensures a high-probability bound for the other. Although the original bound by Gyorgy & Joulani (2021) was stated with $d_{\max}$ instead of $\sigma_{\max}$, we can replace the former with the latter by observing that, in the analysis of Gyorgy & Joulani (2021, Theorem 4.1), they only use $d_{\max}$ to upper bound the number of outstanding observations. Note that $\sigma_{\max}$ is never larger than $d_{\max}$, indicating it is a well-behaved term that is not vulnerable to a few large delays. See Masoudian et al. (2022, Lemma 3) for a refined quantification of the relation between $\sigma_{\max}$ and $d_{\max}$.

If we consider a fixed confidence level $\delta \in (0,1)$, then we can make the learning rate η_t and the implicit exploration term γ_t in `DAda-Exp3` depend on the specific value of δ so as to achieve an improved regret bound (see Appendix B.2). This allows us to show that in the BIO setting with adversarial action-state mappings and stochastic losses, the regret $\mathcal{R}_T$ of `DAda-Exp3` is upper bounded by

$$2\sqrt{2KTC_{K,6\delta}} + 2\sqrt{D_T C_{K,6\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{2}{\delta} \qquad (5)$$

with probability at least $1-\delta$, where

$$C_{K,\delta} = 3\ln K + \ln\frac{12}{\delta} \ .$$

Next, we state the regret bound for `MetaBIO`. We remark that we initialize `DAda-Exp3` with confidence parameter $\delta/2$ so as to guarantee the high-probability bound as in (5) with probability at least $1-\delta/2$ as required.

Theorem 4.3. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *using* `DAda-Exp3`*, then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T \leq 2\sqrt{2KTC_{K,3\delta}} + 7\sqrt{ST\ln\frac{4ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{4}{\delta} \ . \qquad (6)$$

We begin the analysis of Theorem 4.3 by decomposing the regret into two parts: (i) the regret $\mathcal{R}_T$ of `DAda-Exp3` with losses $\widetilde{\theta}_t(S_t)$, and (ii) the gap $R_T - \mathcal{R}_T$, corresponding to the cumulative error of the estimates fed to `DAda-Exp3`. For the first part, we follow an approach similar to Gyorgy & Joulani (2021) and apply Neu (2015, Lemma 1) to obtain a concentration bound for the loss estimates defined using importance weighting along with implicit exploration. When using the actual losses, the application of Neu (2015, Lemma 1) is straightforward. However, when the mean loss estimate $\widetilde{\theta}_t(S_t)$ is used rather than the actual loss, there is a potential dependency between the chosen action A_t and $\widetilde{\theta}_t(S_t)$. In Appendix B.3 we carefully design a filtration to show that we may indeed use the high-probability regret bound of `DAda-Exp3` in order to upper bound the first part (regret $\mathcal{R}_T$ defined in terms of the estimates $\widetilde{\theta}_t$).

The second part requires to bound the cumulative error of our estimator in (2) for the observed states $\{S_t\}_{t\in[T]}$. To this end, we use the Azuma-Hoeffding inequality to control the error of these estimates. Doing so causes a $\widetilde{\mathcal{O}}(\sqrt{ST})$ term to appear in the regret bound. The detailed proof of this part is in Appendix B.4, together with the proof of Theorem 4.3.

The presence of the $\widetilde{\mathcal{O}}(\sqrt{ST})$ term in the regret bound implies that, when $S \gg \max\{\mathcal{D}_T/T, K\}$, using intermediate feedback leads to no advantage over ignoring it. So we ideally want to recover the original bound in (5) when this happens. `MetaAdaBIO` solves this issue and gives the following regret guarantee. The proof of this result is deferred to Appendix B.5. We remark that, to achieve this bound, before the eventual switch we use algorithm `DAda-Exp3` with confidence parameter set to $\delta/3$ so as to guarantee a high-probability bound on R_{t^*} with probability at least $1-\delta/2$ over the first t^* rounds that `DAda-Exp3` runs by itself.

Theorem 4.4. *Let $\delta \in (0,1)$. If we run* `MetaAdaBIO` *with* `DAda-Exp3`*, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T \leq 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 6\sqrt{KTC_{K,2\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta} \ . \qquad (7)$$

If we consider any upper bound $d_{\max}$ on the delays

$(d_t)_{t\in[T]}$, we can further observe that the regret R_T of MetaAdaBIO (with DAda-Exp3) satisfies

$$R_T = \widetilde{\mathcal{O}}\Big(\sqrt{KT} + \min\Big\{\sqrt{S}\big(\sqrt{T} + d_{\max}\big), \sqrt{d_{\max}T}\Big\}\Big)$$

with high probability. This also follows from the fact that, as previously mentioned, we can bound the total delay of MetaBIO by $\mathcal{D}_\Phi \le 2Sd_{\max}^2$.

Given the previous regret bounds, we observe that we may further improve the dependency on the delays by adopting the idea of skipping rounds with large delays when computing the learning rates. This "skipping" idea was introduced by Thune et al. (2019) and has been leveraged by Gyorgy & Joulani (2021) to show that DAda-Exp3 can achieve a refined high-probability regret bound—see Gyorgy & Joulani (2021, Theorem 5.1). As a consequence, we can indeed provide an improved bound in our setting by following similar steps as in the proof of Theorem 4.3. The only main change is the adoption of the version of DAda-Exp3 that uses the skipping procedure.

Corollary 4.5. *Let $\delta \in (0,1)$. If we run* MetaBIO *with* DAda-Exp3 *with skipping (Gyorgy & Joulani, 2021, Theorem 5.1), then the regret of* MetaBIO *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T = \mathcal{O}\Bigg(\sqrt{KTC_{K,\delta}} + \sqrt{ST\ln\frac{ST}{\delta}} + \ln\frac{1}{\delta} + \sqrt{C_{K,\delta}\ln K}\min_{R\subseteq\Phi}\Big\{|R| + \sqrt{\mathcal{D}_{\Phi\setminus R}\ln K}\Big\}\Bigg).$$

This result could also be extended in a similar way to MetaAdaBIO, so as to achieve the best result from the presence of intermediate feedback.

So far, we have provided some high-probability guarantees for the regret of both MetaBIO and MetaAdaBIO, by which we can derive some expectation bounds as well (e.g., by setting $\delta \approx 1/T$). However, using the empirical mean estimators $\widehat{\theta}_t$ as the mean loss estimators at time t and working directly with the expected regret allows us to improve the achievable bound by a polylogarithmic factor. Hence, for the expected regret we use Tsallis-INF (Zimmert & Seldin, 2020), a learning algorithm for the standard delayed bandit problem that uses a hybrid regularizer to deal with delays and gives a minimax-optimal expected regret bound. The proof of this expected regret upper bound is in Appendix B.6.

Proposition 4.6. *If we execute* MetaAdaBIO *with* Tsallis-INF *(Zimmert & Seldin, 2020), and use the switching condition $\sqrt{8\mathfrak{D}_t\ln K} > 6\sqrt{ST\ln(2ST)}$ at each round $t\in[T]$, where $\mathfrak{D}_t = \sum_{j=1}^t \sigma_j$, then the regret of* MetaAdaBIO *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$\mathbb{E}[R_T] \le 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi\ln K} + 2\min\Big\{6\sqrt{ST\ln(2ST)}, \sqrt{8\mathcal{D}_T\ln K}\Big\}.$$

Remark 4.7. In MetaBIO, we can replace T by t^2 in the definition of the confidence intervals for (2) and remove the need for prior knowledge of the time horizon T. In MetaAdaBIO, we could use a doubling trick to avoid the prior knowledge of T in the switching condition. On the other hand, it is not required to know the number of states S for expectation bounds on the regret of MetaBIO. However, removing the prior knowledge of S in the high-probability regret bounds is challenging. Indeed, to the best of our knowledge, there is no result in BIO that avoids prior knowledge on the number of states. Lifting this requirement in the high-probability analysis is thus an interesting question for future work.

5. Lower Bounds

The lower bounds in this section are for the expected regret $\mathbb{E}[R_T]$. Since our algorithms provide high-probability guarantees, the upper bounds also apply to the expected regret. Throughout this section we will make use of constant delay i.e. $d_t = d$ for all $t\in[T]$. We will first prove a general $\sqrt{KT}$ lower bound for all algorithms in BIO, after which we specialize to particular cases.

We start by proving a $\Omega(\sqrt{KT})$ lower bound for any algorithm in our setting and for any combination of stochastic or adversarial action-state mappings and loss vectors. The construction is a reduction to the standard bandits lower bound construction.

Theorem 5.1. *Irrespective to whether the action-state mappings and loss vectors are stochastic or adversarial, there exists a sequence of losses such that any (possibly randomized) algorithm in BIO suffers regret $\mathbb{E}[R_T] = \Omega(\sqrt{KT})$.*

Proof. Our construction only uses two states h_1 and h_2. The loss vectors, which are deterministic and do not change over time, are defined as follows: $\ell_t(h_1) = 1$ and $\ell_t(h_2) = 0$ for all $t \ge 0$. The stochastic action-state mapping, which is also constant over time, is given by

$$s_t(a) = \begin{cases} h_1 & \text{with probability } p_a \\ h_2 & \text{with probability } 1-p_a \end{cases}$$

for all $a\in\mathcal{A}$ and $t\ge 0$, where the probabilities p_a are to be determined. Thus, the loss of an arm a is $\ell_t(s_t(a)) = \ell_t(h_1) = 1$ with probability p_a and $\ell_t(s_t(a)) = \ell_t(h_2) = 0$ with probability $1-p_a$. Since the loss is determined by the state, the learner receives bandit feedback without

delay. We can then choose p_a for $a \in \mathcal{A}$ to mimic the standard $\Omega(\sqrt{KT})$ distribution-free bandit lower bound—e.g., see Slivkins et al. (2019, Chapter 2). By Yao's minimax principle, the same lower bound also applies to the case with adversarial action-state mappings. Since the loss vectors are deterministic, this covers all possible cases in BIO. □

Adversarial action-state mapping and stochastic losses. We first prove a lower bound $\sqrt{ST}$ for any number $K \geq 2$ of actions. However, we do need a minor generalization of our setting to allow correlation between unseen losses. Specifically, we allow all pairs of losses $\ell_j(s), \ell_{j'}(s')$ of distinct states $s \neq s'$ to be correlated if $j > j'$ and $j - j' \leq d$, while we guarantee the i.i.d. nature of losses for any fixed state. Since $\mathbb{E}[\ell_t(S_t)] = \mathbb{E}[\theta(S_t)]$, this does not affect the analysis for the upper bound on the regret of our algorithms since $\mathbb{E}[R_T] \leq \mathbb{E}[\mathcal{R}_T]$ (see Lemma A.3). However, for a high-probability upper bound, we need to relate R_T and $\mathcal{R}_T$, which now leads to an additive $\widetilde{\mathcal{O}}(\sqrt{ST})$ term rather than an additive $\widetilde{\mathcal{O}}(\sqrt{T})$ term as in Equation (3).

