Banker Online Mirror Descent: A Universal Approach for Delayed Online Bandit Learning

Jiatai Huang [*] [1] **Yan Dai** [*] [1] **Longbo Huang** [1]

Abstract

We propose Banker Online Mirror Descent (`Banker-OMD`), a novel framework generalizing the classical Online Mirror Descent (OMD) technique in the online learning literature. The `Banker-OMD` framework almost completely decouples feedback delay handling and the task-specific OMD algorithm design, thus facilitating the design of new algorithms capable of efficiently and robustly handling feedback delays. Specifically, it offers a general methodology for achieving $\widetilde{\mathcal{O}}(\sqrt{T}+\sqrt{D})$-style regret bounds in online bandit learning tasks with delayed feedback, where T is the number of rounds and D is the total feedback delay. We demonstrate the power of `Banker-OMD` by applications to two important bandit learning scenarios with delayed feedback, including delayed scale-free adversarial Multi-Armed Bandits (MAB) and delayed adversarial linear bandits. `Banker-OMD` leads to the first delayed scale-free adversarial MAB algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{K}L(\sqrt{T}+\sqrt{D}))$ regret and the first delayed adversarial linear bandit algorithm achieving $\widetilde{\mathcal{O}}(\text{poly}(n)(\sqrt{T}+\sqrt{D}))$ regret. As a corollary, the first application also implies $\widetilde{\mathcal{O}}(\sqrt{KT}L)$ regret for non-delayed scale-free adversarial MABs, which is the first to match the $\Omega(\sqrt{KT}L)$ lower bound up to logarithmic factors and can be of independent interest.

1. Introduction

Multi-armed bandit (MAB) is a classical online learning problem with partial information feedback. In the MAB problem, an agent is given a set of arms and needs to choose one each time, after which the agent suffers a loss determined by the environment. The agent's objective is to minimize the expected difference between his/her total loss and the total loss of a fixed best arm, which is called the regret.

Among the many techniques successfully applied to MAB algorithm design, Online Mirror Descent (OMD) (Warmuth and Jagota, 1997) and Follow-The-Regularized-Leader (FTRL) (Gordon, 1999) have been proven to be powerful tools, especially in adversarial MAB settings — where at each round t, an adversary arbitrarily picks a loss for each arm when the agent is making a decision. The OMD/FTRL perspective offers an alternative algorithmic form of classical adversarial MAB algorithms originating from the idea of exponential weighting, particularly the well-known EXP3 algorithm (Auer et al., 2002b). It also leads to the development of new MAB algorithms such as Tsallis-INF (Zimmert and Seldin, 2019) and BROAD-OMD (Wei and Luo, 2018) by using different regularizers. At last, the OMD/FTRL perspective also generalizes to other online learning tasks, e.g., adversarial linear bandits (Abernethy et al., 2008; Audibert et al., 2014) or Markov Decision Processes (Jin et al., 2020).

However, in many important practical learning problems, action feedback may not arrive immediately after execution. For instance, in the web advertisement scenario (Li et al., 2010), after the server selects a set of ads and renders the page for a user, the signal about user reactions may arrive after the server picks ads for other incoming users. Similar situations can also happen in parallel computing (Chen and Xu, 2019) where jobs need to be allocated to some work node before previous jobs finish or medical experiments (Wason and Trippa, 2014) where patients need to get treated before the information of previous treatment is gathered.

In such cases where feedback is arbitrarily delayed, the classical OMD approach can no longer be directly applied to guarantee similar performance. Technically, this is because the classical OMD/FTRL framework heavily relies on techniques to rearrange potential terms into a telescoping sum so that terms from *adjacent* rounds cancel. Thus, delayed feedback makes this real-time cancellation impossible. While it is still possible to design algorithms for delayed MABs (Zimmert and Seldin, 2020; Gyorgy and Joulani, 2021) or delayed MDPs (Jin et al., 2022; Dai et al., 2022) using OMD or similar ideas, one naturally asks the following question:

[*]Equal contribution [1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. Correspondence to: Longbo Huang <longbohuang@mail.tsinghua.edu.cn>.

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

Is there a general framework for transforming existing task-specific OMD algorithms into ones that can handle the "delayed version" of the corresponding tasks?

In this paper, we provide a positive answer to this question by presenting a novel OMD framework, `Banker-OMD`.

1.1. Our Contribution

We first present `Banker-OMD`, a novel framework generalizing the classical OMD framework to handle feedback delays efficiently. Unlike classical OMD that plans a new action based only on the potential term of the previous *single* round, `Banker-OMD` proposes a new perspective to take *all* previous rounds into account, providing a robust rule to plan new actions when action feedback is delayed. The `Banker-OMD` framework achieves $\widetilde{\mathcal{O}}((\sqrt{T}+\sqrt{D})f(T))$ regret for various delayed bandit learning tasks, where T is the number of rounds, D is the total delay, and $f(T)$ is a task-specific factor irrelevant to delays. More importantly, our framework keeps the flexibility of the classical OMD framework — it depends on neither the choice of the regularizer nor the loss estimators and can be equipped with various OMD techniques (e.g., doubling tricks).

To demonstrate the power of `Banker-OMD`, we apply it to two important bandit learning scenarios with delayed feedback. The first one is delayed scale-free adversarial MAB (Section 6): in addition to the usual delays (i.e., the feedback of t-th round can suffer a delay of d_t), the losses can fall into a general *unknown* range $[-L, L]$ instead of the restrictive range $[0, 1]$ (Zimmert and Seldin, 2020; Thune et al., 2019). As we will explain, such a generalization is very important in real-world scenarios such as advertisement recommendations. We design algorithms that can achieve $\widetilde{\mathcal{O}}(\sqrt{K}L(\sqrt{T}+\sqrt{D}))$ regret. Additionally, a small-loss style bound (Neu, 2015) is also provided for better data adaptivity. Consequently, when applying our algorithm to the non-delayed scale-free adversarial MAB problem, our bound dominates the SOTA $\widetilde{\mathcal{O}}(\sqrt{KL_2}+L_\infty\sqrt{KT})=\widetilde{\mathcal{O}}(K\sqrt{T}L)$ (Putta and Agrawal, 2022) and matches the $\Omega(\sqrt{KT}L)$ lower bound (Auer et al., 2002b) up to logarithmic factors, which can be of independent interest (see Remark 6.4 for more discussions).

The second one is delayed adversarial linear bandits (Section 7), where the feedback can be *arbitrarily* delayed and the losses are adversarial, significantly strengthening the uniformly delayed setting (Ito et al., 2020) or stochastic linear bandit setting (Zhou et al., 2019; Vernade et al., 2020). In this setting, an `Banker-OMD`-based $\widetilde{\mathcal{O}}(\text{poly}(n)(\sqrt{T}+\sqrt{D}))$-regret algorithm is yielded.

To our knowledge, we are the first to handle feedback delays in adversarial scale-free MABs and the first to allow *arbitrary* (i.e., unrestricted and also unknown) delays in adversarial linear bandits, respectively. For a better comparison, we also present an overview of our algorithms together with several related works in Table 1 (in the appendix).

1.2. Related Work

Due to space limitations, only the most relevant works are discussed. Appendix A.1 gives a more thorough discussion.

For the vanilla delayed adversarial MABs (with $[0,1]$-bounded losses), Bistritz et al. (2019); Thune et al. (2019); Zimmert and Seldin (2020) achieved $\widetilde{\mathcal{O}}(\sqrt{D}+\sqrt{KT})$ regret, which is optimal compared to the $\Omega(\sqrt{KT}+\sqrt{D\log K})$ lower bound (Cesa-Bianchi et al., 2016) up to logs. However, our work *does not* build upon any of them. In Appendix A.2, we discuss why we develop a new framework.

For scale-free learning, most works in this line consider full-information feedback (Orabona and Pál, 2018) or stochastic MABs (Hadiji and Stoltz, 2023). The only upper bound for adversarial MABs is $\widetilde{\mathcal{O}}(K\sqrt{T}L)$ (Putta and Agrawal, 2022), which does not consider feedback delays and still has a $\sqrt{K}$ gap with the $\Omega(\sqrt{KT}L)$ lower bound (Auer et al., 2002b). See Remark 6.4 for more discussions on this.

For delayed linear bandits, Ito et al. (2020) studied adversarial linear bandits with *known uniform* delays and achieved $\widetilde{\mathcal{O}}(\sqrt{n(n+d)T})$ regret. All other works (see, e.g., (Zhou et al., 2019; Vernade et al., 2020)) only consider stochastic losses. On the contrary, our algorithm can handle both adversarial losses and *arbitrary and unknown* feedback delays.

2. Problem Setup: Delayed Adversarial MAB

Notations. For $n \geq 1$, we denote the set $\{1, 2, \ldots, n\}$ by $[n]$. We denote the probability simplex over $[n]$ by $\triangle^{[n]}$. We use $\mathbf{0}$ to denote the zero vector. We use $\mathbf{1}_i$ to denote the one-hot vector with 1 on the i-th coordinate, i.e., $(\mathbf{1}_i)_j = \mathbb{1}[i = j]$. We use $\widetilde{\mathcal{O}}$ and $\widetilde{\Theta}$ to ignore all logarithmic dependencies. Let f be a strictly convex function defined on some convex domain $A \subseteq \mathbb{R}^K$. For any $x, y \in A$, if $\nabla f(x)$ exists, denote the Bregman divergence between y and x induced by f as

$$D_f(y,x) \triangleq f(y) - f(x) - \langle \nabla f(x), y - x\rangle.$$

We use $f^*(y) \triangleq \sup_{x\in\mathbb{R}^K}\{\langle y, x\rangle - f(x)\}$ to denote the Fenchel conjugate of f. At last, we use $\overline{f}$ to denote the restriction of function f on $\triangle^{[K]}$, i.e., $\overline{f}(x) = f(x)$ if $x \in \triangle^{[K]}$ and $\overline{f}(x) = \infty$ otherwise.

This section focuses on delayed adversarial MABs (Zimmert and Seldin, 2020), which we use to introduce our `Banker-OMD` framework. For the more general delayed scale-free adversarial MAB setting and delayed adversarial linear bandit setting, please refer to Sections 6 and 7.

In a delayed adversarial MAB, there are $K \geq 2$ available actions (arms) and $T \geq 1$ rounds. For each round $t \in [T]$, there is a loss vector $l_t \in [0,1]^K$ obliviously determined by an adversary. Meanwhile, there is also an oblivious delay d_t associated with this round (also decided by the adversary).[1] We define $D \triangleq \sum_{t=1}^T d_t$ following the convention (Bistritz et al., 2019). Losses and delays are all *hidden* to the agent.

At the beginning of t-th round, the agent chooses an action $A_t \in [K]$. The feedback (t, l_{t,A_t}) will be revealed to the agent at the end of the $t+d_t$-th round. The agent can decide A_t based on all historical actions, all arrived observations, and any private randomness. Note that the agent does **not** have direct access to d_t's, though it can infer the delay d_t *after* its feedback has arrived, i.e., after d_t rounds.

The agent's objective is to minimize the difference between its total loss, i.e., $\sum_{t=1}^T l_{t,A_t}$ and the minimal possible total loss incurred by a single action. Formally speaking, we use pseudo-regret (also referred to as regret for convenience) as the performance metric for any MAB algorithm.

Definition 2.1. The pseudo-regret of an MAB algorithm is

$$\mathfrak{R}_T \triangleq \max_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^T l_{t,A_t} - \sum_{t=1}^T l_{t,i}\right],$$

where the expectation is taken to both the algorithm's internal randomness and randomness from the environment.

3. Vanilla OMD Framework

We review the vanilla Online Mirror Descent (OMD) framework for non-delayed MAB problems in Algorithm 1. It has been widely used in many bandit learning works (e.g., (Abernethy et al., 2008; 2015; Audibert et al., 2014)). Note that we use action scales σ_t instead of learning rates η_t for the ease of presentation, whose relationship is $\sigma_t = \eta_t^{-1}$.

Algorithm 1 Vanilla OMD for MAB

```
Input: Number of arms $K$, rounds $T$, Legendre regularizer
    $\Psi : \mathbb{R}_+^K \to \mathbb{R}$ (defined in Definition B.2), initial action
    $x_1 \in \triangle^{[K]}$, action scales $\sigma_1, \ldots, \sigma_T$.
for $t = 1, 2, \ldots, T$ do
    Sample $A_t \in [K]$ according to $x_t$. Pull arm $A_t$.
    Receive $l_{t,A_t}$. Calculate $\widetilde{l}_t \leftarrow \frac{l_{t,A_t}}{x_{t,A_t}} \mathbf{1}_{A_t}$.
    Set $x_{t+1} \leftarrow \nabla\overline{\Psi}^*(\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t)$.
```

Before presenting our framework, we first sketch the standard analysis of an OMD-based algorithm. By the OMD framework itself, we can conclude that (see, e.g., Theorem 28.4 by Lattimore and Szepesvári (2020)):

$$\mathbb{E}[\langle l_t, x_t - y\rangle] = \mathbb{E}[\langle \widetilde{l}_t, x_t - y\rangle] \qquad (\forall y \in \triangle^{[K]})$$

$$\le \mathbb{E}[\sigma_t D_\Psi(y, x_t) - \sigma_t D_\Psi(y, z_t) + \sigma_t D_\Psi(x_t, \widetilde{z}_t)], \quad (1)$$

where z_t and $\widetilde{z}_t$ are defined as

$$z_t = \nabla\overline{\Psi}^*\big(\nabla\Psi(x_t) - \sigma_t^{-1}\widetilde{l}_t\big),$$
$$\widetilde{z}_t = \nabla\Psi^*\big(\nabla\Psi(x_t) - \sigma_t^{-1}\widetilde{l}_t\big). \quad (2)$$

Eq. (1) then gives the following by telescoping sums:

$$\mathfrak{R}_T \le \sigma D_\Psi(y, x_1) + \sum_{i=1}^T \sigma\, \mathbb{E}[D_\Psi(x_t, \widetilde{z}_t)]. \quad (3)$$

When Ψ satisfies certain conditions, one can find constants C_1 and C_2 such that (see, e.g., (Abernethy et al., 2015)):

- $D_\Psi(y, x_1) \le C_1$ for any $y \in \triangle^{[K]}$, and
- $\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)] \le \frac{C_2}{\sigma_t}$ for any $x_t \in \triangle^{[K]}$ and $\sigma_t > 0$.

In such cases, we say (Ψ, x_1) is (C_1, C_2)-*regular*. Setting $\sigma_t \equiv \sqrt{\frac{C_2}{C_1}T}$ then gives the following according to Eq. (3):

$$\mathfrak{R}_T \le \sqrt{\frac{C_2}{C_1}T}C_1 + T\left(\sqrt{\frac{C_2}{C_1}T}\right)^{-1} C_2$$
$$= \mathcal{O}\left(\sqrt{C_1 C_2 T}\right). \quad (4)$$

However, as noticed, the classical OMD framework heavily relies on the availability of l_{t,A_t} at the end of round t (based on which we compute $\widetilde{z}_t$). Consequently, it's not easy to make an OMD-based algorithm capable of feedback delays.

To resolve this issue, the previous solution in the literature is to carefully design the loss estimators $\widetilde{l}_t$ and/or the regularizer Ψ (see, e.g., (Zimmert and Seldin, 2020; Putta and Agrawal, 2022)). However, such an idea limits the use of OMD to complicated applications like scal-free delayed adversarial MABs or delayed adversarial linear bandits. We, on the other hand, design an enhanced framework that can easily handle feedback delays, leaving the regularizers and estimators to the user, as we will see in Sections 6 and 7.

4. `Banker-OMD` for Delayed Bandit Feedback

We are ready to introduce our `Banker-OMD` framework. We begin by inspecting the classical Eq. (1) for the vanilla OMD framework, with a new set of terminologies for the terms This gives a better understanding of our framework.

4.1. Terminologies in Regret Decomposition

To begin with, we sum Eq. (1) over $t = 1, 2, \ldots, T$ for an upper bound of $\mathfrak{R}_T$ and rename the terms as follows:

$$\mathfrak{R}_T \le \sum_{t=1}^T \underbrace{\mathbb{E}[\sigma_t D_\Psi(y, x_t)]}_{\text{withdrawal}} - \sum_{t=1}^T \underbrace{\mathbb{E}[\sigma_t D_\Psi(y, z_t)]}_{\text{saving}}$$

[1] For simplicity, we focus on the standard arm-independent delay model in the main text. Our framework is actually capable of the more general *arm-dependent* delay model; see Appendix A.3.

$$+\sum_{t=1}^{T}\underbrace{\mathbb{E}[\sigma_t D_\Psi(x_t,\widetilde{z}_t)]}_{\text{immediate cost}}, \tag{5}$$

where z_t and $\widetilde{z}_t$ are defined in Eq. (2).

We now explain the idea behind the names in Eq. (5). Let us first look at the third term $\sigma_t D_\Psi(x_t,\widetilde{z}_t)$. As mentioned in Section 3, one can bound $\mathbb{E}[\sigma_t D_\Psi(x_t,\widetilde{z}_t)\mid\mathcal{F}_{t-1}]\le\frac{C_2}{\sigma_t}=\mathcal{O}(\sigma_t^{-1})$ when (Ψ,x_0) is chosen to be "good enough." In other words, they can be bounded without affecting the first two terms. Hence, we refer to them as "immediate costs" as they immediately contribute to $\mathfrak{R}_T$ once σ_t's are decided.

Then consider $\sum_{t=1}^T\mathbb{E}[\sigma_t D_\Psi(y,x_t)]-\sum_{t=1}^T\mathbb{E}[\sigma_t D_\Psi(y,z_t)]$. If we choose a constant scale $\sigma_t=\sigma$ and pick $x_{t+1}\leftarrow z_t$ (as in Section 3), it becomes a telescoping sum bounded by $\sigma D_\Psi(y,x_1)-\sigma D_\Psi(y,z_s)$. Therefore, we can view it in a step-by-step manner: In round t, we manage to reduce some regret by incurring a negative term $-\sigma D_\Psi(y,z_t)$. However, we will immediately use this term to pay the new regret $\sigma D_\Psi(y,x_{t+1})$. After that, we compute z_{t+1} and introduce a new negative term $-\sigma D_\Psi(y,z_{t+1})$ for subsequent rounds. Thus, the $-\sigma D_\Psi(y,z_t)$ terms can be viewed as bank "savings," which ensures that one can "withdraw" $\sigma D_\Psi(y,x_{t+1})$ to pay the cost of x_{t+1} in the future, without overdrafting.

4.2. Beyond Telescoping: Utilizing Multiple Savings

We generalize the banker idea above to tackle the challenge of potentially absent feedback: We use *more than one* "saving terms" accumulated in previous rounds to make up the current "withdrawal" $\widetilde{\sigma}D_\Psi(y,x)$ – in contrast to vanilla OMD where only a *single* saving is used.

Formally, we focus on a specific round $t\in[T]$. We aim to design the action x_t and the action scale σ_t wisely so that the withdrawal $\sigma_t D_\Psi(y,x_t)$ is covered by several previous savings from rounds $1\le t_1<t_2<\cdots<t_h<t$, namely $\sigma_{t_1}D_\Psi(y,z_{t_1})$, $\sigma_{t_2}D_\Psi(y,z_{t_2})$, ..., $\sigma_{t_h}D_\Psi(y,z_{t_h})$.[2] In this section, we treat $t_1,t_2,\ldots,t_h$ as some given sequence – its decision rule is the focus of the next section.

In this case, we have Lemma 4.1, where the choice of x in Eq. (6) is analog to z_t used in vanilla OMD (*c.f.* Eq. (2)).

Lemma 4.1. *For any $z_1,\ldots,z_h\in\triangle^{[K]}$, $\sigma_1,\ldots,\sigma_h>0$ and Legendre convex $\Psi:\mathbb{R}_+^K\to\mathbb{R}$, let $\widetilde{\sigma}=\sum_{i=1}^h\sigma_i$ and*

$$x=\nabla\overline{\Psi}^*\left(\sum_{i=1}^h\frac{\sigma_i}{\widetilde{\sigma}}\nabla\Psi(z_i)\right), \tag{6}$$

then we have

$$\sigma D_\Psi(y,x)\le\sum_{i=1}^m\sigma_i D_\Psi(y,z_i),\quad\forall y\in\triangle^{[K]}.$$

In other words, once we set σ_t as $\sum_{i=1}^h\sigma_{t_i}$ and construct x_t from Eq. (6), we automatically ensure $\sigma_t D_\Psi(y,x_t)\le\sum_{i=1}^h\sigma_{t_i}D_\Psi(y,z_{t_i})$ – i.e., the t-th withdrawal is covered by the savings from rounds $t_1,t_2,\ldots,t_h$ – as desired.

4.3. "Over-Drafting" in Case of Saving Shortage

In the presence of delayed feedback, the total accumulated savings might not be always enough. For example, there is no feedback in the first few rounds, but we still need to decide x_t in real time. To cover the remaining part of the withdrawal, we propose to "over-draft" some more savings via investing on a *default action* $x_0\in\triangle^{[K]}$. Specifically, if we want to pick an action scale $\sigma_t>\sum_{i=1}^h\sigma_{t_i}$. Let the shortage be $b_t=\sigma_t-\sum_{i=1}^h\sigma_{t_i}$. We then add and subtract $\mathbb{E}[b_t]D_\Psi(y,x_0)$ to the RHS of Eq. (5), which gives

$$\begin{aligned}\mathfrak{R}_T\le&\sum_{t=1}^T\underbrace{\mathbb{E}[\sigma_t D_\Psi(y,x_t)]}_{\text{withdrawal}}-\sum_{t=1}^T\underbrace{\mathbb{E}[\sigma_t D_\Psi(y,z_t)]}_{\text{saving}}\\&-\underbrace{\mathbb{E}[b_t]D_\Psi(y,x_0)}_{\text{imaginary saving}}+\sum_{t=1}^T\underbrace{\mathbb{E}[\sigma_t D_\Psi(x_t,\widetilde{z}_t)]}_{\text{immediate cost}}\\&+\underbrace{\mathbb{E}[b_t]D_\Psi(y,x_0)}_{\text{investment cost}}.\end{aligned} \tag{7}$$

As desired, we can use the negative term $-b_t D_\Psi(y,x_0)$ as an "imaginary" saving – in addition to the h "actual" savings $\sum_{i=1}^h\sigma_{t_i}D_\Psi(y,z_{t_i})$ – to apply Lemma 4.1. This allows us to pay the withdrawal due to the following action:

$$x_t=\nabla\overline{\Psi}^*\left(\sum_{i=1}^h\frac{\sigma_{t_i}}{\sigma_t}\nabla\Psi(z_{t_i})+\frac{b_t}{\sigma_t}\nabla\Psi(x_0)\right), \tag{8}$$

which exactly uses up the h actual savings $\sigma_{t_i}D_\Psi(y,z_{t_i})$ together with the imaginary one $-b_t D_\Psi(y,x_0)$.

We call b_t an "investment": When the total savings are insufficient for a new action, we "invest" in some x_0 to make up the difference and proceed. As illustrated in Eq. (7), such investments are not for free. They each introduces an *investment cost* $\mathbb{E}[b_t]D_\Psi(y,x_0)$ to our total regret.

Intuitively, if for each $t\in[T]$ we can assign a sequence of previous savings $\{(\sigma_{t,i},z_{t,i})\}_{i=1}^{h_t}$ together with an investment b_t, we can set $\sigma_t=b_t+\sum_{i=1}^{h_t}\sigma_{t,i}$, determine x_t according to Eq. (8), and yield the following regret bound:

$$\mathfrak{R}_T\lesssim\underbrace{\mathbb{E}[B_T]\cdot D_\Psi(y,x_0)}_{\text{total investment cost}}+\sum_{t=1}^T\underbrace{\mathbb{E}[\sigma_t D_\Psi(x_t,\widetilde{z}_t)]}_{\text{immediate cost}}, \tag{9}$$

where $B_T\triangleq\sum_{t=1}^T b_t$ is called the *total investment*.

[2] The vanilla OMD covered in the previous section therefore reduces to the special case of $h=1$ and $t_1=t-1$.

Algorithm 2 `Banker-OMD` Framework

```
Input: Number of arms $K$, regularizer and default investment $(\Psi, x_0)$, subroutine to pick action scales $S$
Output: A sequence of actions $A_1, A_2, \ldots \in [K]$
Initialize $B_0 \leftarrow 0$    ▷ maintain the total investment
for $t = 1, 2, \ldots, T$ do
    $a_t \leftarrow$ missing.    ▷ whether feedback of $t$ has arrived
    $\sigma_t \leftarrow S(t, \{(A_s, a_s, l_{s,A_s}, \sigma_s, v_s)\}_{s<t})$.    ▷ a user-specified rule of deciding the action scale $\sigma_t$
    $v_t \leftarrow \sigma_t$.    ▷ coefficient of the saving term $D_\Psi(y, z_t)$
    Initialize $b_t \leftarrow \sigma_t$. For all $s < t$ such that $a_s =$ arrived, set $\sigma_{t,s} = \min\{v_s, b_t\}$ and $b_t \leftarrow b_t - \sigma_{t,s}$.    ▷ determine the amount of "investment" and "savings" to form $x_t$ by minimizing $b_t$ while ensuring Eq. (10)
    $B_t \leftarrow B_{t-1} + b_t$.    ▷ update total investment $B_t$
    for $s = 1, 2, \ldots, t-1$ do
        $v_s \leftarrow v_s - \sigma_{t,s}$.    ▷ spend $\sigma_{t,s}$ units of savings
    $x_t \leftarrow \nabla\overline{\Psi}^*(\frac{1}{\sigma_t}\sum_{s=1}^{t-1}\sigma_{t,s}\nabla\Psi(z_s) + \frac{b_t}{\sigma_t}\nabla\Psi(x_0))$.    ▷ decide $x_t$ according to Lemma 4.1 and Eq. (8)
    Sample $A_t \in [K]$ according to $x_t$, pull arm $A_t$.
    for upon receiving each new feedback $(s, l_{s,A_s})$ do
        $\widetilde{l}_s \leftarrow \frac{l_{s,A_s}}{x_{s,A_s}}\mathbf{1}_{A_s}$. $z_s \leftarrow \nabla\overline{\Psi}^*(\nabla\Psi(x_s) - \frac{1}{\sigma_s}\widetilde{l}_s)$.
        $a_s \leftarrow$ arrived.    ▷ saving $-\sigma_s D_\Psi(y, x_s)$ available
```

4.4. Formal Framework: `Banker-OMD`

In this section, we focus on assigning $\{(\sigma_{t,i}, z_{t,i})\}_{i=1}^{h_t}$ and b_t to every round $1 \le t \le T$ so that Eq. (9) is ensured.

For each round $s < t$, we denote its corresponding "remaining savings" as v_s, i.e., the coefficient before $D_\Psi(y, z_s)$ in the current regret upper bound is $-v_s$. Thus, v_s is initialized to 0 and becomes σ_s once the s-th feedback arrives (so z_s is revealed and the saving $\sigma_s D_\Psi(y, z_s)$ becomes available). As sketched in the previous section, we must decompose each action scale σ_t into an investment b_t together with the sum of several previous savings – if we use $\sigma_{t,s}$ units of saving from round s when constructing σ_t, then we need

$$b_t + \sum_{s=1}^{t-1} \sigma_{t,s} = \sigma_t, \quad \text{and} \quad \sigma_{t,s} \le v_s, \tag{10}$$

As b_t causes extra investment cost, we shall minimize b_t while ensuring Eq. (10) – in our framework, we only allocate b_t if using up all previous v_s's is still insufficient.

We formally present our framework in Algorithm 2, which we call `Banker-OMD`. As we sketched in the previous section, we shall expect its regret to be controlled by the total investment B_T, the action scales σ_t, and the remaining savings v_t – which is formalized in the following theorem.

Theorem 4.2 (`Banker-OMD` Regret Bound)**.** *At the end of any round T, for any $y \in \triangle^{[K]}$, we have*

$$\sum_{t=1}^{T}\langle \widetilde{l}_t, x_t - y\rangle \le \underbrace{B_T \cdot D_\Psi(y, x_0)}_{\text{total investment cost}} + \underbrace{\sum_{t=1}^{T}\sigma_t D_\Psi(x_t, \widetilde{z}_t)}_{\text{total immediate costs}} - \underbrace{\sum_{t=1}^{T} v_t D_\Psi(y, z_t)}_{\text{remaining savings}}, \tag{11}$$

where $B_T, \widetilde{l}_1, \ldots, \widetilde{l}_t, x_1, \ldots, x_t, z_1, \ldots, z_t$ are the variable values produced by Algorithm 2 at the end of round T and $\widetilde{z}_t$ is defined as $\widetilde{z}_t = \nabla\Psi^*(\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t)$.