In the proof of the $\sqrt{ST}$ lower bound, we leverage the fact that losses are independent only across time steps for a fixed state, while they may depend on the losses of the other states. Note that our lower bound holds even when the learner knows the action-state assignments beforehand.

Theorem 5.2. *Suppose that the action-state mapping is adversarial and the losses are stochastic and that $d_t = d$ for all $t \in [T]$. If $T \geq \min\{S, d\}$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\sqrt{\min\{S, d\}T}\big)$.*

We provide a sketch of the proof of Theorem 5.2 (see Appendix C for the full proof). First, suppose that $S \leq 2d$. For the construction of the lower bound we only consider two actions and equally split the states over these two actions. Then, we divide the T time steps in blocks of length $S/2 \leq d$. In each block, each state has the same loss. Since the block length is smaller then the delay, we have effectively created a two-armed bandit problem with $T' = T/(S/2)$ rounds and loss range $[0, S/2]$, for which we can prove a $\Omega\big(S\sqrt{T'}\big) = \Omega\big(\sqrt{ST}\big)$ lower bound by showing an equivalent lower bound for the full information setting. If $S > 2d$, we use the same construction with only $2d$ states, and obtain a $\Omega\big(\sqrt{dT}\big)$ lower bound.

Finally, we can show the following lower bound, whose proof can be found in Appendix C.

Theorem 5.3. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq d + 1$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret*

$$\mathbb{E}[R_T] = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big) \ .$$

This term is also present in the dynamic regret bound of `NSD-UCRL2`, but it is necessarily incurred from their analysis even in the stationary case (Vernade et al., 2020, Theorem 1).

This last lower bound implies that the regret of our algorithm is near-optimal. Since the lower bound of Theorem 5.1 applies to the case where the action-state mapping is adversarial and the losses are stochastic, we find the following result as a corollary of Theorem 5.1, Theorem 5.2, and Theorem 5.3.

Corollary 5.4. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq 1 + \min\{S, d\}$, then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret*

$$\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{\min\{S, d\}T}, (d+1)\sqrt{S}\big\}\big) \ .$$

Stochastic action-state mappings and adversarial losses. In this case we recover the standard lower bound for adversarial bandits with bounded delay.

Theorem 5.5. *Suppose that the action-state mapping is stochastic, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists a stochastic action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{dT}\big\}\big)$.*

Proof. Since by Theorem 5.1 we already know that any algorithm must suffer $\Omega\big(\sqrt{KT}\big)$ regret, we only need to show a $\Omega(\sqrt{dT})$ lower bound. We use two states, h_1 and h_2. Our action-state mapping is deterministic and, for all $t \geq 0$, assigns $s_t(a) = h_1$ to all but one action $a^\star$, to which the mapping assigns $s_t(a^\star) = h_2$. We now have constructed a two-armed bandit problem with delayed feedback and T rounds, for which a $\Omega(\sqrt{dT})$ lower bound is known (Cesa-Bianchi et al., 2019). □

Adversarial action-state mappings, adversarial losses. Since we can recover the construction of the lower bound in Theorem 5.5, we have the following result.

Corollary 5.6. *Suppose that the action-state mapping is adversarial, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists an action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{dT}\big\}\big)$.*

6. Experiments

We empirically compare our algorithm `MetaBIO` with the following baselines: `DAda-Exp3` (Gyorgy & Joulani, 2021) for adversarial delayed bandits without intermediate observations (which we used to instantiate the algorithm $\mathcal{B}$), the standard `UCB1` algorithm (Auer et al., 2002a) for stochastic bandits without delays and intermediate observations, and `NSD-UCRL2` (Vernade et al., 2020) for non-stationary stochastic action-state mappings and stochastic losses. We run all experiments with a time horizon of $T = 10^4$. All our plots show the cumulative regret of the algorithms considered as a function of time. The performance of each algorithm is averaged over 20 independent runs in every experiment, and the shaded areas consider a range centered around the mean with half-width corresponding to the empirical standard deviation of these 20 repetitions. In the first two experiments, we consider both fixed delays $d \in \{50, 100, 200\}$ and random delays $d_t \sim \text{Laplace}(50, 25)$ sampled i.i.d. from the Laplace distribution with $\mathbb{E}[d_t] = 50$.

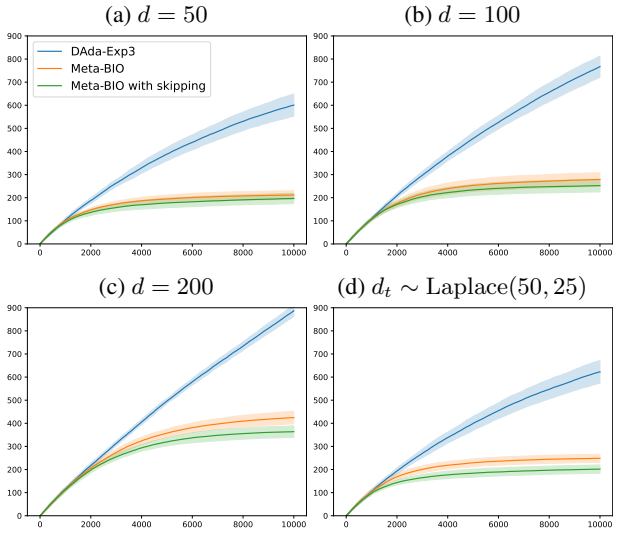

Figure 1. Cumulative regret over time for the stochastic action-state mapping when delays are fixed or random.

Experiment 1: stochastic action-state mappings. Here we use a stationary version of the experiments in (Vernade et al., 2020)—see Table 1 in Appendix D for details. We set $K = 4$ and $S = 3$, while we repeat this experiment for the previously mentioned values of delays. Figure 1 shows that, across all delay regimes, `MetaBIO` largely improves on the performance of `DAda-Exp3` by exploiting intermediate observations.

Experiment 2: adversarial action-state mappings. In this construction, we simulate the adversarial mapping using a construction adapted from (Zimmert & Seldin, 2021): we alternate between two stochastic mappings while keeping the loss means fixed. We set $K = 4$, $S = 3$, and we consider multiple instances for the different values of delays as in the previous experiment. The interval between two consecutive changes in the distribution of action-state mappings grows exponentially. See Table 2 in Appendix D for details. Figure 2 shows that `MetaBIO` and `MetaBIO` with "skipping" outperform both `UCB1` and `NSD-UCRL2`.

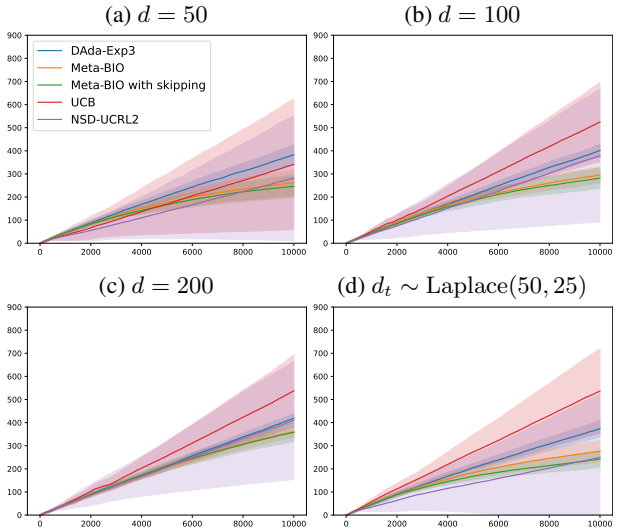

Figure 2. Cumulative regret over time for the adversarial action-state mapping when delays are fixed or random. All algorithms have small variance except for `UCB1` and `NSD-UCRL2`.

Experiment 3: utility of intermediate observations. Here we set $K = 8$, $d = 100$, and investigate how the performance of `MetaBIO` changes when the number S of states varies in $\{4, 6, 8, 10, 12\}$. The mean loss is always 0.2 for the optimal state and 1 for the others. The optimal action always maps to the optimal state. The suboptimal actions map to the optimal state with probability 0.6 and map to a random suboptimal state with probability 0.4. This implies that the expected loss of each arm remains constant when the number of states changes. Figure 3 shows that the regret gap between `MetaBIO` and `DAda-Exp3` shrinks as the number of states increases. This observation confirms our theoretical findings about the dependency of the regret on the number of states, which lead to a larger improvement the fewer they are.

Experiment 4: performance of `MetaAdaBIO` when $S < d$. We use the same setting as in Experiment 1 with delay $d = 20$.[2] Figure 4 shows the performance of `MetaAdaBIO` compared with both `DAda-Exp3` and

[2]Compared to the switching condition used for the analysis of `MetaAdaBIO`, we replace $49ST \ln \frac{8ST}{\delta}$ with ST. This change allows the switching condition to be triggered more easily to provide a better visualization of the behaviour of `MetaAdaBIO`, while it only introduces a polylog factor in its regret bound.