Theorem 4.2 offers a regret bound for `Banker-OMD` under general parameter selection and serves as the basis for achieving $\widetilde{\mathcal{O}}(\sqrt{T} + \sqrt{D})$-style regret bounds by choosing proper configurations (see Theorem 4.6 for an easy example). The proof of Theorem 4.2 generally follows from our informal intuition, i.e., inductively applying Lemma 4.1. A formal version can be found in Appendix B.

Imitating Eq. (4) for vanilla OMD, we give the following corollary by assuming (C_1, C_2)-regularity in Theorem 4.2.

Corollary 4.3. *If (Ψ, x_0) is (C_1, C_2)-regular, i.e., $D_\Psi(y, x_0) \le C_1$ for all $y \in \triangle^{[K]}$ and $\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t) \mid \mathcal{F}_{t-1}] \le \frac{C_2}{\sigma_t}$ for all $t \in [T]$, then the `Banker-OMD` framework presented in Algorithm 2 ensures*

$$\mathfrak{R}_T \le C_1 \cdot \mathbb{E}[B_T] + C_2 \cdot \mathbb{E}\left[\sum_{t=1}^{T}\sigma_t^{-1}\right].$$

Remark 4.4. If there is no delay and we set $\sigma_t \equiv \sigma$, then `Banker-OMD` coincides with the vanilla OMD with the same σ – which is indeed our inspiration for `Banker-OMD`.

4.5. Tuning the Action Scales Painlessly

One may notice that in Algorithm 2, the regularizer Ψ and the action scales σ_t remain unspecified. In fact, our `Banker-OMD` framework inherits the flexibility of vanilla OMD – the users can arbitrarily pick Ψ and σ_t as they want. For example, Ψ can be any popular regularizer including negative entropy, log-barrier (Wei and Luo, 2018), and Tsallis entropy (Abernethy et al., 2015). The action scale σ_t can also be freely decided, e.g., they may depend on the current round index t and the statistics of experienced delays.[3]

In this section, we exemplify a possible tuning of σ_t which automatically ensures $\widetilde{\mathcal{O}}(\sqrt{T} + \sqrt{D})$-style regret bounds in delayed bandit learning problems. In the following, we assume (Ψ, x_0) to be (C_1, C_2)-regular, which solely depends on the the choice of (Ψ, x_0) picked by the user.

[3] As a side note, we *do not* require σ_t's to be monotone – even more flexible than the vanilla OMD framework.

Let $\mathfrak{d}_t$ denote the number of feedback that has not arrived at the beginning of time t. We define $\mathfrak{D}_t = \sum_{s=1}^t \mathfrak{d}_s$ as the cumulative experienced delay up to time t – both $\mathfrak{d}_t$ and $\mathfrak{D}_t$ can be easily maintained at run-time. Since $\sum_{t=1}^T \frac{1}{\sqrt{t}} = \mathcal{O}(\sqrt{T})$ and $\sum_{t=1}^T \frac{\mathfrak{d}_t}{\sqrt{\mathfrak{D}_t}} = \mathcal{O}(\sqrt{\mathfrak{D}_T}) = \mathcal{O}(\sqrt{D})$ (Lemma B.1), one only needs to tune the action scales σ_t as

$$\sigma_t = \left(\frac{1}{\sqrt{t}} + \frac{\mathfrak{d}_t}{\sqrt{\mathfrak{D}_t}}\right)^{-1}, \quad \forall t \geq 1$$

to upper bound the total immediate costs as

$$C_2 \sum_{t=1}^T \sigma_t^{-1} = \mathcal{O}\left(C_2(\sqrt{T} + \sqrt{D})\right).$$

For the total investment part, we make the following observation, which follows from a direct analysis of Line 6:

Lemma 4.5. *Let T_0 be the largest number such that $b_{T_0} > 0$. Then, at the end of time T, we can decompose B_T as*

$$B_T = B_{T_0} = \sigma_{T_0} + \sum_{t=1}^{T_0-1} \mathbb{1}[t + d_t \geq T_0]\sigma_t.$$

Suppose that at the beginning of round T_0, there are still m feedback on the way, each corresponding to rounds $t_1 < \cdots < t_m$. Since these m feedback contributes at least $\binom{m+1}{2}$ to the total delay D, we have $m = \mathcal{O}(\sqrt{D})$. Thus, under the action scale $\sigma_t = \left(\frac{1}{\sqrt{t}} + \frac{\mathfrak{d}_t}{\sqrt{\mathfrak{D}_t}}\right)^{-1}$, we have

$$\begin{aligned} B_T &= \sigma_{T_0} + \sum_{i=1}^m \sigma_{t_i} \leq \sqrt{t_1} + \sum_{i=2}^m \frac{\sqrt{\mathfrak{D}_{t_i}}}{\mathfrak{d}_{t_i}} + \frac{\sqrt{\mathfrak{D}_{T_0}}}{\mathfrak{d}_{T_0}} \\ &\leq \sqrt{T} + \sum_{i=1}^m \frac{\sqrt{D}}{i} = \mathcal{O}\left(\sqrt{T} + \sqrt{D}\log D\right). \end{aligned}$$

Therefore, we obtain the following theorem:

Theorem 4.6 (Main Theorem: `Banker-OMD` with Delays)**.** *When setting σ_t as follows in Algorithm 2:*

$$\sigma_t = \left(\frac{1}{\sqrt{t}} + \mathfrak{d}_t\sqrt{\frac{\ln(\mathfrak{D}_t + 1)}{\mathfrak{D}_t}}\right)^{-1}, \quad \forall 1 \leq t \leq T,$$

we have $B_T \leq \mathcal{O}(\sqrt{T} + \sqrt{D\log D})$ at the end of round T. Moreover, if in addition (Ψ, x_0) is (C_1, C_2)-regular, setting

$$\sigma_t = \sqrt{\frac{C_2}{C_1}}\left(\frac{1}{\sqrt{t}} + \mathfrak{d}_t\sqrt{\frac{\ln(\mathfrak{D}_t + 1)}{\mathfrak{D}_t}}\right)^{-1}, \quad \forall 1 \leq t \leq T$$

guarantees

$$\mathfrak{R}_T = \mathcal{O}\left(\sqrt{C_1 C_2}(\sqrt{T} + \sqrt{D\log D})\right).$$

Remark 4.7. This theorem only provides an easy way of getting $\widetilde{\mathcal{O}}(\sqrt{T} + \sqrt{D})$ regret. As we said, the action scales are up to the user – as we will demonstrate shortly, a different and more sophisticated design of σ_t is used in Section 6.

5. Application 0: Delayed Adversarial MAB

For a sanity check and a simple yet illustrative example, we consider the standard delayed adversarial MAB setting. Inspired by the `Tsallis-INF` algorithm for adversarial MABs (Zimmert and Seldin, 2019), we use the $1/2$-Tsallis entropy regularizer $\Psi(x) = -\sum_{i=1}^K 2\sqrt{x_i}$. To ensure regularity, we pick $x_0 = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$, which ensures (Ψ, x_0) is $(\mathcal{O}(\sqrt{K}), \mathcal{O}(\sqrt{K}))$-regular (Abernethy et al., 2015). At last, as suggested by Theorem 4.6, we determine σ_t^{-1} as $\frac{1}{\sqrt{t}} + \mathfrak{d}_t\sqrt{\frac{\ln(\mathfrak{D}_t+1)}{\mathfrak{D}_t}}$. Plugging all these configurations into Algorithm 2, we get the algorithm `Banker-TINF`, whose performance guarantee is stated in Corollary 5.1.

Corollary 5.1. *`Banker-TINF` applied to delayed adversarial MABs ensures $\mathfrak{R}_T = \mathcal{O}\left(\sqrt{KT} + \sqrt{KD\log D}\right)$.*

Remark 5.2. While the delay-related term $\mathcal{O}(\sqrt{KD\log D})$ is slightly worse than the SOTA result $\mathcal{O}(\sqrt{D\log K})$ (Zimmert and Seldin, 2020), we see that `Banker-OMD` conveniently translates algorithm design into finding a regularizer with appropriate regularity, which is a task decoupled from handling feedback delays. As we will present in the following sections, this feature enables us to conveniently adapt existing OMD-based algorithms from a large variety of non-delayed bandits settings to their delayed counterparts.

6. Application I: Delayed Scale-Free MAB

In prior sections of this paper, we assumed by default that all losses are $[0, 1]$-bounded. However, in practice, it is often hard to quantize the feedback to a limited range in a deterministic way. For example, one advertisement may cause unexpectedly great reactions in the market, either positive or negative. Therefore, the $[0, 1]$-bounded-loss assumption may fail to capture such subtle but important scenarios — moreover, as one will see shortly, algorithm design in such a setting is hard, even in absence of feedback delays.

In a scale-free adversarial MAB problem (without feedback delays), all losses $l_{t,a}$'s are within some bounded interval $[-L, L]$ and L is *unknown* to the agent.[4] Due to the $\Omega(\sqrt{KT})$ lower bound for standard adversarial MABs (Auer et al., 2002b), a trivial lower bound $\Omega(\sqrt{KTL})$ exists. On the other hand, the best-known upper bound by Putta and Agrawal (2022) only reduce to $\widetilde{\mathcal{O}}(K\sqrt{T}L)$ in the worst

[4] The assumption on the boundedness of the interval is without loss of generality because L is kept as a secret: viewing from hindsight, such an $L < \infty$ always exists.

case — a $\sqrt{K}$ gap with the lower bound exists even without feedback delays (see Remark 6.4 for more discussion).

A scale-free MAB algorithm is more robust to extreme feedback, but we want more — e.g., being robust to feedback delays. This results in the novel setting which we call delayed scale-free adversarial MAB, where the feedback of each round t will only be delivered after d_t rounds; meanwhile the losses still fall into an unknown bounded range $[-L, L]$ instead of $[0, 1]$. This is again the reality: in advertisement recommendations, the effect of deploying an advertisement also cannot be immediately observed.

6.1. High-Level Design Ideas

In this section, we outline the design idea of our algorithm. For this section, we shall assume the losses to be non-negative — this assumption enables us to use the $1/2$-Tsallis entropy regularizer, following the idea of Corollary 4.3.[5]

As scale-free losses already produce adequate difficulties (Putta and Agrawal, 2022), we first present our solution for non-delayed settings before presenting the delay-robust algorithms. Assume that we face a non-delayed instance where $d_t \equiv 0$ and we additionally have a *forecast* oracle about $\lVert l_t \rVert_\infty$ at the beginning of each round t (i.e., $l_{t,A_t} \in [0, \lVert l_t \rVert_\infty]$ always holds). With the non-delay assumption and the forecast, we can follow the spirit of non-delayed adversarial MAB algorithms (Zimmert and Seldin, 2019) for an algorithm with $\mathcal{O}(L\sqrt{KT})$ regret, as follows:

Idealized Setting #1 (no delay, with $\lVert l_t \rVert_\infty$ forecast). With the $1/2$-Tsallis entropy regularizer, the total immediate costs become $\sqrt{K} \sum \sigma_t^{-1} l_{t,A_t^2}$. Analogue to the $[0, 1]$-bounded case (Zimmert and Seldin, 2019), we want it to be of order σ_t after applying the summation lemma (Lemma B.1). Thus,

$$\sigma_t = \left(1 + \sum_{s=1}^{t} \lVert l_s \rVert_\infty^2\right)^{1/2}$$

ensures the total immediate costs to be bounded by

$$\begin{aligned}
&\sqrt{K} \sum_{t=1}^{T} \lVert l_t \rVert_\infty^2 \left(1 + \sum_{s=1}^{t} \lVert l_s \rVert_\infty^2\right)^{-1/2} \\
&= \mathcal{O}\left(\sqrt{K} \sqrt{1 + \sum_{s=1}^{T} \lVert l_s \rVert_\infty^2}\right) = \mathcal{O}(\sqrt{KT}L).
\end{aligned}$$

The same σ_t also ensures the total investment cost to be $\sigma_T \sqrt{K} = \mathcal{O}(\sqrt{KT}L)$. Hence, $\mathfrak{R}_T = \mathcal{O}(\sqrt{KT}L)$.

The above reasoning shows that we can easily adapt a classical $[0, 1]$-valued MAB algorithm to scale-free settings with the help of accurate forecasts. In fact, with such forecasts available, we can also generalize `Banker-TINF` for delayed adversarial MABs (Corollary 5.1) to scale-free tasks.

Idealized Setting #2 (has delay, with $\lVert l_t \rVert_\infty$ forecast). Inspired by Theorem 4.6, we set

$$\sigma_t = (\mathfrak{d}_t + 1)^{-1} \cdot \left(1 + \sum_{s=1}^{t} (\mathfrak{d}_s + 1) \lVert l_s \rVert_\infty^2\right)^{1/2},$$

where we weigh the delays in $\mathfrak{D}_t$ by the loss magnitudes. The total immediate costs then become proportional to

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\sigma_t^{-1} l_{t,A_t}^2] \\
&\le \sum_{t=1}^{T} \lVert l_t \rVert_\infty^2 (\mathfrak{d}_t + 1) \cdot \left(1 + \sum_{s=1}^{t} (\mathfrak{d}_s + 1) \lVert l_s \rVert_\infty^2\right)^{-1/2} \\
&= \mathcal{O}\left(\sqrt{\sum_{t=1}^{T} (\mathfrak{d}_t + 1) \lVert l_t \rVert_\infty^2}\right) = \mathcal{O}(\sqrt{D + T}L),
\end{aligned}$$

where we again used Lemma B.1 together with the fact that $\sum_t (\mathfrak{d}_t + 1) \le D + T$. Following the proof of Theorem 4.6, we can bound the total investment as $B_T = \widetilde{\mathcal{O}}\left(\sqrt{\sum_{t=1}^{T} (\mathfrak{d}_t + 1) \lVert l_t \rVert_\infty^2}\right) = \widetilde{\mathcal{O}}(\sqrt{D + T}L)$. Combining these two parts then gives $\mathfrak{R}_T = \widetilde{\mathcal{O}}(\sqrt{K(D + T)}L)$.