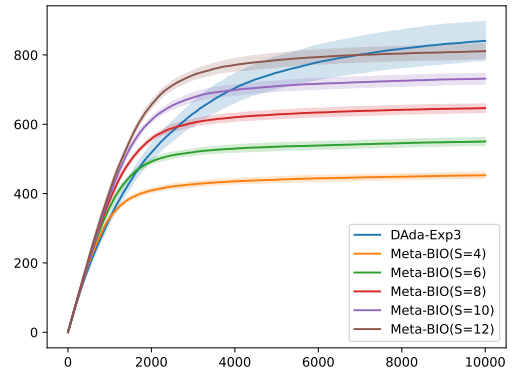

Figure 3. Cumulative regret over time of both `DAda-Exp3` and `MetaBIO` with different numbers of states $S \in \{4, 6, 8, 10, 12\}$.

`MetaBIO`. Before the switching point, `MetaAdaBIO` runs `DAda-Exp3` (up to independent internal randomization). Afterwards, `MetaAdaBIO` switches to `MetaBIO` (which in turn runs `DAda-Exp3` as a subroutine) and quickly aligns with its performance. Note that, at the switching time, `MetaAdaBIO` uses (via `MetaBIO`) the same instance of `DAda-Exp3` that was already running, rather than starting a new instance. It can be shown that our analysis of `MetaAdaBIO` applies to this variant as well without changes in the order of the bound.

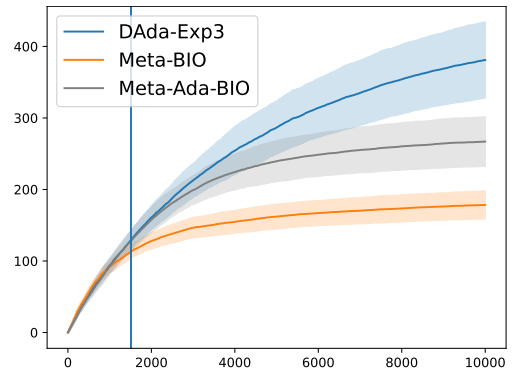

Figure 4. Cumulative regret over time of `DAda-Exp3`, `MetaBIO` and `MetaAdaBIO`. The vertical blue line marks the switching point of `MetaAdaBIO`.

Experiment 5: performance of `MetaAdaBIO` when $S > d$. We use a setting that is almost identical to that of Experiment 3 (Section 6), except we set $d = 4$ and $S = 14$. The performance of the three algorithms is shown in Figure 5. We can observe that `MetaAdaBIO` does not switch to `MetaBIO` and its performance is thus the same as that of `DAda-Exp3`, whereas `MetaBIO` incurs a larger regret.

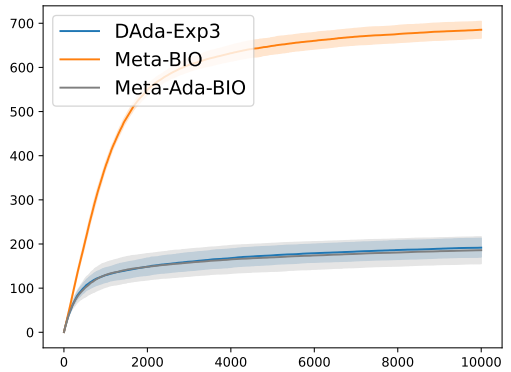

Figure 5. Cumulative regret over time of `DAda-Exp3`, `MetaBIO` and `MetaAdaBIO` when $S > d$.

7. Future Work

The work of Vernade et al. (2020) also considers a non-stationary action-state mapping and derive regret bounds for the switching regret. Preliminary results suggest that, as long as there is an algorithm that can provide bounds on the switching regret with delayed feedback, our ideas also transfer to this setting. Unfortunately, there is currently no algorithm that can provide bounds on the switching regret with delayed feedback and we leave this as a promising direction for future work.

Acknowledgements

EE, NCB, and HQ are partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR), by the EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence), and by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence). SM acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199. YS acknowledges partial support by the Independent Research Fund Denmark, grant number 9040-00361B. This work was mostly done while DvdH was at the University of Milan partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and partially supported by Netherlands Organization for Scientific Research (NWO), grant number VI.Vidi.192.095.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3), 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), 2002b.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. H. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2019.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. No weighted-regret learning in adversarial bandits with delays. *Journal of Machine Learning Research*, 23, 2022.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20, 2019.

Eick, S. G. The two-armed bandit with delayed responses. *The Annals of Statistics*, 1988.

Gyorgy, A. and Joulani, P. Adapting to delays and data in adversarial multi-armed bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 2021.

Van der Hoeven, D. and Cesa-Bianchi, N. Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Joulani, P., Gyorgy, A., and Szepesvári, C. Online learning under delayed feedback. In *International Conference on Machine Learning*, 2013.

Mann, T. A., Gowal, S., György, A., Hu, H., Jiang, R., Lakshminarayanan, B., and Srinivasan, P. Learning from delayed outcomes via proxies with applications to recommender systems. In *International Conference on Machine Learning*, 2019.

Masoudian, S., Zimmert, J., and Seldin, Y. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2022.

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

Neu, G., György, A., Szepesvári, C., and Antos, A. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, 2010.

Neu, G., György, A., Szepesvári, C., and Antos, A. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59, 2014.

Simon, R. Adaptive treatment assignment methods and clinical trials. *Biometrics*, 33, 1977.

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

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933.

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

Vernade, C., György, A., and Mann, T. A. Non-stationary delayed bandits with intermediate observations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 2020.

Zimmert, J. and Seldin, Y. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics*, 2020.

Zimmert, J. and Seldin, Y. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22, 2021.

A. Auxiliary Results

Lemma A.1. *Consider any algorithm that picks actions $(A_t)_{t\in[T]}$ in the adversarial delayed bandits problem with intermediate feedback with arbitrary action-state mappings $(s_t)_{t\in[T]}$ and i.i.d. loss vectors $(\ell_t)_{t\in[T]}$. Then, for any given $\delta \in (0,1)$,*

$$R_T - \mathcal{R}_T \le \sqrt{2T\ln(2/\delta)} \qquad \text{and} \qquad \mathcal{R}_T - R_T \le \sqrt{2T\ln(2K/\delta)}$$

individually hold with probability at least $1-\delta$.

Proof. First, observe that we can relate the two notions of regret as

$$R_T = \mathcal{R}_T + \sum_{t=1}^{T}\big(\theta(S_t) - \ell_t(S_t)\big) + \underbrace{\min_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell_t(s_t(a)) - \min_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a))}_{(\triangle)} \ .$$

By Azuma-Hoeffding inequality, we can show that each side of

$$-\sqrt{\frac{T}{2}\ln\left(\frac{1}{\delta'}\right)} \le \sum_{t=1}^{T}\big(\theta(S_t) - \ell_t(S_t)\big) \le \sqrt{\frac{T}{2}\ln\left(\frac{1}{\delta'}\right)} \tag{8}$$

holds with probability at least $1-\delta'$. Now, define

$$a^*_\ell \in \operatorname*{arg\,min}_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell_t(s_t(a)) \qquad \text{and} \qquad a^*_\theta \in \operatorname*{arg\,min}_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a)) \ .$$

On the one hand, observe that

$$(\triangle) \le \sum_{t=1}^{T}\ell_t(s_t(a^*_\theta)) - \sum_{t=1}^{T}\theta(s_t(a^*_\theta)) \le \sqrt{\frac{T}{2}\ln\left(\frac{1}{\delta'}\right)} \ ,$$

where the last inequality holds with probability at least $1-\delta'$ by Azuma-Hoeffding inequality. On the other hand, we can show that

$$(\triangle) \ge \sum_{t=1}^{T}\ell_t(s_t(a^*_\ell)) - \sum_{t=1}^{T}\theta(s_t(a^*_\ell)) =: (\diamond) \ .$$

However, in this case a^*_ℓ depends on the entire sequence $\ell_1,\dots,\ell_T$. We thus need to use a union bound in order to show that

$$\mathbb{P}\left((\diamond) \le -\sqrt{\frac{T}{2}\ln\left(\frac{K}{\delta'}\right)}\right) \le \sum_{a\in\mathcal{A}}\mathbb{P}\left(\sum_{t=1}^{T}\ell_t(s_t(a)) - \sum_{t=1}^{T}\theta(s_t(a)) \le -\sqrt{\frac{T}{2}\ln\left(\frac{K}{\delta'}\right)}\right) \le \delta' \ ,$$

where the last inequality follows by Azuma-Hoeffding inequality. We conclude the proof by setting $\delta' = \delta/2$. □

Lemma A.2. *The estimates $(\widehat{\theta}_t)_{t=1}^T$ defined in Equation (2) are such that $|\widehat{\theta}_t(s) - \theta(s)| \le \frac{1}{2}\varepsilon_t(s)$ simultaneously holds for all $t\in[T]$ and all $s\in\mathcal{S}$ with probability at least $1-\delta/2$.*

Proof. In a similar way as in Vernade et al. (2020), define $X_m(s)$ to be the empirical mean estimate for $\theta(s)$ which uses the first $m\in[T]$ observed losses corresponding to state $s\in\mathcal{S}$. Notice that $\widehat{\theta}_t(s) = X_{N'_t(s)}(s)$, while we define $\varepsilon'_m(s) = \sqrt{\frac{2}{m}\ln(\frac{4ST}{\delta})}$ so that $\varepsilon_t(s) = \varepsilon'_{N'_t(s)}(s)$. We can additionally observe that $\mathbb{E}[X_m(s)] = \theta(s)$. Then, we can use Azuma-Hoeffding inequality to show that

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{s\in\mathcal{S}}\bigcap_{t\in[T]}\left\{|\widehat{\theta}_t(s) - \theta(s)| \le \frac{1}{2}\varepsilon_t(s)\right\}\right) &\ge \mathbb{P}\left(\bigcap_{s\in\mathcal{S}}\bigcap_{m\in[T]}\left\{|X_m(s) - \theta(s)| \le \frac{1}{2}\varepsilon'_m(s)\right\}\right) \\
&\ge 1 - 2\sum_{s\in\mathcal{S}}\sum_{m=1}^{T} e^{-\frac{1}{2}\varepsilon'_m(s)^2 m} \\
&= 1 - \frac{\delta}{2} \ ,
\end{aligned}$$

where we also used a union bound in the second inequality. □

Lemma A.3. *Consider any algorithm that picks actions $(A_t)_{t\in[T]}$ in the BIO setting with adversarial action-state mappings $(s_t)_{t\in[T]}$ and stochastic loss vectors $(\ell_t)_{t\in[T]}$. Assume that the losses for any fixed state are i.i.d., whereas pairs of losses $\ell_j(s), \ell_{j'}(s')$ of distinct states $s \neq s'$ might be correlated when $j > j'$ and $j - j' \leq d_{j'}$. Then, it holds that $\mathbb{E}[R_T] \leq \mathbb{E}[\mathcal{R}_T]$, where the expectation is with respect to the stochasticity of the losses and the randomness of the algorithm.*