Actual Setting. In the actual situation without the $\lVert l_t \rVert_\infty$ oracle, we introduce a doubling trick to maintain an upper bound $\widehat{L}_t$ of the maximum observed feedback l_{s,A_s}. We then replace all unknown $\lVert l_s \rVert_\infty$'s in σ_t with $\widehat{L}_t$, pretending that it is a prediction returned by an ideal oracle, i.e.,

$$\sigma_t = (\mathfrak{d}_t + 1)^{-1} \cdot \left(1 + \sum_{s=1}^{t} (\mathfrak{d}_s + 1) \widehat{L}_s^2\right)^{1/2}. \tag{12}$$

While maintaining the estimations $\widehat{L}_t$, if we receive feedback l_{s,A_s} exceeding the current $\widehat{L}_t$ (i.e., we underestimated the real losses), we update the upper bound $\widehat{L}$ to $2l_{s,A_s}$ and *skip* that round by setting the loss estimator $\widetilde{l}_s$ to $\mathbf{0}$ (see Line 11). As there is at most $\mathcal{O}(\sqrt{D})$ feedback on the way, each doubling only causes $\mathcal{O}(\sqrt{D})$ rounds to be skipped. Their losses can be bounded by the new value of $\widehat{L}_t$, which means the total regret caused by skipping is at most $\sum_{\widehat{L}_t} \mathcal{O}(\sqrt{D}L_t) = \mathcal{O}(\sqrt{D}L)$ — informally, by doing so, we get a "weak oracle" that upper bounds each $|l_{t,A_t}|$ within a constant multiplicative error, which is sufficient for $\widetilde{\mathcal{O}}(\sqrt{T} + \sqrt{D})$-style regret as we show in Section 6.3.

6.2. `Banker-SFTINF` for Non-Negative Losses

We formalize the ideas presented in the last section into Algorithm 3, with a regret guarantee stated in Theorem 6.1.

[5] For the general loss case, the log-barrier regularizer can be used to derive an even better regret guarantee (which adapts to the losses), albeit with a more complicated analysis; see Section 6.4.

Algorithm 3 `Banker-SFTINF` for Delayed Scale-Free Adversarial MAB with Non-Negative Losses

```
Initialize $\widehat{L}_1 = 1$. Deploy Banker-OMD (with modifications) with $x_0 = 1/K$ and the 1/2-Tsallis entropy regularizer $\Psi(x) = -2\sum_{i=1}^K \sqrt{x_i}$ as follows.
for $t = 1, 2, \ldots, T$ do
    Set $a_t \leftarrow$ missing and $\mathfrak{d}_t \leftarrow \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s=\text{missing}\}}$.
    Set $D_t \leftarrow \sum_{s=1}^{t}(\mathfrak{d}_s + 1)\widehat{L}_s^2$.  ▷ implement Eq. (12)
    Calculate $\sigma_t = \left((\mathfrak{d}_t + 1)\sqrt{\frac{\ln(3+D_t/\widehat{L}_t^2)}{3+D_t}}\right)^{-1}$.
    Pick new action $x_t$ as Lines 5 – 10 of Algorithm 2.
    Play $A_t \sim x_t$ and initialize $\widehat{L}_{t+1} \leftarrow \widehat{L}_t$.
    for upon receiving each new feedback $(s, l_{s,A_s})$ do
        $\widehat{L}_{t+1} \leftarrow \max\{\widehat{L}_{t+1}, 2l_{s,A_s}\}$.  ▷ doubling trick
        if $l_{s,A_s} > \widehat{L}_s$ then
            $\widetilde{l}_s \leftarrow \mathbf{0}$; $a_s \leftarrow$ skipped.  ▷ skip $s$
        else
            $\widetilde{l}_s \leftarrow \frac{l_{s,A_s}}{x_{s,A_s}}\mathbf{1}_{A_s}$; $a_s \leftarrow$ arrived.  ▷ keep $s$
        Calculate $z_s \leftarrow \nabla\overline{\Psi}^*(\nabla\Psi(x_s) - \frac{1}{\sigma_s}\widetilde{l}_s)$.
```

Theorem 6.1. *When the losses are non-negative, the* `Banker-SFTINF` *algorithm in Algorithm 3 ensures*

$$\mathfrak{R}_T = \mathcal{O}\left(\sqrt{K(D+T)\log(D+T)} \cdot L\right) = \widetilde{\mathcal{O}}\left(\sqrt{K(D+T)}L\right).$$

Remark 6.2. As noticed by Cesa-Bianchi et al. (2016), any delayed adversarial MAB algorithm must suffer $\Omega(\sqrt{KT} + \sqrt{D\log K})$ regret when losses are $[0,1]$-bounded. Thus, in the scale-free setting, any reasonable algorithm must suffer $\Omega((\sqrt{KT} + \sqrt{D\log K}) \cdot L)$ total regret (even assuming a known L). Hence, regarding K as a small constant independent of T, Algorithm 3 is only $\mathcal{O}(\sqrt{\log(D+T)}) = \widetilde{\mathcal{O}}(1)$ times worse than the theoretical lower bound.

6.3. Analysis Sketch of `Banker-SFTINF`

We sketch the analysis of `Banker-SFTINF`. A formal version is presented in Appendix C. Denote by $\mathcal{E}_t$ the event that $a_t =$ skipped" eventually. Let $\widehat{l}_t \triangleq \frac{l_{t,A_t}}{x_{t,A_t}}\mathbf{1}_{A_t}$ be the unnullified estimator. We decompose the regret as follows:

$$\begin{aligned}\mathfrak{R}_T &= \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\neg\mathcal{E}_t}\langle\widehat{l}_t, x_t - \mathbf{1}_{i^*}\rangle] + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t}\langle\widehat{l}_t, x_t - \mathbf{1}_{i^*}\rangle] \\ &\le \underbrace{\sum_{i=1}^T \mathbb{E}[\langle\widetilde{l}_t, x_t - \mathbf{1}_{i^*}\rangle]}_{\text{Banker-OMD controlled regret}} + \underbrace{\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} l_{t,A_t}]}_{\|l_t\|_\infty \text{ estimation error}} \qquad (13)\end{aligned}$$

because $\widetilde{l}_t = \mathbb{1}_{\neg\mathcal{E}_t}\widehat{l}_t$ by definition. Since $\widetilde{l}_t$'s are the losses fed into `Banker-OMD`, the first term decomposes into the total investment cost $\mathbb{E}[B_T] \cdot D_\Psi(y, x_0)$ and the immediate costs $\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\neg\mathcal{E}_t}\sigma_t D_\Psi(x_t, \widetilde{z}_t)]$, thanks to Theorem 4.2.

First, consider the total investment B_T. Let the last new investment be made at T_0 (Lemma 4.5). Suppose that there are m feedback corresponding to rounds $T_1, T_2, \ldots, T_m$ absent at T_0 (i.e., $T_i + d_{T_i} \ge T_0$). Then B_T is bounded by

$$\begin{aligned}B_T &\le \sum_{i=0}^m \frac{1}{\mathfrak{d}_{T_i} + 1}\sqrt{3 + D_{T_i}} \\ &= \widetilde{\mathcal{O}}\left(\sqrt{(D+T)L^2}\right),\end{aligned}$$

because $m = \mathcal{O}(\sqrt{D})$ and $\sum_{t=1}^T(\mathfrak{d}_t + 1) = \mathcal{O}(D+T)$.

For immediate costs, we can prove $\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)] \le \sqrt{K}\sigma_t^{-1}\|l_t\|_\infty^2$ analogue to the $[0,1]$-bounded cases (formally stated in Lemma C.2). Moreover, we have

$$\begin{aligned}\sum_{t=1}^T \sigma_t^{-1}\|l_t\|_\infty^2 &\le \sqrt{\sum_{t=1}^T(\mathfrak{d}_t + 1)\widehat{L}_t^2} \\ &\le \sqrt{(D+T)}L\end{aligned}$$

due to the summation lemma Lemma B.1 and $\|l_t\|_\infty \le \widehat{L}_t$.

Besides, the other term in Eq. (13) can never exceed $O(\sqrt{D}L)$ — intuitively, each "skipping" will only cause no more than $\sqrt{D}$ feedback on the way to be skipped. Therefore, Eq. (13) gives $\mathfrak{R}_T = \widetilde{\mathcal{O}}(\sqrt{K(D+T)}L)$, as claimed.

6.4. `Banker-SFLBINF` for General Losses

As mentioned, the 1/2-Tsallis entropy regularizer only works for non-negative losses. Fortunately, for the general-loss case, we can use the log-barrier regularizer $\Psi(x) = -\sum_{i=1}^K \ln x_i$ to derive the `Banker-SFLBINF` algorithm (Algorithm 4 in Appendix D.1). It not only allows us to derive the same regret guarantee as Theorem 6.1 when losses can be negative (up to logarithmic factors), but also gives a small-loss style (Neu, 2015) adaptive regret bound. The bound is stated in Theorem 6.3 and proved in Appendix D.

Theorem 6.3. `Banker-SFLBINF` *in Algorithm 4 ensures*

$$\mathfrak{R}_T = \widetilde{\mathcal{O}}\left(\sqrt{K\,\mathbb{E}[\widetilde{\mathfrak{L}}_T^2]} + \sqrt{KD}L\right),$$

where $\widetilde{\mathfrak{L}}_T^2 \triangleq \sum_{t=1}^T(\mathfrak{d}_t + 1)l_{t,A_t}^2$ *and* $\mathfrak{d}_t$ *is the number of feedback absent at the beginning of round* t*. In particular,*

$$\begin{aligned}\mathfrak{R}_T &= \mathcal{O}\left(\sqrt{K(D+T)}\log TL + \sqrt{D}L\log L\right) \\ &= \widetilde{\mathcal{O}}\left(\sqrt{K(D+T)}L\right).\end{aligned}$$

Remark 6.4. When running on non-delayed instances, `Banker-SFLBINF` enjoys an $\widetilde{\mathcal{O}}(\sqrt{KTL} + KL)$ regret guarantee. Meanwhile, the SOTA bounds are $\widetilde{\mathcal{O}}(\sqrt{KL_2} + L_\infty\sqrt{KT})$ or $\widetilde{\mathcal{O}}(\sqrt{KL_2}+L_\infty\sqrt{KL_1})$ (Putta and Agrawal, 2022), which both reduce to $\widetilde{\mathcal{O}}(K\sqrt{TL})$ in the worst case. Hence, our algorithm dominates theirs by a factor of $\sqrt{K}$ in the worst case, closing the gap with the $\Omega(\sqrt{KTL})$ lower bound (Auer et al., 2002b). More importantly, our bound is *uniformly* better than theirs because $\widetilde{\mathfrak{L}}_T^2 = \sum_{t=1}^T l_{t,A_t}^2 \le \sum_{t=1}^T \|l_t\|_2^2 = L_2$. Besides, a technical comparison with (Putta and Agrawal, 2022) is presented in Appendix A.4.

7. Application II: Delayed Linear Bandits

In a delayed adversarial linear bandit, there is a convex action set $\mathcal{C} \subseteq \mathbb{R}^n$ where n is called the ambient dimension. Correspondingly, there is a loss set $\overline{\mathcal{C}} = \{l \in \mathbb{R}^n \mid |\langle l, x\rangle| \le 1, \forall x \in \mathcal{C}\}$. At each round t, the agent picks an action $A_t \in \mathcal{C}$. At the same time, an adversary picks a loss vector $l_t \in \overline{\mathcal{C}}$ and a delay d_t. Then the agent suffers a loss of $\widehat{l}_t = \langle l_t, A_t\rangle$ which is revealed to the agent at the end of the $t + d_t$-th round. The objective of the agent is still to minimize the pseudo-regret, now defined by

$$\mathfrak{R}_T \triangleq \sup_{y\in\mathcal{C}} \mathbb{E}\left[\sum_{t=1}^T \langle l_t, A_t - y\rangle\right].$$

OMD, as in many other problems, has been widely adopted in non-delayed linear bandit problems (Abernethy et al., 2008; Bubeck et al., 2012; Bubeck and Eldan, 2015). In particular, Abernethy et al. (2008) proposed an OMD-based algorithm BOLO for adversarial linear bandits. It uses an $\mathcal{O}(n)$-self-concordant barrier (see Definition E.3 for an exact definition) together with a sampling scheme supported on the Dikin ellipsoid (which we will introduce in Algorithm 5). This algorithm achieves $\mathcal{O}(n^{3/2}\sqrt{T\log T})$ regret.

Now, we illustrate how to easily convert the BOLO algorithm into a delay-robust version via our Banker-OMD framework. We first pick an $\mathcal{O}(n)$-self-concordant Ψ and set $x_0 = \nabla\Psi^*(\mathbf{0})$; according to Abernethy et al. (2008), this ensures (Ψ, x_0) to be $(\mathcal{O}(n\log T), \mathcal{O}(n^2))$-regular under the sampling scheme of BOLO. After that, we pick the action scales according to Theorem 4.6 – resulting in the Banker-BOLO algorithm (presented as Algorithm 5 in Appendix E.1). Our Banker-BOLO algorithm then works in delayed adversarial linear bandits and ensures the following:

Theorem 7.1. *When Ψ is an $\mathcal{O}(n)$-self-concordant barrier over $\mathcal{C}$, Banker-BOLO in Algorithm 5 ensures*

$$\mathfrak{R}_T = \widetilde{\mathcal{O}}\left(n^{3/2}\sqrt{T} + n^2\sqrt{D}\right),$$

which is only $\widetilde{\mathcal{O}}(n^2\sqrt{D})$ more than the non-delayed regret $\widetilde{\mathcal{O}}(n^{3/2}\sqrt{T})$ achieved by BOLO (Abernethy et al., 2008).

To our knowledge, Banker-BOLO is the first algorithm achieving $\widetilde{\mathcal{O}}(\text{poly}(n)(\sqrt{T} + \sqrt{D}))$ regret for adversarial linear bandits with arbitrary delays.[6] The previous feedback delay model on adversarial linear bandits only allows *constant* delay lengths (Ito et al., 2020); moreover, this constant length needs to be revealed to the agent *beforehand*. Therefore, our setting is far more general than the previous one. And – more importantly – such achievements are attained by simply plugging a non-delayed adversarial linear bandit algorithm (BOLO) into our Banker-OMD framework.

8. Conclusion and Future Directions

In this paper, we propose the Banker-OMD framework for bandit learning problems with feedback delays. It almost decouples feedback delay handling and the task-specific OMD algorithm design. We illustrate the power of our framework by studying two different problems — we achieve the first near-optimal performance in both settings. Our result also extends to non-delayed scale-free adversarial MABs.