Proof. We know that $\mathbb{E}[\ell_t(s_t(a))] = \theta(s_t(a))$ for any fixed $a \in \mathcal{A}$ and all $t \in [T]$. We further observe that

$$\mathbb{E}[\ell_t(S_t)] = \mathbb{E}\Big[\mathbb{E}[\ell_t(s_t(A_t)) \mid A_t]\Big] = \mathbb{E}[\theta(S_t)]$$

holds for all $t \in [T]$, as A_t is independent of losses that can be correlated with ℓ_t. Now, define

$$a^*_\ell \in \operatorname*{arg\,min}_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell_t(s_t(a)) \qquad \text{and} \qquad a^*_\theta \in \operatorname*{arg\,min}_{a\in\mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a)) \ .$$

Then, we conclude the proof by showing that

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] &= \sum_{t=1}^{T} \mathbb{E}[\ell_t(S_t)] - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t(a^*_\ell))\right] \\ &\geq \sum_{t=1}^{T} \mathbb{E}[\ell_t(S_t)] - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t(a^*_\theta))\right] = \sum_{t=1}^{T} \mathbb{E}[\theta(S_t)] - \sum_{t=1}^{T} \theta(s_t(a^*_\theta)) = \mathbb{E}[R_T] \ .\end{aligned}$$

□

B. High-Probability Regret Bound

B.1. Total delay bound

Lemma 4.1 (Total actual delay)**.** *If* `MetaBIO` *is run with any algorithm $\mathcal{B}$ on delays $(d_t)_{t\in[T]}$, then $\widetilde{\mathcal{D}}_T \leq \mathcal{D}_\Phi$.*

Proof of Lemma 4.1. For any $s \in \mathcal{S}$, we define $\mathcal{T}_s = \{t \in [T] : S_t = s\}$ to be the set of all rounds when the state observed by the learner corresponds to s. Denote by t_s the last time step $t \in \mathcal{T}_s$ such that $N_t(s) < \sigma_t$ and let $\mathcal{C}_s = \{t \in \mathcal{T}_s : t \leq t_s\}$ be those rounds in $\mathcal{T}_s$ that come no later than t_s. According to the choice of t_s, all the rounds in $\mathcal{T}_s$ for which learner waits for the respective delayed loss, must belong to $\mathcal{C}_s$, while the learner incurs $\widetilde{d}_t = 0$ delay for rounds $t \in \mathcal{T}_s \setminus \mathcal{C}_s$. Now we partition $\mathcal{C}_s$ into two sets: the observed set $\mathcal{C}_s^{\text{obs}} = \{t \in \mathcal{C}_s : t + d_t \leq t_s\}$ and the outstanding set $\mathcal{C}_s^{\text{out}} = \{t \in \mathcal{C}_s : t + d_t > t_s\}$. From the choice of t_s, we can see that the number of rounds in $\mathcal{C}_s^{\text{obs}}$ is

$$|\mathcal{C}_s^{\text{obs}}| \leq N_{t_s}(s) < \sigma_{t_s} \leq \sigma_{\max} \ ,$$

and the number of rounds in C_s^{out} is

$$|\mathcal{C}_s^{\text{out}}| \leq \sigma_{t_s} \leq \sigma_{\max} \ .$$

Therefore, we have $|\mathcal{C}_s| \leq 2\sigma_{\max}$. So if we define $\mathcal{C}_{\text{all}} = \bigcup_{s\in\mathcal{S}} \mathcal{C}_s$, then $|\mathcal{C}_{\text{all}}| \leq \min\{2S\sigma_{\max}, T\} = |\Phi|$. This also implies that

$$\sum_{t=1}^{T} \widetilde{d}_t \leq \sum_{t\in\mathcal{C}_{\text{all}}} d_t \leq \sum_{t\in\Phi} d_t$$

by definition of Φ. □

B.2. Improved Regret for DAda-Exp3 for Fixed δ

We follow the analysis of Theorem 4.1 in Gyorgy & Joulani (2021, Appendix A) and our goal is to use the knowledge of $\delta \in (0,1)$ to tune the learning rates $(\eta_t)_{t\in[T]}$ and the implicit exploration terms $(\gamma_t)_{t\in[T]}$, accordingly. Let $d_1, \ldots, d_T$ be the sequence of delays perceived by DAda-Exp3, and let $D_T = \sum_{t=1}^T d_t$ be its total delay. Furthermore, let σ_t be the number of outstanding observations of DAda-Exp3 at the beginning of round $t \in [T]$. Suppose that we take $\gamma_t = c\eta_t$ with $c > 0$ for all $t \in [T]$, then following the same analysis as in Gyorgy & Joulani (2021, Appendix A), we end up with the following regret bound that holds with probability at least $1 - 2\delta'$ for any $\delta' \in (0, 1/2)$:

$$\begin{aligned}\mathcal{R}_T &\leq \frac{\ln(K)}{\eta_T} + \sum_{t=1}^T \eta_t(\sigma_t + (c+1)K) + \frac{\ln(K/\delta')}{2c\eta_T} + \frac{\sigma_{\max} + c + 1}{2c}\ln(1/\delta') \\ &= \frac{1}{\eta_T}\left(\ln(K) + \frac{\ln(K/\delta')}{2c}\right) + \sum_{t=1}^T \eta_t(\sigma_{t-1} + (c+1)K) + \frac{\sigma_{\max}+1}{2c}\ln(1/\delta') + \frac{\ln(1/\delta')}{2} \quad .\end{aligned}$$

Therefore, by taking $\eta_t^{-1} = \sqrt{\frac{(c+1)Kt + \sum_{j=1}^t \sigma_j}{2\ln(K) + \frac{1}{c}\ln(K/\delta')}}$, we get the following bound with probability at least $1 - 2\delta'$:

$$\mathcal{R}_T \leq 2\sqrt{\left((c+1)KT + \sum_{t=1}^T \sigma_t\right)\left(2\ln(K) + \frac{\ln(K/\delta')}{c}\right)} + \frac{\sigma_{\max}+1}{2c}\ln(1/\delta') + \frac{\ln(1/\delta')}{2} \quad .$$

We know that $\sum_{t=1}^T \sigma_t = D_T$ by definition of σ_t. Then, we can set $c = 1$ to obtain that the regret $\mathcal{R}_T$ (as per the original notion of regret used in Gyorgy & Joulani (2021)) is

$$\mathcal{R}_T \leq 2\sqrt{2KT(3\ln(K) + \ln(1/\delta'))} + 2\sqrt{D_T(3\ln(K) + \ln(1/\delta'))} + \frac{\sigma_{\max}+2}{2}\ln(1/\delta') \tag{9}$$

with probability at least $1 - 2\delta'$.

From Lemma A.1, we have that

$$R_T \leq \mathcal{R}_T + \sqrt{2T\ln(2/\delta')} \tag{10}$$

holds with probability at least $1 - \delta'$. So, combining Equations (9) and (10), and setting $\delta = 3\delta'$, we can upper bound our notion of regret R_T as

$$R_T \leq 2\sqrt{2KT(3\ln(K) + \ln(3/\delta))} + \sqrt{2T\ln(6/\delta)} + 2\sqrt{D_T(3\ln(K) + \ln(3/\delta))} + \frac{\sigma_{\max}+2}{2}\ln(3/\delta) \tag{11}$$

with probability at least $1 - \delta$.

B.3. Reduction to DAda-Exp3 via MetaBIO

Based on the reduction via MetaBIO, we require that $\mathcal{B}$ guarantee a regret bound

$$\widehat{\mathcal{R}}_T^{\mathcal{B}} = \sum_{t=1}^T \widetilde{\theta}_t(S_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^T \widetilde{\theta}_t(s_t(a)) \tag{12}$$

that holds with high probability when the losses experienced by $\mathcal{B}$ are of the form $\widetilde{\theta}_t\big(s_t(a)\big)$. Note that, even though the action-state mappings $s_1, \ldots, s_T$ are unknown to the learner, we can provide those losses as long as $\mathcal{B}$ requires bandit feedback only. Indeed, we can compute $\widetilde{\theta}_t(S_t)$ defined in Equations (1) and (2), while we cannot determine $s_t(a)$ for all actions $a \in \mathcal{A}$ that are not A_t. As mentioned in Section 4, in this work we consider DAda-Exp3 (Gyorgy & Joulani, 2021) as algorithm $\mathcal{B}$ used by MetaBIO. In what follows, we refer to this specific choice for the algorithm $\mathcal{B}$.

The analysis of DAda-Exp3 for the high-probability bound (Theorem 4.2) is such that most steps only require that the loss of each action is bounded in $[0, 1]$. Then, those steps apply for any such sequence of loss vectors. However, the crucial part of that analysis that requires attention is the application of Lemma 1 from Neu (2015). We restate it below for reference.

Before that, we introduce the notation required for stating the result. We consider a learner choosing actions $A_1, \dots, A_T$ according to probability distributions $p_1, \dots, p_T$ over actions. We denote by $\mathcal{F}_{t-1}$ the observation history of the learner until the beginning of round t. The result uses importance-weighted estimates for the losses $\ell_1, \dots, \ell_T$ with implicit exploration, where the implicit exploration parameter is $\gamma_t \geq 0$ for each time t. These loss estimates are defined as

$$\widetilde{\ell}_t(a) = \frac{\mathbb{1}[A_t = a]}{p_t(a) + \gamma_t} \ell_t(a) \qquad \forall t \in [T], \forall a \in \mathcal{A} \ . \tag{13}$$

Lemma B.1 (Neu (2015, Lemma 1))**.** *Let γ_t and $\alpha_t(a)$ be nonnegative $\mathcal{F}_{t-1}$-measurable random variables such that $\alpha_t(a) \leq 2\gamma_t$, for all $t \in [T]$ and all $a \in \mathcal{A}$. Let $\widetilde{\ell}_t(a)$ be as in (13). Then,*

$$\sum_{t=1}^{T} \sum_{a=1}^{K} \alpha_t(a) \big(\widetilde{\ell}_t(a) - \ell_t(a) \big) \leq \ln(1/\delta)$$

holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

In our case, we require an analogous result that work when loss vectors correspond with our estimates $\widetilde{\theta}_1, \dots, \widetilde{\theta}_T$. However, these estimate have a dependency with the past actions chosen by the learner. This requires some nontrivial changes in the proof of Neu (2015, Lemma 1).