However, the most important contribution of our work is to illustrate the power of Banker-OMD in decoupling feedback delays from the algorithm design (i.e., deciding the regularizer Ψ and action scale σ_t's). Using our framework, algorithms for non-delayed problems like Tsallis-INF (Zimmert and Seldin, 2019) or BOLO (Abernethy et al., 2008) can be easily tuned to handle delays. Thus, we expect our framework to be used in many other delayed bandit learning problems.

Moreover, Banker-OMD allows the learning scales to be non-monotonic (see Algorithm 2), which is more flexible than vanilla OMD even in non-delayed settings. Our framework is also capable of handling arm-dependent delays (see Appendix A.3). We leave further investigations to the future.

Besides regret, the space complexity of Banker-OMD can also be improved: The current version requires memorizing the actions for all rounds with absent feedback (which can be as large as $\mathcal{O}(\sqrt{DK})$), while the algorithms specially designed for delayed MABs only need $\mathcal{O}(K)$ space (Zimmert and Seldin, 2020). See Appendix A.5 for more discussions.

Acknowledgements

This work is supported by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108403 and the Tsinghua Precision Medicine Foundation 10001020109. We want to thank the anonymous reviewers for their insightful comments.

[6] Meanwhile, the lower bound is still under-explored. The only lower bound for delayed adversarial linear bandits is $\Omega(n\sqrt{T} + \sqrt{ndT})$ under the uniform and known delay assumption (Ito et al., 2020). As our setting is more general, $(\sqrt{T} + \sqrt{D})$ factors are unavoidable, though the optimal dependency on n remains unknown.

References

Jacob Abernethy, Elad E Hazan, and Alexander Rakhlin. 2008. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*. 263–273.

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. 2015. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems* 28 (2015), 2197–2205.

Alekh Agarwal and John C Duchi. 2011. Distributed delayed stochastic optimization. *Advances in neural information processing systems* 24 (2011).

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. 2017. Corralling a band of bandit algorithms. In *Conference on Learning Theory*. PMLR, 12–38.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. 2014. Regret in online combinatorial optimization. *Mathematics of Operations Research* 39, 1 (2014), 31–45.

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

Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. 2002a. Adaptive and self-confident on-line learning algorithms. *J. Comput. System Sci.* 64, 1 (2002), 48–75.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. 2019. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*. 11349–11358.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. 2012. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 41–1.

Sébastien Bubeck and Ronen Eldan. 2015. The entropic barrier: a simple and optimal universal self-concordant barrier. In *Conference on Learning Theory*. PMLR, 279–279.

Nicoló Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. 2016. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*. PMLR, 605–622.

Lixing Chen and Jie Xu. 2019. Task replication for vehicular cloud: Contextual combinatorial bandit with delayed feedback. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 748–756.

Yan Dai, Haipeng Luo, and Liyu Chen. 2022. Follow-the-Perturbed-Leader for Adversarial Markov Decision Processes with Bandit Feedback. *Advances in Neural Information Processing Systems* 35 (2022).

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. 2014. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research* 15, 1 (2014), 1281–1316.

Thomas Desautels, Andreas Krause, and Joel W Burdick. 2014. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research* 15 (2014), 3873–3923.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. 2011. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. 169–178.

Geoffrey J Gordon. 1999. Regret bounds for prediction problems. In *Proceedings of the twelfth annual conference on Computational learning theory*. 29–40.

Andras Gyorgy and Pooria Joulani. 2021. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*. PMLR, 3988–3997.

Hédi Hadiji and Gilles Stoltz. 2023. Adaptation to the Range in K–Armed Bandits. *Journal of Machine Learning Research* 24, 13 (2023), 1–33.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. 2021. Delayed Feedback in Episodic Reinforcement Learning. *arXiv preprint arXiv:2111.07615* (2021).

Jiatai Huang and Longbo Huang. 2021. Robust Wireless Scheduling under Arbitrary Channel Dynamics and Feedback Delay. In *2021 33th International Teletraffic Congress (ITC-33)*. IEEE, 1–8.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. 2020. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems* 33 (2020).

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. 2020. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*. PMLR, 4860–4869.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. 2022. Near-Optimal Regret for Adversarial MDP with Delayed Bandit Feedback.

Advances in Neural Information Processing Systems 35 (2022).

Tiancheng Jin and Haipeng Luo. 2020. Simultaneously Learning Stochastic and Adversarial Episodic MDPs with Known Transition. *Advances in Neural Information Processing Systems* 33 (2020).

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. 2013. Online learning under delayed feedback. In *International Conference on Machine Learning*. 1453–1461.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. 2022. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7281–7289.

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

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.

Chris Mesterharm. 2005. On-line learning with delayed label feedback. In *International Conference on Algorithmic Learning Theory*. Springer, 399–413.

Arkadi Nemirovski. 1979. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody* 15, 1 (1979).

Arkadi Nemirovski. 2004. Interior point polynomial time methods in convex programming. *Lecture notes* (2004).

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. 1983. Problem complexity and method efficiency in optimization. (1983).

Yurii Nesterov and Arkadii Nemirovskii. 1994. *Interior-point polynomial algorithms in convex programming*. SIAM.

Gergely Neu. 2015. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*. PMLR, 1360–1375.

Francesco Orabona and Dávid Pál. 2018. Scale-free online learning. *Theoretical Computer Science* 716 (2018), 50–69.

Laurent Orseau and Marcus Hutter. 2021. Isotuning With Applications To Scale-Free Online Learning. *arXiv preprint arXiv:2112.14586* (2021).

Sudeep Raja Putta and Shipra Agrawal. 2022. Scale-Free Adversarial Multi Armed Bandits. In *International Conference on Algorithmic Learning Theory*. PMLR, 910–930.

Ralph Tyrell Rockafellar. 2015. *Convex analysis*. Princeton university press.

Shai Shalev-Shwartz and Yoram Singer. 2007a. Online learning: Theory, algorithms, and applications. (2007).

Shai Shalev-Shwartz and Yoram Singer. 2007b. A primal-dual perspective of online learning algorithms. *Machine Learning* 69, 2 (2007), 115–142.

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

Dirk Van Der Hoeven and Nicolò Cesa-Bianchi. 2022. Nonstochastic Bandits and Experts with Arm-Dependent Delays. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. 2020. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*. PMLR, 9712–9721.

Manfred K Warmuth and Arun K Jagota. 1997. Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *Electronic proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics*, Vol. 326. Citeseer.

James MS Wason and Lorenzo Trippa. 2014. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in medicine* 33, 13 (2014), 2206–2221.

Chen-Yu Wei and Haipeng Luo. 2018. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*. PMLR, 1263–1291.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. 2019. Learning in Generalized Linear Contextual Bandits with Stochastic Delays. In *Advances in Neural Information Processing Systems*, Vol. 32.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. 2019. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*. PMLR, 7683–7692.

Julian Zimmert and Yevgeny Seldin. 2019. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 467–475.

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

Supplementary Materials

A. More Discussions

A.1. More Related Works

Online Mirror Descent. Online mirror descent is a concept that originated in classical optimization and first developed by Nemirovski (1979) and Nemirovskij and Yudin (1983). It was introduced to the context of online learning no later than Warmuth and Jagota (1997); Gordon (1999); Shalev-Shwartz and Singer (2007a;b). Since then, OMD has been used in various online learning tasks with adversarial reward, such as MABs (Audibert et al., 2014; Abernethy et al., 2015; Wei and Luo, 2018; Zimmert and Seldin, 2019), semi-bandits (Audibert et al., 2014; Zimmert et al., 2019), linear bandits (Abernethy et al., 2008; Bubeck et al., 2012; Audibert et al., 2014) and Markov Decision Processes (MDP) (Jin et al., 2020; Jin and Luo, 2020). For more discussions, one may refer to the references therein.

Delayed MAB. Joulani et al. (2013) first studied adapting existing stochastic MAB algorithm to delayed settings with i.i.d. delays. Cesa-Bianchi et al. (2016) studied adversarial MABs with uniform delays d and derived a regret lower bound of

Table 1. Overview of Our Algorithms and Some Related Works

Algorithm	Setting	Regret Upper bound
Zimmert and Seldin (2020)	Delayed Adversarial MAB $\Omega(\sqrt{KT}+\sqrt{D\log K})$ (Cesa-Bianchi et al., 2016)	$\widetilde{\mathcal{O}}(\sqrt{KT}+\sqrt{D})$
`Banker-TINF` (**Corollary 5.1**)		$\widetilde{\mathcal{O}}(\sqrt{K(D+T)})$
`Banker-SFTINF` (**Algorithm 3**)	Scale-free Delayed Adversarial MAB $\Omega((\sqrt{KT}+\sqrt{D\log K})L)$ (Cesa-Bianchi et al., 2016)	$\widetilde{\mathcal{O}}(\sqrt{K(D+T)}L)$ for non-negative losses
`Banker-SFLBINF` (**Algorithm 4**)		$\widetilde{\mathcal{O}}\left(\sqrt{K\,\mathbb{E}[\widetilde{\mathfrak{L}}_T^2]}+\sqrt{KDL}\right)=\widetilde{\mathcal{O}}(\sqrt{K(D+T)}L)$
	Scale-free Adversarial MAB $\Omega(\sqrt{KT}L)$ (Auer et al., 2002b)	$\widetilde{\mathcal{O}}\left(\sqrt{K\,\mathbb{E}[\sum_{t=1}^T l_{t,A_t}^2]}\right)=\widetilde{\mathcal{O}}(\sqrt{KT}L)$
Putta and Agrawal (2022)		$\widetilde{\mathcal{O}}\left(\sqrt{K\sum_{t=1}^T\lVert l_t\rVert_2^2}+L\sqrt{KT}\right)=\widetilde{\mathcal{O}}(K\sqrt{T}L)$
`Banker-BOLO` (**Algorithm 5**)	Delayed Adversarial Linear Bandit $\Omega(n\sqrt{T}+\sqrt{nD})$ (Ito et al., 2020)[a]	$\widetilde{\mathcal{O}}(n^{3/2}\sqrt{T}+n^2\sqrt{D})$
Ito et al. (2020)		$\widetilde{\mathcal{O}}(n\sqrt{T}+\sqrt{n}\sqrt{dT})$ with known uniform delays d

[a] Precisely, Ito et al. (2020) gave an $\Omega(n\sqrt{T}+\sqrt{ndT})$ bound for the uniform delay case where the delay d is known before-hand.

$\Omega(\max\{\sqrt{KT},\sqrt{dT\log K}\})$. Recent works (Bistritz et al., 2019; Thune et al., 2019; Zimmert and Seldin, 2020) show that the total overhead due to feedback delays can be controlled in $\mathcal{O}(\sqrt{D\log K})$ (where D is the total delay, which is dT if the delays are uniform) by introducing a moving negative entropy component in the OMD regularizer. Remarkably, all these works assumed $[0,1]$-bounded losses, i.e., $l_{t,a}\in[0,1],\forall t\in[T],a\in[K]$. Thus, our delayed scale-free MAB setting extends existing results to allow general losses.

Delayed Linear Bandits. Zhou et al. (2019) first studied stochastic linear bandits with i.i.d. delays. Vernade et al. (2020) then studied stochastic linear bandits with random but only partially observable delays. Ito et al. (2020) studied adversarial linear bandits with uniform delay d, proposing an $\widetilde{\mathcal{O}}(\sqrt{n(n+d)T})$-regret algorithm and showing a near-matching $\Omega(\sqrt{n(n+d)T})$ lower bound.

Learning with Feedback Delays. Aside from aforementioned delayed MABs and delayed linear bandits, feedback delays are also considered in stochastic Markov Decision Processes (MDPs), see, e.g., (Lancewicki et al., 2022; Howson et al., 2021; Jin et al., 2022; Dai et al., 2022). General online optimizations with feedback delays and full information (Mesterharm, 2005; Agarwal and Duchi, 2011) or bandit feedback (Dudik et al., 2011; Desautels et al., 2014) are also extensively studied in the literature. Finally, feedback delays in wireless network optimizations are recently handled by Huang and Huang (2021), via the application of a preliminary version of our `Banker-OMD` framework.

Scale-free Learning. Scale-free settings were first studied in full-information case (e.g., (De Rooij et al., 2014; Orabona and Pál, 2018)). The most recent work in this line (Orseau and Hutter, 2021) studies non-delayed full-information scale-free online learning. For MABs, Hadiji and Stoltz (2023) proposed an $\mathcal{O}(L_\infty\sqrt{KT\log K})$ algorithm for stochastic ones. Putta and Agrawal (2022) then considered the more challenging adversarial case, yielding two adaptive bounds, $\widetilde{\mathcal{O}}(\sqrt{KL_2}+L_\infty\sqrt{KT})$ and $\widetilde{\mathcal{O}}(\sqrt{KL_2}+L_\infty\sqrt{KL_1})$, which both become $\widetilde{\mathcal{O}}(K\sqrt{T}L_\infty)$ in the worst case. Here, L_p denotes the p-norm of the flattened loss matrix, i.e., $L_p=(\sum_{t=1}^T\sum_{i=1}^K l_{t,i}^p)^{1/p}$. Importantly, no previous work considered delayed feedback in scale-free MABs, and we give the first attempt to solve this problem.

A.2. Why Not Use Existing Works to Solve the Delayed Scale-free Adversarial MAB Problem?

In this section, we explain why we are not satisfied with simply using existing works (e.g., (Zimmert and Seldin, 2020)) to solve our delayed scale-free adversarial MAB setting (via techniques like doubling trick). The main reason is that handling delays and handling scale-free losses are actually coherent — a largely delayed large loss can make the loss estimation inaccurate for a long phase. As a result, to obtain an analysis for the modified algorithm, it still needs to resolve a few challenging technical difficulties.

For example, consider the representative work by Zimmert and Seldin (2020). This work introduces the notation $\widehat{L}_{t,i_t}^{\text{miss}}$ to denote the sum of estimated losses of the outstanding prior steps. Originally, their analysis upper-bounds the expectation of $\widehat{L}_{t,i_t}^{\text{miss}}$ by $\mathfrak{d}_t$. As a result, the dependence of D in final regret bound comes from summing up $\widehat{L}_{t,i_t}^{\text{miss}}$ and then taking expectation. While this argument works well when $L=1$, such calculation is no longer valid if they get divided by $\widehat{L}_t$'s, as a heavily delayed large loss will affect up to $\Theta(D)$ subsequent $\widehat{L}_{t,i_t}^{\text{miss}}$'s. Moreover, when there may be negative losses, $\widehat{L}_{t,i_t}^{\text{miss}}$ may contain negative components, which means the step (f) in the proof of Lemma 3 (which uses second-order remainder upper bounds of Bergman divergences) no longer holds.