Before that, we introduce some crucial definitions for this proof. Let $\rho(t) = t + d_t$ be the arrival time for the realized loss $\ell_t(S_t)$ of the state S_t observed at time $t \in [T]$. Let $\widetilde{\rho}(t) = t + \widetilde{d}_t$ be instead the arrival time perceived by algorithm $\mathcal{B}$ relative to its choice of A_t at time t, i.e., when $\mathcal{B}$ receives $\widetilde{\theta}_t(S_t)$. This also means that $\widetilde{\theta}_t(S_t)$ is only defined at time $\widetilde{\rho}(t) \leq \rho(t)$.

Let $\pi \colon [T] \to [T]$ be the permutation of $[T]$ that orders rounds according to their value of $\widetilde{\rho}$. In other words, π satisfies the following property:

$$\pi(r) < \pi(t) \quad \iff \quad \widetilde{\rho}(r) < \widetilde{\rho}(t) \vee (\widetilde{\rho}(r) = \widetilde{\rho}(t) \wedge r < t) \qquad \forall r, t \in [T] \ . \tag{14}$$

This permutation allows us to sort rounds according to the order in which `MetaBIO` feeds $\mathcal{B}$ with a respective estimate for the mean loss. In particular, the r-th round in this order corresponds with the round $t_r = \pi^{-1}(r)$, for any $r \in [T]$. Hence, we can equivalently define the round t_r as the round such that its estimate $\widetilde{\theta}_{t_r}(S_{t_r})$ for the mean loss $\theta(S_{t_r})$ is the r-th estimate received by $\mathcal{B}$.

Define

$$\mathcal{F}_r = \{(j, A_j, S_j, \ell_j(S_j)) \mid j \in [T], \pi(j) \leq r\} \qquad \forall r \in [T] \tag{15}$$

as the information observed by $\mathcal{B}$ by the end to the time step when we feed it the estimate relative to round t_r. Note that this defines a filtration, as $\mathcal{F}_{r-1} \subseteq \mathcal{F}_r$ for all $r \in [T]$, which has some desirable properties thanks to the ordering π we consider. In particular, we have that $\widetilde{d}_{t_r}, \varepsilon_{t_r}, p_{t_r}, N'_{t_r}$ are $\mathcal{F}_{r-1}$-measurable random variables by the way we define them. This property is also due to the fact that N_{t_r} and $\mathcal{L}'_{t_r}$ are determined when conditioning on $\mathcal{F}_{r-1}$. Moreover, we are now interested in the following importance-weighted loss estimates with implicit exploration:

$$\widetilde{\ell}_t(a) = \frac{\mathbb{1}[A_t = a]}{p_t(a) + \gamma_t} \widetilde{\theta}_t(s_t(a)) \qquad \forall t \in [T], \forall a \in \mathcal{A} \ . \tag{16}$$

Corollary B.2. *Let γ_{t_r} and $\alpha_{t_r}(a)$ be non-negative $\mathcal{F}_{r-1}$-measurable random variables such that $\alpha_{t_r}(a) \leq 2\gamma_{t_r}$, for all $r \in [T]$ and all $a \in \mathcal{A}$. Let $\widetilde{\ell}_t(a)$ be as in (16). Then,*

$$\sum_{t=1}^{T} \sum_{a=1}^{K} \alpha_t(a) \big(\widetilde{\ell}_t(a) - \widetilde{\theta}_t(s_t(a)) \big) \leq \ln(1/\delta)$$

holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Proof. We follow the proof of Neu (2015, Lemma 1) by considering any realization $\ell_1, \dots, \ell_T$ of the losses. The main difference is that, when defining the supermartingale as in the original proof, we need to consider the terms of the sum in the

order denoted by π instead of the increasing order of t. For this reason, we rewrite the sum from the statement by following the order given by π:

$$\sum_{r=1}^{T}\sum_{a=1}^{K}\alpha_{t_r}(a)\big(\widetilde{\ell}_{t_r}(a)-\widetilde{\theta}_{t_r}(s_{t_r}(a))\big) \ .$$

At this point, we need prove that $\mathbb{E}\big[\widetilde{\ell}_{t_r}(a)\,\big|\,\mathcal{F}_{r-1}\big]\le\widetilde{\theta}_{t_r}(s_{t_r}(a))$, where we recall that $t_r=\pi^{-1}(r)$. Also recall that ε_{t_r}, p_{t_r} and γ_{t_r} are $\mathcal{F}_{r-1}$-measurable. This property allows us to prove the inequality with the conditional expectation of $\widehat{\theta}_t$ instead of the one with the actual optimistic estimates $\widetilde{\theta}_t$, by the definition of the latter. In other words, we now need to prove that $\mathbb{E}\big[\widehat{\ell}_{t_r}(a)\,\big|\,\mathcal{F}_{r-1}\big]\le\widehat{\theta}_{t_r}(s_{t_r}(a))$, where $\widehat{\ell}_t(a)=\frac{\mathbb{1}[A_t=a]}{p_t(a)+\gamma_t}\widehat{\theta}_t(s_t(a))$.

We can consider two cases depending on whether $\widetilde{d}_{t_r}<d_{t_r}$ is true or not (and, thus, we are in the case $\widetilde{d}_{t_r}=d_{t_r}$). In the first case, note that the realized losses used for computing $\widehat{\theta}_{t_r}(s_{t_r}(a))$ correspond to time steps in $\mathcal{L}'_{t_r}(s_{t_r}(a))$, for which there is a corresponding tuple in $\mathcal{F}_{r-1}$. Therefore, we have that $\widehat{\theta}_{t_r}(s_{t_r}(a))$ is $\mathcal{F}_{r-1}$-measurable, and we can show that

$$\mathbb{E}\Big[\widehat{\ell}_{t_r}(a)\mathbb{1}[\widetilde{d}_{t_r}<d_{t_r}]\,\Big|\,\mathcal{F}_{r-1}\Big]=\mathbb{E}\Big[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\,\Big|\,\mathcal{F}_{r-1}\Big]\frac{\mathbb{1}[\widetilde{d}_{t_r}<d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))}\ell_j(s_{t_r}(a)) \ .$$

In the second case, we have that $\widetilde{d}_{t_r}=d_{t_r}$, which implies that $t_r\in\mathcal{L}'_{t_r}(s_{t_r}(a))$ in the case $A_{t_r}=a$. This means that we have a corresponding tuple in $\mathcal{F}_{r-1}$ only for rounds in $\mathcal{L}'_{t_r}(s_{t_r}(a))\setminus\{t_r\}$. Nonetheless, this does not pose an issue since we have the indicator $\mathbb{1}[A_{t_r}=a]$, and thus $S_{t_r}=s_t(a)$. Indeed, we have that

$$\begin{aligned}
\mathbb{E}\Big[\widehat{\ell}_{t_r}(a)\mathbb{1}[\widetilde{d}_{t_r}=d_{t_r}]\,\Big|\,\mathcal{F}_{r-1}\Big]&=\mathbb{E}\Bigg[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\cdot\frac{\mathbb{1}[\widetilde{d}_{t_r}=d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))}\ell_j(s_{t_r}(a))\,\Bigg|\,\mathcal{F}_{r-1}\Bigg]\\
&=\mathbb{E}\Big[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\,\Big|\,\mathcal{F}_{r-1}\Big]\frac{\mathbb{1}[\widetilde{d}_{t_r}=d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{\substack{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))\\ j\neq t_r}}\ell_j(s_{t_r}(a))\\
&\quad+\mathbb{E}\Big[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\,\Big|\,\mathcal{F}_{r-1}\Big]\frac{\mathbb{1}[\widetilde{d}_{t_r}=d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\ell_{t_r}(s_{t_r}(a))\\
&=\mathbb{E}\Big[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\,\Big|\,\mathcal{F}_{r-1}\Big]\frac{\mathbb{1}[\widetilde{d}_{t_r}=d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))}\ell_j(s_{t_r}(a))
\end{aligned}$$

and therefore the inequality

$$\mathbb{E}\Big[\widehat{\ell}_{t_r}(a)\,\Big|\,\mathcal{F}_{r-1}\Big]=\mathbb{E}\Big[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\,\Big|\,\mathcal{F}_{r-1}\Big]\widehat{\theta}_{t_r}(s_{t_r}(a))\le\widehat{\theta}_{t_r}(s_{t_r}(a))$$

is true because $\mathbb{1}[\widetilde{d}_t<d_t]+\mathbb{1}[\widetilde{d}_t=d_t]=1$ for all $t\in[T]$, and by definition of $\widehat{\theta}_t$.

As already mentioned, this is equivalent to proving that $\mathbb{E}\big[\widetilde{\ell}_{t_r}(a)\,\big|\,\mathcal{F}_{r-1}\big]\le\widetilde{\theta}_{t_r}(s_{t_r}(a))$ holds. By using a notation similar to the original proof, if we define $\widetilde{\lambda}_r=\sum_{a=1}^{K}\alpha_{t_r}(a)\widetilde{\ell}_{t_r}(a)$ and $\lambda_r=\sum_{a=1}^{K}\alpha_{t_r}(a)\widetilde{\theta}_{t_r}(s_{t_r}(a))$, the process $(Z_r)_{r\in[T]}$ with $Z_r=\exp(\sum_{j=1}^{r}(\widetilde{\lambda}_j-\lambda_j))$ is a supermartingale with respect to $(\mathcal{F}_r)_{r\in[T]}$ which has the same properties as in the proof of Neu (2015, Lemma 1). This concludes the current proof by following a similar reasoning as in the original one. □

Thanks to this result, we can conclude that the adoption of `DAda-Exp3` for the reduction via `MetaBIO` can guarantee a high-probability regret bound on $\widehat{\mathcal{R}}_T^{\mathcal{B}}$ as stated in Theorem 4.2, but with total delay $\widetilde{\mathcal{D}}_T=\sum_{t=1}^{T}\widetilde{d}_t$ instead of $\mathcal{D}_T$.

B.4. Regret of `MetaBIO`

By Lemma A.2, we have that

$$R_T\le\sum_{t=1}^{T}\widetilde{\theta}_t(S_t)-\min_{a\in\mathcal{A}}\sum_{t=1}^{T}\widetilde{\theta}_t(s_t(a))+\sum_{t=1}^{T}\varepsilon_t(S_t)=\widehat{\mathcal{R}}_T^{\mathcal{B}}+\sum_{t=1}^{T}\varepsilon_t(S_t) \tag{17}$$

with probability at least $1-\delta/2$, where $\widehat{R}_T^{\mathcal{B}}$ (Equation (12)) is the regret of algorithm $\mathcal{B}$ when fed with $(\widetilde{\theta}_t \circ s_t)_{t\in[T]}$ as losses.

Lemma B.3. *Conditioning on the event as stated in Lemma A.2, the sum of errors suffered from* `MetaBIO` *by using the loss estimates* $(\widetilde{\theta}_t)_{t\in[T]}$ *from Equations* (1) *and* (2) *is*

$$\sum_{t=1}^{T} \varepsilon_t(S_t) \leq \left(4+2\sqrt{2}\right)\sqrt{ST\ln\left(\frac{4ST}{\delta}\right)} \ .$$

Proof. First, observe that we can rewrite the sum of errors as

$$\sum_{t=1}^{T} \varepsilon_t(S_t) = \sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{1}[\widetilde{d}_t < d_t] + \sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{1}[\widetilde{d}_t = d_t] \ .$$

We now provide an upper bound for the first sum of errors. For any $s \in \mathcal{S}$, we define $\mathcal{T}_s = \{t \in [T] : S_t = s\}$ to be the set of all rounds when the state observed by the learner corresponds to s. We can bound it as

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{1}[\widetilde{d}_t < d_t] &= \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s} \varepsilon_t(s)\mathbb{1}[\widetilde{d}_t < d_t] \\
&= \sqrt{2\ln\left(\frac{4ST}{\delta}\right)}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s}\sqrt{\frac{1}{N'_t(s)}}\mathbb{1}[\widetilde{d}_t < d_t] \\
&\leq 2\sqrt{\ln\left(\frac{4ST}{\delta}\right)}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s}\sqrt{\frac{1}{M_t(s)}}\mathbb{1}[\widetilde{d}_t < d_t] && \text{(because } N'_t(s) \geq \tfrac{1}{2}M_t(s)\text{)} \\
&\leq 4\sqrt{\ln\left(\frac{4ST}{\delta}\right)}\sum_{s\in\mathcal{S}}\sqrt{M_T(s)} && \text{(since } M_t(s) \text{ is increasing over } \mathcal{T}_s\text{)} \\
&\leq 4\sqrt{ST\ln\left(\frac{4ST}{\delta}\right)} \ ,
\end{aligned}$$

where the second inequality holds because $N'_t(S_t) = N_t(S_t) \geq \frac{1}{2}M_t(S_t)$ when $\widetilde{d}_t < d_t$ since $M_t(S_t) \leq N_t(S_t) + \sigma_t$, while the last one follows by Jensen's inequality and the fact that $\sum_{s\in\mathcal{S}} M_T(s) = T$.

As a last step, we provide an upper bound to the second sum. Let $J_s = \{r \in \mathcal{T}_s : \widetilde{d}_r = d_r\}$ and notice that $|J_s| \leq |\mathcal{T}_s| = M_T(s)$. Observe that $\rho(t) = \widetilde{\rho}(t)$ for each round t such that $\widetilde{d}_t = d_t$, and thus by Equation (14) we have that

$$\pi(r) < \pi(t) \iff \rho(r) < \rho(t) \vee (\rho(r) = \rho(t) \wedge r < t)$$

for all $r, t \in [T]$ such that $\widetilde{d}_r = d_r$ and $\widetilde{d}_t = d_t$. Define $\nu_s : J_s \to \left[|J_s|\right]$ by

$$\nu_s(t) = |\{r \in J_s \ : \ \pi(r) \leq \pi(t)\}| \qquad \forall t \in J_s \ .$$

Observe that $\nu_s(t) \leq N'_t(s) = |\mathcal{L}'_t(s)|$ for all $s \in \mathcal{S}$ and all $t \in J_s$. This is due to the fact that $\nu_s(t)$ counts a subset of $\mathcal{L}'_t(s)$; to be precise, we have that $\nu_s(t) = |\mathcal{L}'_t(s) \cap J_s|$. Moreover, notice that the condition $\pi(r) \leq \pi(t)$ defines a total order over J_s. Hence, $\nu_s(t)$ counts the number of elements of J_s preceding $t \in J_s$ (including t itself) in this total order.

This implies that ν_s is a bijection between J_s and $\big[|J_s|\big]$. Then, using a similar reasoning as before, we show that

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{1}[\widetilde{d}_t = d_t] &= \sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}_s} \sqrt{\frac{1}{N'_t(s)}}\mathbb{1}[\widetilde{d}_t = d_t] \\
&= \sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in J_s} \sqrt{\frac{1}{N'_t(s)}} && \text{(by definition of } J_s\text{)} \\
&\le \sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in J_s} \sqrt{\frac{1}{\nu_s(t)}} && \text{(since } \nu_s(t) \le N'_t(s) \text{ for } t \in J_s\text{)} \\
&\le 2\sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sqrt{|J_s|} && \text{(since } \nu_s(t) \text{ is bijective)} \\
&\le 2\sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sqrt{M_T(s)} && \text{(since } |J_s| \le M_T(s)\text{)} \\
&\le 2\sqrt{2ST\ln\left(\frac{4ST}{\delta}\right)} \ . && \text{(by Jensen's inequality)}
\end{aligned}$$

□

Theorem 4.3. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *using* `DAda-Exp3`*, then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$\begin{aligned}
R_T \le\ & 2\sqrt{2KTC_{K,3\delta}} + 7\sqrt{ST\ln\frac{4ST}{\delta}} \\
& + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{4}{\delta} \ .
\end{aligned} \tag{6}$$

Proof of Theorem 4.3. By Equation (17), the regret R_T can be bounded as

$$R_T \le \widehat{\mathcal{R}}_T^{\mathcal{B}} + \sum_{t=1}^{T} \varepsilon_t(S_t) \le \widehat{\mathcal{R}}_T^{\mathcal{B}} + 7\sqrt{ST\ln\frac{4ST}{\delta}}$$

with probability at least $1-\delta/2$, where the last inequality follows by Lemma B.3. From what we argued in Appendix B.3, we can upper bound $\widehat{\mathcal{R}}_T^{\mathcal{B}}$ using the high-probability regret bound of `DAda-Exp3`. Notice that the delays incurred by `DAda-Exp3` via `MetaBIO` are those given when providing the estimates $(\widetilde{\theta}_t)_{t\in[T]}$. We denote these delays by $\widetilde{d}_1, \dots, \widetilde{d}_T$, and the total delay perceived by `DAda-Exp3` is thus $\widetilde{\mathcal{D}}_T = \sum_{t=1}^{T} \widetilde{d}_t$. Hence, from the improved bound for `DAda-Exp3` in Equation (9), we have that

$$\widehat{\mathcal{R}}_T^{\mathcal{B}} \le 2\sqrt{2KT(3\ln(K)+\ln(4/\delta))} + 2\sqrt{\widetilde{\mathcal{D}}_T(3\ln(K)+\ln(4/\delta))} + \frac{\sigma_{\max}+2}{2}\ln(4/\delta)$$

holds with probability at least $1-\delta/2$. The combination of the above two inequalities, together with Lemma 4.1, concludes the proof. □

B.5. Regret of `MetaAdaBIO`

Theorem 4.4. *Let $\delta \in (0,1)$. If we run* `MetaAdaBIO` *with* `DAda-Exp3`*, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$\begin{aligned}
R_T \le\ & 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} \\
& + 6\sqrt{KTC_{K,2\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta} \ .
\end{aligned} \tag{7}$$

Proof of Theorem 4.4. Let $t^* \in [T]$ be the last round before `MetaAdaBIO` switches from `DAda-Exp3` to `MetaBIO`, i.e., the last round that satisfies $\mathfrak{D}_{t^*} C_{K,4\delta} \leq 49ST \ln \frac{8ST}{\delta}$. Then, define $a^* \in \arg\min_a \sum_{t=1}^{T} \theta(s_t(a))$. We may decompose regret as

$$\begin{aligned} R_T &= \sum_{t=1}^{t^*} \Big(\theta(S_t) - \theta(s_t(a^*))\Big) + \sum_{t=t^*+1}^{T} \Big(\theta(S_t) - \theta(s_t(a^*))\Big) \\ &\leq \underbrace{\sum_{t=1}^{t^*} \theta(S_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{t^*} \theta(s_t(a))}_{R_{t^*}} + \underbrace{\sum_{t=t^*+1}^{T} \theta(S_t) - \min_{a\in\mathcal{A}} \sum_{t=t^*+1}^{T} \theta(s_t(a))}_{R_{t^*:T}} \ . \end{aligned}$$

The incurred delay until time t^* is $\mathfrak{D}_{t^*}$. Thus, from Equation (11), we get that the following bound

$$R_{t^*} \leq 2\sqrt{2Kt^* C_{K,2\delta}} + \sqrt{2t^* \ln \frac{12}{\delta}} + 2\sqrt{\mathfrak{D}_{t^*} C_{K,2\delta}} + \frac{\sigma_{\max}+2}{2} \ln \frac{6}{\delta} \tag{18}$$

holds with probability at least $1-\delta/2$, where we recall that $C_{K,\delta} = 3\ln K + \ln(12/\delta)$. If our algorithm never switches, then $t^* = T$ and we get the bound in (18) for R_T. Note that this is no greater than the upper bound in the statement as $\sqrt{\mathfrak{D}_T C_{K,2\delta}} \leq 7\sqrt{ST\ln(8ST/\delta)}$ by definition of t^* in this case.