Therefore, as resolving such issues are challenging and does not provide generality and flexibility on other tasks (like delayed adversarial linear bandits), we instead propose a new framework of handling delayed bandit learning problems — as we illustrated in the main text, our framework allows various nice properties and solves many different problems (most of which are open problems in the literature). That's why we do not directly adopt existing delayed MAB papers.

A.3. Arm-Dependent Delays and `Banker-OMD`

Throughout the paper, we discussed about feedback delays that only depends on time indices (i.e., they are irrelevant to the actions actually taken). This model, namely arm-independent delay model, is the most common one in the literature (Thune et al., 2019; Bistritz et al., 2019; Zimmert and Seldin, 2020; Gyorgy and Joulani, 2021). However, in realistic applications, such an assumption is often too restrictive. For example, in medical experiments, different medicine may take different time to show its effect.

To resolve this issue, Van Der Hoeven and Cesa-Bianchi (2022) recently proposed a more general setting called *arm-dependent delays*, where there is a delay matrix $\delta \in \mathbb{N}^{T\times K}$ (also chosen by an oblivious adversary) to give out delay lengths for each (round, arm) *combination*. Formally, the actual delay at time t is determined by $d_t \leftarrow \delta_{t,A_t}$.

Van Der Hoeven and Cesa-Bianchi (2022) pointed out that existing adversarial MAB works on *arm-indepdendent* delays, e.g., (Thune et al., 2019; Zimmert and Seldin, 2020) are *not capable* of dealing with the new setting. They also proposed a new algorithm guaranteeing

$$\mathfrak{R}_T \le \sqrt{KT} + \sqrt{\ln K \cdot \mathbb{E}\left[\sum_{t=1}^T \langle x_t, \rho_t\rangle\right]} + \rho^*, \tag{14}$$

where ρ_t is a K-dimensional vector such that $\rho_{t,a}$ denotes the number of feedback of action a that is still on the way at the beginning of round t, i.e., $\rho_{t,a} \triangleq \sum_{s=1}^{t-1} \mathbb{1}_{A_s=a, s+\delta_{s,a}\ge t}$. Moreover, $\rho^* \triangleq \max_{t\in[T],a\in[K]} \rho_{t,a}$ denotes the maximum number of missing observations of some action.

Although our `Banker-OMD` framework is presented with the assumption of arm-independent delays, we claim here that our results also apply to arm-dependently delayed settings as long as we redefine the symbol D as the sum of all delay lengths *actually experienced*, namely $D \triangleq \sum_{t=1}^T \delta_{t,A_t}$. In other words, D becomes a random variable rather than a obliviously chosen constant — therefore, we can attain a regret bound in terms of $\mathbb{E}[D]$, recovering Eq. (14).

Technically speaking, the reason for the difference between our result and the previous ones (Thune et al., 2019; Zimmert and Seldin, 2020) is that, rather than directly bound the regret (which involves expectation), our core result Theorem 4.2 upper bounds the *sample path* loss difference $\sum_{t=1}^T \langle \widetilde{l}_t, x_t - y\rangle$. When we bound the total investment namely B_T, we also derive a sample path upper bound for B_T: As we only make use of bounds like $\sum_{i\le n} 1/i = \Theta(\log n)$ and monotonicity of functions like $x/\ln x$, whether D is a random variable does not affect this bound of B_T. Moreover, as for each immediate cost term $\sigma_t D_\Psi(x_t, \widetilde{z}_t)$, we either directly derive a sample path upper bound (e.g., for log-barrier Ψ, Lemma D.2), or derive a bound for the conditional expectation $\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t) \mid \mathcal{F}_{t-1}]$, we can always obtain a $\mathcal{O}(\sigma_t^{-1})$ bound given appropriate regularity conditions of Ψ. We then sum up these bounds and then apply summation lemmas to get bounds in D. Again, until now, whether D is a random variable does not matter. At this point, we have derived an upper bound for $\sum_{t=1}^T \langle \widetilde{l}_t, x_t - y\rangle$ in D. Luckily, this bound is a concave function in D (as it only involves square roots and logarithms). Taking expectations on both sides, we can then get a regret bound in $\mathbb{E}[D]$, just like Eq. (14).

A.4. Technical Comparison with (Putta and Agrawal, 2022)

Compared to our work, Putta and Agrawal (2022) did not explicitly estimate the actual loss range L (whereas we did by introducing $\widehat{L}_t$'s). Hence, instead of our tuning $\sigma_t \propto (1 + \sum_{s=1}^{t-1} l_{s,A_s}^2 + \widehat{L}_t^2)^{1/2} \approx (1 + \sum_{s=1}^{t} l_{s,A_s}^2)^{1/2}$, they can only set $\sigma_t \propto (1 + \sum_{s=1}^{t-1} l_{s,A_s}^2)^{1/2}$ – thus, they can only use the following summation lemma different from Lemma B.1:

$$\sum_{i=1}^n \frac{x_i}{\sqrt{1+\sum_{j=1}^{i-1} x_j}} \le \mathcal{O}\left(\sqrt{1+\sum_{i=1}^n x_i} + \max_{1\le i\le n} x_i\right).$$

The maximum of x_i causes an extra term that is proportional to the maximum reciprocal of probabilities to pull each arm in their regret bound. Consequently, their algorithm must include *explicit uniform exploration*, i.e., playing $(1-\gamma_t)x_t + \frac{\gamma_t}{K}$ instead of x_t for the t-th step. Even for non-delayed settings, $\gamma_t = \Omega(t^{-1/2})$ is needed for $\widetilde{\mathcal{O}}(\sqrt{TL})$ style total regret. Therefore, if one directly adapts their technique to delayed settings, the regret bound would involve a term proportional to

$$\max_{1\le t\le T} \sum_{s\le t: a_s = \text{“missing” at } t} x_{s,A_s}^{-1},$$

which could be as large as $\Theta(\sqrt{DT})$ in the worst case, making their framework incapable to feedback delays.

In a nutshell, there are essentially two improvements in our delayed scale-free MAB results: the first one is due to the *novel learning rates* (which improves the regret in non-delayed settings), and the second one is due to *our proposed Banker-OMD framework* (which easily generalizes the non-delayed regret bounds to delay-robust ones).

A.5. Computational and Space Complexity of `Banker-OMD`

Computational Complexity. Notice that the loop in Lines 8-9 of Algorithm 2 is to greedily spend up previous savings to meet the desired new action scale σ_t. Thus, by maintaining a linked list of all “non-exhausted” savings, the loop only needs to scan amortized $\mathcal{O}(1)$ previous time indices to decide each new x_t. Hence the algorithm displayed in Algorithm 2 can be implemented with an amortized time complexity of $\mathcal{O}(K)$ per action – which matches vanilla OMD.

Space Complexity. While the computational complexity of Algorithm 2 is the same as vanilla OMD, the space complexity is much larger as we need to memorize all previous actions. Fortunately, we can also adopt a slightly different decision rule of $\sigma_{t,s}$ and b_t (which is not included in the current version for ease of presentation), which further reduces the space complexity to $\mathcal{O}(\sqrt{D}K)$ while maintaining a per-step $\mathcal{O}(K)$ amortized time complexity:

1. Calculate the sum of all available savings $v = \sum_{s<t, a_s = \text{arrived}} v_s$.
2. If $v < \sigma_t$, fill up the remaining proportion by $b_t = \sigma_t - v$. Otherwise, let $b_t = 0$. This step ensures that Lemma 4.5 (the upper bound on the total investment B_T) still holds.
3. Finally, for each $s < t$ where $a_s =$ arrived, we set $\sigma_{t,s} = \min(v, \sigma_t)\frac{v_s}{v}$. In other words, we set $\sigma_{t,s} \propto v_s$ such that $\sum_{s<t, a_s = \text{arrived}} \sigma_{t,s} = \min(v,\sigma_t)$, which ensures Eq. (10).

Under the new policy, x_t in Line 10 of Algorithm 2 becomes

$$x_t = \nabla\overline{\Psi}^*\left(\frac{\min(v,\sigma_t)}{\sigma_t} \sum_{s<t, a_s = \text{arrived}} \frac{v_s}{v}\nabla\Psi(z_s) + \frac{b_t}{\sigma_t}\nabla\Psi(x_0)\right).$$

Therefore, by tracking $\sum_{s<t, a_s = \text{arrived}} \frac{v_s}{v}\nabla\Psi(z_s)$, we can obtain x_t by applying the mirror map $\overline{\Psi}^*(\cdot)$ to the current $\sum_{s<t, a_s = \text{arrived}} \frac{v_s}{v}\nabla\Psi(z_s)$. Because we always have $v_s \propto \sigma_s$, each update does not involve calculating the summation again and thus can be done in $\mathcal{O}(K)$ time – again the same as vanilla OMD. Moreover, in this version, we only need to memorize the actions whose feedback is absent, giving $\mathcal{O}(\sqrt{D}K)$ space consumption.

B. Technical Details for Online Mirror Descent and `Banker-OMD`

Throughout the paper, we use

$$\mathcal{F}_t = \sigma\left(A_1, \ldots, A_t, l_{1,A_1}\mathbb{1}[d_1 + 1 \le t], \ldots, l_{t,A_t}\mathbb{1}[d_t + t \le t]\right)$$

to denote the filtration of σ-algebra when studying random quantities indexed by time. Below, we present a summation lemma that we heavily use when tuning the learning scales.

Lemma B.1 (Summation Lemma (Auer et al., 2002a, Lemma 3.5))**.** *The following holds for any non-negative $x_1, x_2, \ldots, x_T$'s:*

$$\sum_{t=1}^T \frac{x_t}{\sqrt{1+\sum_{s=1}^t x_s}} \le 2\sqrt{1+\sum_{t=1}^T x_t}.$$

B.1. Legendre Functions and Their Properties

Definition B.2 (Legendre Functions)**.** For a strictly convex f defined on a convex domain $A \subseteq \mathbb{R}^K$, we say f is *Legendre* if (i) $int(A)$ is non-empty, (ii) f is differentiable on $int(A)$, and (iii) $\lim_{n\to\infty} \|\nabla f(x_n)\|_2 \to \infty$ for any sequence $(x_n)_{n=1}^\infty$ in $int(A)$ converging to some $x \in \partial\, int(A)$.

The proof of the following lemmas can be found in any textbook for convex analysis, e.g., (Rockafellar, 2015).

Lemma B.3. *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex set, $f : \mathcal{C} \to \mathbb{R}$ be a Legendre function. Then,*

1. *∇f is a bijection between $int(\mathcal{C})$ and $int(dom(f^*))$ with the inverse $(\nabla f)^{-1} = \nabla f^*$;*
2. *$D_f(y, x) = D_{f^*}(\nabla f(x), \nabla f(y))$ for all $x, y \in int(\mathcal{C})$;b*
3. *the convex conjugate f^* is Legendre.*

We now give a formal proof to the single-step OMD regret bound Eq. (1), stated in the following lemma.

Lemma B.4. *For any $\sigma > 0$, $x, y \in \triangle^{[K]}$, $l \in \mathbb{R}_+^K$ and Legendre function $\Psi : \mathbb{R}_+^K \to \mathbb{R}$, we have*

$$\langle l, x - y\rangle \le \sigma D_\Psi(y, x) - \sigma D_\Psi(y, z) + \sigma D_\Psi(x, \widetilde{z}) \tag{15}$$

where

$$z = \operatorname*{arg\,min}_{x' \in \triangle^{[K]}} \langle l, x'\rangle + \sigma D_\Psi(x', x), \quad \widetilde{z} = \operatorname*{arg\,min}_{x' \in \mathbb{R}_+^K} \langle l, x'\rangle + \sigma D_\Psi(x', x), \tag{16}$$

or equivalently,

$$z = \nabla\overline{\Psi}^*(\nabla\Psi(x) - \frac{1}{\sigma} l), \quad \widetilde{z} = \nabla\Psi^*(\nabla\Psi(x) - \frac{1}{\sigma} l). \tag{17}$$

Proof. Ψ is Legendre hence $\nabla\Psi$ explodes on $\partial\mathbb{R}_+^K$, which guarantees that the minimizer x' in the definition of $\widetilde{z}$ in Eq. (16) will lie in $int(\mathbb{R}_+^K)$ and $\frac{\partial}{\partial x'}[\langle l, x'\rangle + \sigma D_\Psi(x', x)] = 0$. The bijection property in Lemma B.3 then asserts this $\arg\min$ definition is equivalent to the definition in Eq. (17) using mirror maps $\nabla\Psi$ and $\nabla\Psi^*$. Since $\overline{\Psi}$ is a Legendre function on $\triangle^{[K]}$, similar argument suggests that the definitions for z in Eq. (16) and Eq. (17) are equivalent.