Otherwise, we use the switching condition $\sqrt{\mathfrak{D}_{t^*} C_{K,2\delta}} \leq 7\sqrt{ST\ln(8ST/\delta)}$ along with the fact that $\sqrt{t^* \ln(12/\delta)} \leq \sqrt{Kt^* C_{K,2\delta}}$ to get

$$R_{t^*} \leq 3\sqrt{2Kt^* C_{K,2\delta}} + 14\sqrt{ST \ln \frac{8ST}{\delta}} + \frac{\sigma_{\max}+2}{2} \ln \frac{6}{\delta} \ . \tag{19}$$

Furthermore, Theorem 4.3 directly gives us an upper bound for $R_{t^*:T}$ since `MetaAdaBIO` runs `MetaBIO` for $t > t^*$ with the confidence parameter set to $\delta/2$. We just need to bound the total incurred delays of these rounds, namely $\widetilde{\mathcal{D}}_{t^*:T}$. Let σ'_t be the outstanding observations for any round $t > t^*$ as perceived by the execution of `MetaBIO` starting after round t^*, that is, when considering only delays $(d_t)_{t>t^*}$. It is immediate to observe that $\sigma'_t \leq \sigma_t$ and thus $\max_{t>t^*} \sigma'_t \leq \max_{t>t^*} \sigma_t$. Moreover, from Lemma 4.1 we have

$$\widetilde{\mathcal{D}}_{t^*:T} \leq \mathcal{D}_{\Phi'} \ ,$$

where Φ' denotes a set of $\min\{T-t^*, 2S\sigma'_{\max}\}$ rounds with the largest delays among $(d_t)_{t>t^*}$, with $\sigma'_{\max} = \max_{t>t^*} \sigma'_t$. So we have

$$\mathcal{D}_{\Phi'} \leq \mathcal{D}_{\Phi}$$

due to the fact that $|\Phi'| = \min\{T-t^*, 2S\sigma'_{\max}\} \leq \min\{T, 2S\sigma_{\max}\} = |\Phi|$. Therefore, from Theorem 4.3 we obtain

$$R_{t^*:T} \leq 2\sqrt{2K(T-t^*)C_{K,3\delta}} + 7\sqrt{ST\ln\frac{8ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{8}{\delta} \tag{20}$$

with probability at least $1-\delta/2$. We conclude the proof by combining Equations (19) and (20) along with the fact that $\sqrt{t^*} + \sqrt{T-t^*} \leq \sqrt{2T}$ to get that the bound

$$R_T \leq 6\sqrt{KTC_{K,2\delta}} + 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta}$$

holds with probability at least $1-\delta$. □

B.6. Expected Regret Analysis of `MetaAdaBIO` with `Tsallis-INF`

Proposition 4.6. *If we execute* `MetaAdaBIO` *with* `Tsallis-INF` *(Zimmert & Seldin, 2020), and use the switching condition* $\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST\ln(2ST)}$ *at each round* $t \in [T]$*, where* $\mathfrak{D}_t = \sum_{j=1}^{t} \sigma_j$*, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$\begin{aligned} \mathbb{E}[R_T] \leq\ & 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi \ln K} \\ & + 2\min\left\{6\sqrt{ST\ln(2ST)}, \sqrt{8\mathcal{D}_T \ln K}\right\} \ . \end{aligned}$$

Proof of Proposition 4.6. We begin by studying of expected regret of `MetaBIO` and we then give a regret analysis of `MetaAdaBIO`. When running `MetaBIO`, we use the unbiased empirical mean estimators $(\widehat{\theta}_t)_{t\in[T]}$ as the mean loss estimates, rather than the lower confidence bounds $(\widetilde{\theta}_t)_{t\in[T]}$. The expected regret is defined as

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} \mathbb{E}[\theta(S_t)] - \sum_{t=1}^{T} \theta(s_t(a^*)) \ ,$$

where $a^* = \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a))$. Here we use a version of `Tsallis-INF` that is tailored for the delayed bandits problem (Zimmert & Seldin, 2020), which guarantees a bound in expectation on the regret

$$\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a) = \sum_{t=1}^{T} \widehat{\theta}_t(S_t) - \sum_{t=1}^{T} \widehat{\theta}_t(s_t(a))$$

against any fixed action $a \in \mathcal{A}$, using the loss estimates $\{\widehat{\theta}_t\}_{t\in[T]}$. Observe that this regret is defined in terms of our estimates, as required in our case. By Zimmert & Seldin (2020, Theorem 1), `Tsallis-INF` guarantees that its expected regret is

$$\mathbb{E}\Big[\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] = \mathbb{E}\left[\sum_{t=1}^{T} \widehat{\theta}_t(S_t) - \sum_{t=1}^{T} \widehat{\theta}_t(s_t(a^*))\right] \leq 4\sqrt{KT} + \sqrt{8\widetilde{\mathcal{D}}_T \ln K} \leq 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} \ ,$$

where the last inequality uses Lemma 4.1. Then, we can focus on our notion of regret and use the above regret bound to obtain that

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\Big[R_T - \widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] + \mathbb{E}\Big[\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\big(\theta(S_t) - \widehat{\theta}_t(S_t)\big)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\big(\widehat{\theta}_t(s_t(a^*)) - \theta(s_t(a^*))\big)\right] + \mathbb{E}\Big[\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] \\
&\leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\big(\theta(S_t) - \widehat{\theta}_t(S_t)\big)\right]}_{\Delta} + \mathbb{E}\left[\sum_{t=1}^{T}\big(\widehat{\theta}_t(s_t(a^*)) - \theta(s_t(a^*))\big)\right] + 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} \ . \qquad (21)
\end{aligned}$$

We know that our mean estimator is unbiased. Therefore, we have that $\mathbb{E}[\widehat{\theta}_t(s_t(a^*))] = \theta(s_t(a^*))$ for any $t \in [T]$, meaning that the second term in the right-hand side of (21) is equal to zero.

On the other hand, we can apply Lemma A.2 to get the following bound for Δ that holds with probability at least $1 - \delta/2$ for any $\delta \in (0, 1)$:

$$\Delta \leq \min\left\{\frac{1}{2}\sum_{t=1}^{T} \varepsilon_t(S_t), T\right\} \ , \qquad (22)$$

where we recall that $\varepsilon_t(s) = \sqrt{\frac{2}{N_t'(s)} \ln \frac{4ST}{\delta}}$. In particular, the inequality $\Delta \leq T$ is true in general. By Lemma B.3, we can bound the right-hand side of (22) as

$$\frac{1}{2}\sum_{t=1}^{T} \varepsilon_t(S_t) \leq \frac{7}{2}\sqrt{ST \ln \frac{4ST}{\delta}}$$

when conditioning on the event as in the statement of Lemma A.2. If we denote such an event as $\mathcal{E}$, we have that $\mathbb{P}\big(\overline{\mathcal{E}}\big) \leq \delta/2$ and that $\mathbb{E}[\Delta \mid \mathcal{E}] \leq \frac{7}{2}\sqrt{ST \ln(4ST/\delta)}$. As a consequence, we notice that

$$\mathbb{E}[\Delta] = \mathbb{E}[\Delta \mid \mathcal{E}]\mathbb{P}(\mathcal{E}) + \mathbb{E}\big[\Delta \mid \overline{\mathcal{E}}\big]\mathbb{P}\big(\overline{\mathcal{E}}\big) \leq \frac{7}{2}\sqrt{ST \ln \frac{4ST}{\delta}} + \frac{\delta}{2}T \leq 5\sqrt{ST \ln(2ST)} + 1$$

where in the last inequality we set $\delta = 2/T$. Since we assume that $S \geq 2$, we can easily observe that $\mathbb{E}[\Delta] \leq 6\sqrt{ST \ln(2ST)}$. Plugging this into Equation (21) gives us

$$\mathbb{E}[R_T] \leq 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 6\sqrt{ST \ln(2ST)} \ . \qquad (23)$$

At this point, we can proceed to the proof of the overall bound on the expected regret of `MetaAdaBIO`. The behaviour of `MetaAdaBIO` follows the same principle as before, but the switching condition is different:

$$\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST\ln(2ST)} \ .$$

Similar to the analysis of `MetaAdaBIO` in Appendix B.5, we decompose the regret into

$$\mathbb{E}[R_T] \leq \underbrace{\sum_{t=1}^{t^*} \mathbb{E}[\theta(S_t)] - \min_{a\in\mathcal{A}} \sum_{t=1}^{t^*} \theta(s_t(a))}_{R_{t^*}} + \underbrace{\sum_{t=t^*+1}^{T} \mathbb{E}[\theta(S_t)] - \min_{a\in\mathcal{A}} \sum_{t=t^*+1}^{T} \theta(s_t(a))}_{R_{t^*:T}} \ ,$$

where t^* is the last round satisfying $\sqrt{8\mathfrak{D}_{t^*}} \leq 6\sqrt{ST\ln(2ST)}$. Then, we have

$$\mathbb{E}[R_{t^*}] \leq 4\sqrt{Kt^*} + \sqrt{8\mathfrak{D}_{t^*}\ln K} \ . \tag{24}$$

If $t^* = T$ then $R_{t^*} = R_T$ and we get the bound in (24), where we note that $\sqrt{8\mathfrak{D}_T \ln K} \leq 6\sqrt{ST\ln(2ST)}$ by definition of t^* in this case, and we can replace $\mathfrak{D}_T$ by $\mathcal{D}_T$. Otherwise, $t^* < T$ and we can apply the bound for `MetaBIO` from (23), along with the fact that the total incurred delay after round t^* is upper bounded by $\mathcal{D}_\Phi$, in order to derive an upper bound for $\mathbb{E}[R_{t^*:T}]$ that is

$$\mathbb{E}[R_{t^*:T}] \leq 4\sqrt{K(T-t^*)} + \sqrt{8\mathcal{D}_\Phi \ln K} + 6\sqrt{ST\ln(2ST)} \ . \tag{25}$$

Finally, if we use the fact that $\sqrt{8\mathfrak{D}_{t^*}} \leq 6\sqrt{ST\ln(2ST)}$ (by definition of t^*) in (24), and combine it with (25), we conclude that

$$\mathbb{E}[R_T] \leq 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 2\min\Big\{6\sqrt{ST\ln(2ST)}, \sqrt{8\mathcal{D}_T \ln K}\Big\} \ ,$$

where we also used the fact that $\sqrt{t^*} + \sqrt{T-t^*} \leq \sqrt{2T}$. □

C. Proofs for the Lower Bounds

Theorem 5.2. *Suppose that the action-state mapping is adversarial and the losses are stochastic and that $d_t = d$ for all $t \in [T]$. If $T \geq \min\{S, d\}$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\sqrt{\min\{S,d\}T}\big)$.*

Proof of Theorem 5.2. Assume without loss of generality that $K = 2$ and let $\mathcal{S} = \{h_1, \dots, h_S\}$ be the finite set of possible states. Let $S' = \lfloor \min\{S/2, d\} \rfloor$ and let $I_1, \dots, I_T$ be the actions chosen by the considered algorithm. Split the T time steps into $m = \lfloor T/S' \rfloor$ blocks $B_1, \dots, B_m$ of equal size S', eventually leaving $\leq S'-1$ extra time steps. We assume with no loss of generality that the last step corresponds to the end of the m-th block. The feedback formed by the losses of the actions chosen by the algorithm in a certain block is received only after the last time step of the same block since $S \leq 2d$. Define $b_i = (i-1)S' + 1$ for all $i \in [m]$. We assume that the learner receives *all* the realized losses $\ell_t(s_t(A))$ for all $t \in B_i$ and all $A \in \{1, 2\}$ at the end of each block, which means that we are in a full information setting, as this only helps the algorithm.

Now, we define a specific sequence of assignments from actions to states, and construct losses so that the expected regret becomes sufficiently large. Let $s_t(A) = h_{2(t-b_i)+A}$ for all $t \in B_i$, all $i \in [m]$ and all $A \in \{1, 2\}$; this means that, for the first time step of any block, actions 1 and 2 will be assigned to states h_1 and h_2 respectively, then to h_3 and h_4 respectively in the next time step of the same block, and so on. Let $\varepsilon = \frac{1}{4}\sqrt{\frac{S'}{2T\ln(4/3)}} \in [0, \frac{1}{4}]$ and let $\theta^{(A)} \in \mathbb{R}^2$ be a vector of mean losses such that $\theta_i^{(A)} = \frac{1}{2} - \mathbb{I}\{i = A\}\varepsilon$, for each $A \in \{1, 2\}$. We simplify the notation with $\mathbb{E}_A[\cdot] = \mathbb{E}\big[\cdot \,\big|\, \theta^{(A)}\big]$ and $\mathbb{P}_A(\cdot) = \mathbb{P}\big(\cdot \,\big|\, \theta^{(A)}\big)$, where the conditioning on $\theta^{(A)}$ means that we sample losses for each state assigned to $i \in \{1, 2\}$ such that they are Bernoulli random variables with mean $\theta_i^{(A)}$. In particular, conditioning on $\theta^{(A)}$, we sample independent Bernoulli random variables $X_1^i, \dots, X_m^i$ with mean $\theta_i^{(A)}$, one for each block, for $i \in \{1, 2\}$. Then, the losses are defined as $\ell_t(s_t(i)) = X_j^i$ for each $t \in B_j$ and each $j \in [m]$.

We can now proceed to show a lower bound for the expected pseudo-regret. Let T_i be the number of times the learner chooses action i over all T time steps. The expected pseudo-regret over the two instances determined by $\theta^{(k)}$ for $k \in \{1, 2\}$ adds up to

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] = \varepsilon(2T - \mathbb{E}_1[T_1] - \mathbb{E}_2[T_2]) \ .$$

Following the standard analysis, we show that the difference $\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2]$ is such that

$$\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2] \leq T \cdot d_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_1) \leq T\sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2)} \ ,$$

where the last step follows by Pinsker's inequality.

Let $\lambda_i = \{(I_t, \ell_t(S_t(1)), \ell_t(S_t(2))) \mid t \in B_i\}$ be the feedback set known to the learner by the end of block B_i, and let $\lambda^i = (\lambda_1, \dots, \lambda_i)$ be the tuple of all feedback sets up to the end of block B_i. Denote by $\mathbb{P}_{k,i}(\cdot)$ the probability measure of feedback tuples λ^i conditioned on $\theta^{(A)}$. By the chain rule for the relative entropy, we can observe that

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2) &= \sum_{i=1}^{m} \sum_{\lambda^{i-1}} \mathbb{P}_1(\lambda^{i-1}) D_{\mathrm{KL}}(\mathbb{P}_{1,i}(\cdot \mid \lambda^{i-1}) \| \, \mathbb{P}_{2,i}(\cdot \mid \lambda^{i-1})) \\ &\leq \sum_{i=1}^{m} \sum_{\lambda^{i-1}} \mathbb{P}_1(\lambda^{i-1}) 16\varepsilon^2 \ln(4/3) \\ &= 16 m \varepsilon^2 \ln(4/3) \ , \end{aligned}$$

where we used the fact that each relative entropy $D_{\mathrm{KL}}(\mathbb{P}_{1,i}(\cdot \mid \lambda^{i-1}) \,\|\, \mathbb{P}_{2,i}(\cdot \mid \lambda^{i-1}))$ corresponds to the sum of the relative entropy between two Bernoulli distributions with means $1/2$ and $1/2 - \varepsilon$ and that between Bernoulli distributions with means $1/2 - \varepsilon$ and $1/2$, respectively, which is upper bounded by $16\varepsilon^2 \ln(4/3)$ for $\varepsilon \in [0, 1/4]$. This follows by an application of the chain rule for the relative entropy, as well as from the fact that the distribution of I_t is the same under both $\mathbb{P}_{1,i}(\cdot \mid \lambda^{i-1})$ and $\mathbb{P}_{2,i}(\cdot \mid \lambda^{i-1})$, for all $t \in B_i$ and any λ^{i-1}. Therefore, we have that

$$\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2] \leq 2\varepsilon T \sqrt{2m \ln(4/3)}$$

which also implies that

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] \geq \varepsilon T \left(1 - 2\varepsilon \sqrt{2 \frac{T}{S'} \ln(4/3)}\right) = \frac{\varepsilon T}{2} \geq \frac{1}{8} \sqrt{\frac{\lfloor S/2 \rfloor T}{2 \ln(4/3)}} \geq \frac{1}{8} \sqrt{\frac{ST}{6 \ln(4/3)}} \ ,$$

where we used the facts that $m \leq T/S'$ and that $\lfloor S/2 \rfloor \geq S/3$ for any integer $S \geq 2$. This means that the expected pseudo-regret of the learner has to be $\frac{1}{16}\sqrt{\frac{ST}{6\ln(4/3)}}$ at least in one of the two instances. Now, for $S > 2d$ we use the same construction, but now we only use $2d$ states, which leads to the promised $\Omega(\sqrt{\min\{S, d\}T})$ lower bound. □

Theorem 5.3. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq d + 1$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret*

$$\mathbb{E}[R_T] = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big) \ .$$

Proof of Theorem 5.3. Let $S' = \min\{\lfloor \frac{S}{2} \rfloor, \lfloor \frac{T}{d+1} \rfloor\} \geq 1$. We consider the first $(d+1)S'$ rounds of the game and divide them into S' blocks $B_1, \dots, B_{S'}$ of same length $d + 1$. In this way, we ensure that the feedback for any time step in some block is revealed to the learner only after its final round.

Without loss of generality, we can assume that the learner observes all the losses of one block immediately after its last time step; this only helps the learner since they would observe only the incurred losses at possibly later rounds otherwise. We can further simplify the problem by assuming that losses are deterministic functions of the states, i.e., $\ell_t \equiv \theta$ for every round t. This also means that the problem turns into an easier, full-information version of our problem with deterministic losses. Now, let the adversary choose the action-state mappings such that for each block index i and each action $a \in \mathcal{A}$, $S_t(a) = S_{t'}(a) \in \{s_{2i-1}, s_{2i}\}$ for all $t, t' \in B_i$. Furthermore, we assume that the losses are chosen such that $\theta(s_{2i-1}) \in \{0, 1\}$ and $\theta(s_{2i}) = 1 - \theta(s_{2i-1})$ for all $i \in [S']$. In this construction, the learner cannot obtain any useful information from the states of a block because of the delays. Moreover, the states observed in one block are not observed again in the other blocks.

It thus suffices to prove a lower bound for a standard full information game with S' rounds and loss range $[0, d+1]$. Hence, we can conclude that the expected regret of any algorithm has to be

$$\mathbb{E}[R_T] = \Omega\Big((d+1)\sqrt{S'}\Big) = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big) \ .$$

□

D. Action-State Mappings and Loss Means Used in the Experiments

Table 1 and Table 2 describe the instances used to generate the data for the experiments of Section 6.

Mean loss	$s=1$	$s=2$	$s=3$
$\theta(s)$	0.2	0.4	0.8
Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	0.8	0.1	0.1
$a=2$	0.4	0.5	0.1
$a=3$	0.3	0.7	0.0
$a=4$	0.5	0.3	0.2

Table 1. Mean losses and stochastic action-state mapping for Experiment 1 in Section 6.

Mean loss	$s=1$	$s=2$	$s=3$
$\theta(s)$	0	1	1
Environment 1			
Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	0.06	0.47	0.47
$a=2$	0	0.50	0.50
$a=3$	0	0.50	0.50
$a=4$	0	0.50	0.50
Environment 2			
Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	1	0	0
$a=2$	0.94	0.03	0.03
$a=3$	0.94	0.03	0.03
$a=4$	0.94	0.03	0.03

Table 2. Mean losses and stochastic action-state mappings for Experiment 2 in Section 6.