The definition of $\widetilde{z}$ in Eq. (17) implies $l = \sigma(\nabla\Psi(x) - \nabla\Psi(\widetilde{z}))$. The first order optimality condition of z in Eq. (16) implies that $\langle \frac{1}{\sigma} l + \nabla\Psi(z) - \nabla\Psi(x), y - z\rangle \ge 0$ for any $y \in \triangle^{[K]}$. Hence we have

$$\begin{aligned}
\langle l, x - y\rangle &= \langle l, x - z\rangle + \langle l, z - y\rangle \\
&\le \sigma\langle \nabla\Psi(x) - \nabla\Psi(\widetilde{z}), x - z\rangle + \sigma\langle \nabla\Psi(z) - \nabla\Psi(x), y - z\rangle \\
&\overset{(a)}{=} \sigma(D_\Psi(z, x) + D_\Psi(x, \widetilde{z}) - D_\Psi(z, \widetilde{z})) - \sigma(D_\Psi(y, z) + D_\Psi(z, x) - D_\Psi(y, x)) \\
&= \sigma D_\Psi(y, x) - \sigma D_\Psi(y, z) + \sigma D_\Psi(x, \widetilde{z}) - \sigma D_\Psi(z, \widetilde{z}) \\
&\le \sigma D_\Psi(y, x) - \sigma D_\Psi(y, z) + \sigma D_\Psi(x, \widetilde{z})
\end{aligned}$$

where (a) uses the following "three-point identity" of Bregman divergences:

$$D_\Psi(a, b) + D_\Psi(b, c) - D_\Psi(a, c) = \langle \nabla\Psi(c) - \nabla\Psi(b), a - b\rangle.$$

□

B.2. Detailed Proofs in Section 4

First of all, we study when the immediate cost terms $\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)]$ in Algorithms 1 and 2 can be uniformly bounded:

Lemma B.5. *If $\Psi(x) = \sum_{i=1}^K f(x_i)$ where $f''(x) = x^{-\alpha}$, $\alpha \ge 1$, then in Algorithms 1 and 2, for any $t \ge 1$ we have*

$$\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t) \mid \mathcal{F}_{t-1}] \le \frac{K}{2\sigma_t}.$$

Proof. In fact, for any choice of Ψ and $t \ge 1$ we have

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) = \sigma_t D_{\Psi^*}(\nabla\Psi(\widetilde{z}_t), \nabla\Psi(x_t))$$

$$
\begin{aligned}
&= \sigma_t D_{\Psi^*}(\nabla\Psi(x_t) - \frac{\widetilde{l}_t}{\sigma_t}, \nabla\Psi(x_t)) \\
&= \sigma_t \left(Psi^*(\nabla\Psi(x_t) - \frac{\widetilde{l}_t}{\sigma_t}) - \Psi^*(\nabla\Psi(x_t)) - \langle x_t, -\frac{\widetilde{l}_t}{\sigma_t}\rangle \right) \\
&= \frac{\|\widetilde{l}_t\|^2_{\nabla^2\Psi^*(\theta_t)}}{2\sigma_t},
\end{aligned} \tag{18}
$$

where in the last step we write the Bregman divergence into a second order Lagrange remainder, θ_t is some element inside the line segment connecting $\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t$ and $\nabla\Psi(x_t)$.

Note that under this particular $\Psi(x) = \sum_{i=1}^K f(x_i)$, we have

- $\nabla^2\Psi^*(\cdot)$ is diagonal,
- The diagonal elements of $\nabla^2\Psi^*(\cdot)$ are non-decreasing on the line segment $[\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t, \nabla\Psi(x_t)]$,

and we can further upper bound Eq. (18) by $\frac{1}{2\sigma_t}\|\widetilde{l}_t\|^2_{\nabla^2\Psi^*(\nabla\Psi(x_t))} = \frac{1}{2\sigma_t}\|\widetilde{l}_t\|^2_{\nabla^2\Psi(x_t)^{-1}}$. Therefore,

$$
\begin{aligned}
\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t) \mid \mathcal{F}_{t-1}] &\le \mathbb{E}[\frac{1}{2\sigma_t}\|\widetilde{l}_t\|^2_{\nabla^2\Psi(x_t)^{-1}} \mid \mathcal{F}_{t-1}] \\
&= \frac{1}{2\sigma_t}\sum_{i=1}^K x_{t,i}^\alpha\, \mathbb{E}[\widetilde{l}_{t,i}^2 \mid \mathcal{F}_{t-1}] \\
&= \frac{1}{2\sigma_t}\sum_{i=1}^K x_{t,i}^\alpha\, \mathbb{E}\left[\frac{l_{t,A_t}^2 \mathbb{1}[A_t = i]}{x_{t,i}^2} \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \frac{1}{2\sigma_t}\sum_{i=1}^K x_{t,i}^{\alpha-1} \\
&\le \frac{K}{2\sigma_t}.
\end{aligned}
$$

□

B.2.1. Proof of Lemma 4.1

We give a formal proof for the "convex combination in the dual space" lemma of Banker-OMD (Lemma 4.1) here.

Proof of Lemma 4.1. Let $\widetilde{x} = \nabla\Psi^*(\sum_{i=1}^h \frac{\sigma_i}{\sigma}\nabla\Psi(z_i))$, we have

$$
\begin{aligned}
\sigma D_\Psi(y, x) &\overset{(a)}{\le} D_\Psi(y, \widetilde{x}) \\
&\overset{(b)}{=} \sigma D_{\Psi^*}(\nabla\Psi(\widetilde{x}), \nabla\Psi(y)) \\
&= \sigma D_{\Psi^*}(\sum_{i=1}^h \frac{\sigma_i}{\sigma}\nabla\Psi(z_i), \nabla\Psi(y)) \\
&\overset{(c)}{\le} \sigma \cdot \sum_{i=1}^h \frac{\sigma_i}{\sigma} D_{\Psi^*}(\nabla\Psi(z_i), \nabla\Psi(y)) \\
&\overset{(d)}{=} \sum_{i=1}^h \sigma_i D_\Psi(y, z_i)
\end{aligned}
$$

where (a) is due to the Pythagorean theorem for Bregman divergences ($D_\Psi(y, \widetilde{x}) = D_\Psi(y, x) + D_\Psi(x, \widetilde{x}) \ge D_\Psi(y, x)$), (b) is due to the duality property of Bregman divergences, (c) is due to the convexity of the first argument of Bregman divergences, and (d) uses again the duality property. □

B.2.2. Proof of Theorem 4.2

It is then easy to see that the general regret bound for `Banker-OMD` (Theorem 4.2) is just a summation over the single-step regret bounds in Lemma 4.1. A formal proof is presented here.

Proof of Theorem 4.2. When $T = 0$ this bound trivially holds. Suppose inequality Eq. (11) holds for all $T \le M$, then the difference of Eq. (11)'s RHS between $T = M + 1$ and $T = M$ is at least

$$
\begin{aligned}
&(B_{T+1} - B_T) D_\Psi(y, x_0) + \sigma_{T+1} D_\Psi(x_{T+1}, \widetilde{z}_{T+1}) - \sum_{i=1}^{T} \triangle v_i D_\Psi(y, z_i) - \sigma_{T+1} D_\Psi(y, z_{T+1}) \\
&= (b_{T+1} D_\Psi(y, x_0) + \sum_{i=1}^{T} \sigma_{T+1,i} D_\Psi(y, z_i)) - \sigma_{T+1} D_\Psi(y, z_{T+1}) + \sigma_{T+1} D_\Psi(x_{T+1}, \widetilde{z}_{T+1}) \\
&\overset{(a)}{\ge} \sigma_{T+1} D_\Psi(y, x_{T+1}) - \sigma_{T+1} D_\Psi(y, z_{T+1}) + \sigma_{T+1} D_\Psi(x_{T+1}, \widetilde{z}_{T+1}) \\
&\overset{(b)}{\ge} \langle \widetilde{l}_{T+1}, x_{T+1} - y \rangle,
\end{aligned}
$$

which is the difference of Eq. (11)'s LHS between $T = M + 1$ and $T = M$. Here we use $\triangle v_i$ to denote the change of the variable v_i in Algorithm 2 from the end of time T to the end of time $T + 1$, (a) is due to Lemma 4.1, (b) is due to Lemma B.4. Thus by induction, we are done. □

B.2.3. Proof of Lemma 4.5

Proof of Lemma 4.5. Lemma 4.5 is a special case of the following fact: at the end of any time T, $B_T = \sum_{i=1}^{T} v_i$. This fact can be easily verified by an induction on T. If `GreedyPick` is used, when we encounter a new round T and observe $b_T > 0$, it is guaranteed that at that moment, any non-zero v_i is corresponding to an action whose feedback is still on the way (including the action A_T we are going to play). □

B.2.4. Proof of Theorem 4.6

The general method for $\mathcal{O}((\sqrt{T} + \sqrt{D \log D}) f(T))$ regret bounds (Theorem 4.6) can be proved following the same proof sketch we presented in the text to deal with $\sigma_t = (\frac{1}{\sqrt{t}} + \frac{\mathfrak{d}_t}{\sqrt{\mathfrak{D}_t}})^{-1}$. To be specific,

Proof of Theorem 4.6. When `Banker-OMD` uses `GreedyPick` and action scales $\sigma_t = \left(\frac{1}{\sqrt{t}} + \mathfrak{d}_t \sqrt{\frac{\ln(\mathfrak{D}_t + 1)}{\mathfrak{D}_t}} \right)^{-1}$, we have

$$
\begin{aligned}
\sum_{t=1}^{T} \sigma_t^{-1} &= \sum_{t=1}^{T} \left(\frac{1}{\sqrt{t}} + \mathfrak{d}_t \sqrt{\frac{\ln(\mathfrak{D}_t + 1)}{\mathfrak{D}_t}} \right) \\
&\le \sum_{t=1}^{T} \left(\frac{1}{\sqrt{t}} + \sqrt{\ln(D+1)} \frac{\mathfrak{d}_t}{\sqrt{\mathfrak{D}_t}} \right) \\
&= \mathcal{O}(\sqrt{T} + \sqrt{D \log D})
\end{aligned}
$$

where in the last step, we use the fact that $\mathfrak{D}_t$ is a cumulative sum of $\mathfrak{d}_1 \dots, \mathfrak{d}_t$ and apply the summation lemma Lemma B.1.

In order to bound B_T, consider the last moment T_0 at which B_t gets increased, and suppose at the beginning of T_0 there are $m = \mathfrak{d}_{T_0}$ feedback on the way, each corresponding to actions at time $t_1 < \cdots < t_m$ respectively. Then it is easy to see that $\mathfrak{d}_{t_i} \ge i - 1$ for any $1 \le i \le m$, for at the beginning of time t_i, actions invoked at time $t_1, \dots, t_{i-1}$ are still on the way. Furthermore, we have $D \ge \mathfrak{d}_{T_0} + \sum_{i=1}^{m} \mathfrak{d}_{t_i} = \binom{m+1}{2}$ hence $m = \mathcal{O}(\sqrt{D})$.

$\sigma_t = \left(\frac{1}{\sqrt{t}} + \mathfrak{d}_t \sqrt{\frac{\ln(\mathfrak{D}_t+1)}{\mathfrak{D}_t}} \right)^{-1}$ guarantees that $\sigma_t \le \sqrt{t}$ and $\sigma_t \le \frac{1}{\mathfrak{d}_t} \sqrt{\frac{\mathfrak{D}_t}{\ln \mathfrak{D}_t + 1}} = \mathcal{O}\left(\sqrt{\frac{D}{\log D}} \right) \frac{1}{\mathfrak{d}_t}$, where the last step is due to the monotonicity of $\sqrt{\frac{x}{\ln x+1}}$ when x is sufficiently large. According to Lemma 4.5, we have

$$
\begin{aligned}
B_T &= \sigma_{T_0} + \sum_{i=1}^{m} \sigma_{t_i} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{\log D}}\right)\frac{1}{\mathfrak{d}_{T_0}} + \sqrt{t_1} + \sum_{i=2}^{m} \mathcal{O}\left(\sqrt{\frac{D}{\log D}}\right)\frac{1}{\mathfrak{d}_{t_i}} \\
&\le \sqrt{T} + \mathcal{O}\left(\sqrt{\frac{D}{\log D}}\right)\left(\sum_{i=2}^{m}\frac{1}{i-1} + \frac{1}{m}\right) \\
&= \mathcal{O}(\sqrt{T} + \sqrt{D\log D}).
\end{aligned}
$$

□

C. Technical Details for `Banker-SFTINF`

By plugging Theorem 4.2 into Eq. (13), we get the following regret decomposition for Algorithm 3:

Lemma C.1. *Algorithm 3 guarantees*

$$
\mathfrak{R}_T \le \underbrace{B_T \sup_{y\in\triangle^{[K]}} D_\Psi(y, x_0))}_{\textit{Total investment}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)]}_{\textit{Immediate costs}} + \underbrace{\sum_{t=1}^{T} \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} l_{t,A_t}]}_{\textit{Skipping error}} \tag{19}
$$

where $\mathcal{E}_t$ denotes the event that the feedback of round t's action is marked to "skipped", σ_t, B_t, x_t, $\widetilde{z}_t$ are variables determined by Algorithm 3.

Below, we will bound the three terms in Eq. (19) within $\mathcal{O}(L\sqrt{K(D+T)\log(D+T)})$ one by one.

C.1. Bound for Total Investment Term

For the total investment term $B_T \sup_{y\in\triangle^{[K]}} D_\Psi(y, x_0))$, recall that we choose Ψ to be the $1/2$-Tsallis entropy function, which guarantees that $\sup_{y\in\triangle^{[K]}} D_\Psi(y, x_0) \le 2\sqrt{K}$. Thus it suffices to bound B_T. Applying Lemma 4.5, suppose the last round on which we make a new investment is T_0 and that time, there are m previous feedback corresponding to rounds $T_1, \dots, T_m$ still on the way, then we have

$$
\begin{aligned}
B_T &= \sum_{i=0}^{m} \sigma_{T_i} \\
&= \sum_{i=0}^{m} \frac{1}{\mathfrak{d}_{T_i}+1}\sqrt{\frac{3+D_{T_i}}{\ln(3+D_{T_i}/\widehat{L}^2_{T_i})}} \\
&\le \left(\sum_{i=0}^{m}\frac{1}{\mathfrak{d}_{T_i}+1}\right)\cdot \max_{1\le t\le T}\sqrt{\frac{3+D_t}{\ln(3+D_t/\widehat{L}^2_t)}} \\
&\le \left(\sum_{i=0}^{m}\frac{1}{i+1}\right)\cdot \max_{1\le t\le T}\widehat{L}_t\sqrt{\frac{3+D_t/\widehat{L}^2_t}{\ln(3+D_t/\widehat{L}^2_t)}} \\
&\overset{(a)}{\le} \mathcal{O}(\log m)\cdot\mathcal{O}\left(L\sqrt{\frac{D+T}{\ln(D+T)}}\right) \\
&\le \mathcal{O}(\log D)\cdot\mathcal{O}\left(L\sqrt{\frac{D+T}{\ln(D+T)}}\right) \\
&\le \mathcal{O}\left(L\sqrt{(D+T)\log D}\right)
\end{aligned}
$$

where in step (a) we simply control the $\widehat{L}_t$ factor out of the square root by L, and utilize the fact that $D_t/\widehat{L}_t \le t + \sum_{s=1}^{t} \mathfrak{d}_t$ for all t and the function $(3+x)/\ln(3+x)$ is increasing. Thus we have show the total investment is within $\mathcal{O}\left(L\sqrt{K(D+T)\log D}\right)$.

C.2. Bound for Immediate Costs

To deal with the immediate costs, we first show each immediate cost term $E[\sigma_t D_\Psi(x_t, \widetilde{z}_t)]$ can be bounded very similarly compared to the $[0,1]$-bounded loss case.

Lemma C.2. *Let Ψ be the $1/2$-Tsallis entropy function, $x_t \in \triangle^{[K]}$, A_t be an independent sample from $[K]$ according to x_t, $\widehat{l}_t \in \mathbb{R}_+^K$, $\widetilde{l}_t = \frac{\widehat{l}_{t,A_t}}{x_{t,A_t}} \mathbf{1}_{A_t}$. Let $\widetilde{z}_t \triangleq \nabla\Psi^*(\nabla\Psi(x_t) - \frac{\widetilde{l}_t}{\sigma_t})$ then we have*

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) \le \frac{\widehat{l}_{t,A_t}^2 x_{t,A_t}^{-1/2}}{\sigma_t}$$

and

$$\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)] \le \sqrt{K}\frac{\|\widehat{l}_t\|_\infty^2}{\sigma_t}.$$

Proof. If suffices to follow the calculation we used in the proof of Lemma B.5:

$$\begin{aligned}
\sigma_t D_\Psi(x_t, \widetilde{z}_t) &\overset{(a)}{=} \sigma_t D_{\Psi^*}(\nabla\Psi(\widetilde{z}_t), \nabla\Psi(x_t)) \\
&= \sigma_t D_{\Psi^*}(\nabla\Psi(x_t) - \frac{\widetilde{l}_t}{\sigma_t}, \nabla\Psi(x_t)) \\
&= \Psi^*(\nabla\Psi(x_t) - \frac{\widetilde{l}_t}{\sigma_t}) - \Psi^*(\nabla\Psi(x_t)) - \langle x_t, -\frac{\widetilde{l}_t}{\sigma_t}\rangle \\
&\overset{(b)}{=} \frac{\|\widetilde{l}_t\|_{\nabla^2\Psi^*(\theta_t)}^2}{2\sigma_t}
\end{aligned}$$

where (a) is due to the duality of Bregman divergences, and (b) regards the Bregman divergence as a second order Lagrange remainder, θ_t is some element inside the line segment connecting $\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t$ and $\nabla\Psi(x_t)$.

Here our Ψ is the $1/2$-Tsallis entropy function, then $\nabla^2\Psi^*(\cdot)$ is diagonal. We can verify that $\Psi_{ii}^{*\prime\prime}(\theta) = \Psi_{ii}''(\Psi_i'^{-1}(\theta))^{-1} = 2((-\theta)^{-2})^{3/2} = -2\theta^{-3}$ is increasing for all i and $\theta < 0$. Also notice that all components of $\widetilde{l}$ are non-negative, hence we have $\nabla^2\Psi^*(\theta_t) \preceq \nabla^2\Psi^*(\nabla\Psi(x_t)) = \nabla^2\Psi(x_t)^{-1}$, and therefore

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) \le \frac{\|\widetilde{l}_t\|_{\nabla^2\Psi(x_t)^{-1}}^2}{2\sigma_t} = \frac{\widetilde{l}_{t,A_t}^2 x_{t,A_t}^{3/2}}{\sigma_t} = \frac{\widehat{l}_{t,A_t}^2 x_{t,A_t}^{-1/2}}{\sigma_t} \le \frac{\|\widehat{l}_t\|_\infty^2 x_{t,A_t}^{-1/2}}{\sigma_t}.$$

Taking expectation on both sides of the inequality, we can further get

$$\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)] \le \frac{\|\widehat{l}_t\|_\infty^2}{\sigma_t} \sum_{i=1}^{K} x_{t,i}^{1/2} \le \sqrt{K}\frac{\|\widehat{l}_t\|_\infty^2}{\sigma_t}.$$

□

For the immediate costs term, the *after-skipping* estimator $\widetilde{l}_t$ used in `Banker-SFTINF` can be regarded as an importance sampling estimate for $\widehat{l}_t \triangleq l_t \cdot \mathbb{1}_{\{l_{t,A_t} \le \widehat{L}_t\}}$, which is a vector in $\mathbb{R}_+^K$, thus Lemma C.2 applies, giving that

$$\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t) \mid \mathcal{F}_{t-1}] \le \frac{\mathbb{E}[\widehat{l}_{t,A_t}^2 x_{t,A_t}^{-1/2}]}{\sigma_t} \le \frac{\widehat{L}_t^2\, \mathbb{E}[x_{t,A_t}^{-1/2}]}{\sigma_t} \le \sqrt{K}\frac{\widehat{L}_t^2}{\sigma_t}.$$

Therefore, it remains to bound $\sum_{t=1}^T \mathbb{E}[\frac{\widehat{L}_t^2}{\sigma_t}]$. We can write

$$\begin{aligned}\sum_{t=1}^T \frac{\widehat{L}_t^2}{\sigma_t} &= \sum_{t=1}^T \frac{(\mathfrak{d}_t+1)\widehat{L}_t^2}{\sqrt{3+D_t}}\sqrt{\ln(3+D_t/\widehat{L}_t^2)} \\ &\overset{(a)}{\le} \mathcal{O}(\sqrt{\log T})\sum_{t=1}^T \frac{(\mathfrak{d}_t+1)\widehat{L}_t^2}{\sqrt{3+D_t}} \\ &\overset{(b)}{\le} \mathcal{O}(\sqrt{\log T})\cdot\mathcal{O}(\sqrt{1+D_T}) \\ &\le \mathcal{O}(L\sqrt{(D+T)\log T}),\end{aligned}$$

where in step (a), we make use of the fact that $D_t/\widehat{L}_t^2 \le t + \sum_{s=1}^t \mathfrak{d}_s$ for all $1 \le t \le T$ to bound $\sqrt{\ln(3+D_t/\widehat{L}_t^2)}$ universally for all t by $\sqrt{\ln(3+D+T)} = \mathcal{O}(\log T)$; in step (b), recall that D_t is just the sum of all $(\mathfrak{d}_s+1)\widehat{L}_s^2$ for $s = 1, 2, \ldots, t$, thus we can apply the summation lemma Lemma B.1. Thus we have verified that the immediate costs are within $\mathcal{O}(L\sqrt{K(D+T)\log T})$.

C.3. Bound for Skipping Error

It remains to bound the skipping error term $\sum \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} l_{t,A_t}]$. For simplicity, denote l_{t,A_t} by $\widehat{l}_t$, also let $\mathcal{T} \triangleq \{t \in [T] : \neg\mathcal{E}_t\}$ denote the set of all skipped time indices. We begin with the following important observation:

Lemma C.3. *For any indices $s, t \in \mathcal{T}$, if $\widehat{l}_t < 2\widehat{l}_s$, we have $s + d_s \ge t$, i.e., the feedback of action A_s arrives later than the decision A_t being made.*

Proof. Otherwise $\widehat{l}_s$ would be used to update $\widehat{L}_{s+d_s+1}$, thus $\widehat{L}_{s+d_s+1} \ge 2\widehat{l}_s$, but $t \ge s + d_s + 1$ so we should have $\widehat{L}_t \ge \widehat{L}_{s+d_s+1} \ge 2\widehat{l}_s$, a contradiction. □

Suppose $|\mathcal{T}| = m$ and denote by $t_1, t_2, \cdots, t_m$ the elements of $\mathcal{T}$, which are sorted in the *arrival order* of their feedback. Consider partitioning $\mathcal{T}$ into non-empty subsets $\mathcal{T}_1 \triangleq \{t_1, t_2, \ldots, t_{m_1}\}, \mathcal{T}_2 \triangleq \{t_{m_1+1}, t_{m_1+2}, \ldots, t_{m_2}\}, \cdots, \mathcal{T}_m \triangleq \{t_{m_{k-1}+1}, \ldots, t_{m_k}\}$ (specially, we denote $m_0 = 0$, $m_k = m$), with the following two properties hold:

- (intra-subset loss upper bound) for each $0 \le i < k$, we have $\widehat{l}_{t_j} < 2\widehat{l}_{t_{m_i+1}}$ for all $m_i < j \le m_{i+1}$;
- (inter-subset loss lower-bound) for each $0 \le i < k-1$, we have $\widehat{l}_{m_{i+1}+1} \ge 2\widehat{l}_{m_i+1}$.

Such partition does exist and can be found by a greedy scan over $\mathcal{T}$. The intra-subset and inter-subset loss requirments allow us to conclude that the number of subsets k is $\mathcal{O}(\log L)$. Besides, for each subset $\mathcal{T}_i$, Lemma C.3 states that all time indices in this subset are no later than the arrival of the feedback of time t_{m_i+1}, namely $t_{m_i+1} + d_{t_{m_i+1}}$. Therefore, $|\mathcal{T}_i| \le d_{t_{m_i+1}} + 1 \le \mathcal{O}(\sqrt{D})$.

After making these observations, one can bound the total losses incurred by rounds in each $\mathcal{T}_i$ by

$$\begin{aligned}\sum_{t\in\mathcal{T}_i} \widehat{l}_t &\overset{(a)}{\le} 2|\mathcal{T}_i|\widehat{l}_{t_{m_i+1}} \\ &\overset{(b)}{\le} 4\cdot 2^{-(k-i)}|\mathcal{T}_i|\widehat{l}_{t_{m_{k-1}+1}} \\ &\overset{(c)}{\le} 4\cdot 2^{-(k-i)}\cdot\mathcal{O}(\sqrt{D})\cdot L \\ &\le \mathcal{O}(2^{-(k-i)}\sqrt{D}L)\end{aligned}$$

where in step (a) we make use of the intra-subset loss relationship in $\mathcal{T}_i$, in step (b) we apply intra-subset loss relationship for $k-i-2$ times to control $\widehat{l}_{t_{m_i+1}}$ by $\widehat{l}_{t_{m_{k-1}+1}}$, in step (c) we apply $|\mathcal{T}_i| \le \mathcal{O}(\sqrt{D})$.

Thus the total loss incurred by $\mathcal{T}$, i.e., the skipping error, is bounded by $\mathcal{O}(\sqrt{D}L)$.

D. Technical Details for `Banker-SFLBINF`

In this section, we introduce our algorithm `Banker-SFLBINF` and also analyze its performance.

D.1. Algorithm Design

Algorithm 4 `Banker-SFLBINF` for Delayed Scale-Free Adversarial MAB

```
1: Initialize $\widehat{L}_1 = 1$, $\mathfrak{D}_0 \leftarrow 1$
2: Run Banker-OMD with $x_0 = {}^1/_K$, the log-barrier regularizer $\Psi(x) = -\sum_{i=1}^K \ln x_i$ and GreedyPick scheduler,
   with some additional processing at each round as below.
3: for $t = 1, 2, \ldots$ do
4:     $a_t \leftarrow$ missing, $\mathfrak{d}_t \leftarrow \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s=\text{missing}\}}$.
5:     $\mathfrak{D}_t \leftarrow \mathfrak{D}_{t-1} + \mathfrak{d}_t$.        ▷ Also maintain ordinary experienced total delays
6:     $D_t \leftarrow \sum_{s\in\{1,\ldots,t\}:a_s=\text{missing}} (\mathfrak{d}_s + 1)\widehat{L}_s^2 + \sum_{s\in\{1,\ldots,t\}:a_s=\text{arrived}} (\mathfrak{d}_s + 1) l_{s,A_s}^2$.
7:     $\sigma_t = \left( (\mathfrak{d}_t + 1)\sqrt{\frac{\ln(3+D_t/\widehat{L}_t^2)}{3+D_t}}\sqrt{K \ln T} \right)^{-1}$.
8:     if $\mathfrak{d}_t \le \sqrt{{}^{\mathfrak{D}_t}/_K}$ then
9:         $\sigma_t \leftarrow \max\{\sigma_t, 2\widehat{L}_t\}$
10:    Pick new action $x_t$ as indicated by Algorithm 2. Play $A_t \sim x_t$.
11:    Initialize $\widehat{L}_{t+1} \leftarrow \widehat{L}_t$.
12:    for upon receiving each new feedback $(s, l_{s,A_s})$ do
13:        $\widehat{L}_{t+1} \leftarrow \max\{\widehat{L}_{t+1}, 2|l_{s,A_s}|\}$.
14:        if $|l_{s,A_s}| > \widehat{L}_s$ or $l_{s,A_s} < -\frac{1}{2}\sigma_s$ then
15:            $\widetilde{l}_s \leftarrow \mathbf{0}$; $a_s \leftarrow$ skipped.        ▷ an additional skipping criterion
16:        else
17:            $\widetilde{l}_s \leftarrow \frac{l_{s,A_s}}{x_{s,A_s}} \mathbf{1}_{A_s}$; $a_s \leftarrow$ arrived
18:        $z_s \leftarrow \nabla\Psi^*(\nabla\Psi(x_s) - \frac{1}{\sigma_s}\widetilde{l}_s)$.
```

Compared to `Banker-SFTINF`, this algorithm uses log-barrier regularizers $\Psi(x) = -\sum_{i=1}^K \ln x_i$ (Abernethy et al., 2015). This regularizer tends to allocate more exploration probabilities to all arms (Agarwal et al., 2017) and leads to various adaptation properties (Wei and Luo, 2018) — which enable us to derive a small-loss style (Neu, 2015) regret bound. Besides the different regularizers, the main differences between Algorithms 3 and 4 are highlighted in blue, including different action scales (Lines 7 and 9) and one more skipping criterion (Line 14).

The change in the regularizer together with the more strict skipping criterion (Line 14) makes it possible to control each immediate cost term $\mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)]$ to by some quantity proportional to $1/\sigma_t$ even when the losses are negative. One can refer to Lemma D.2 in the appendix for more details. The post-processing to the learning scales (Line 9) is a technical modification matching up with Line 14 to make the regret still of order $\mathcal{O}(\sqrt{K})$. At last, the meticulous D_t in Line 6 allows us to additionally derive the small-loss style regret bound.

D.2. Regret Decomposition

The regret decomposition we will use here is slightly different from Eq. (19) because Ψ is now the log-barrier function, which explodes on the boundary of $\Delta^{[K]}$.

Lemma D.1. *Define*

$$\Delta'^{[K]} \triangleq \{x \in \Delta^{[K]} : x_i \ge 1/T, \quad \forall i \in [K]\},$$

then, Algorithm 4 guarantees

$$\mathfrak{R}_T \le \underbrace{\mathcal{O}(KL) + B_T \sup_{y\in\Delta'^{[K]}} D_\Psi(y, x_0))}_{\textit{Total investment}} + \underbrace{\sum_{t=1}^T \mathbb{E}[\sigma_t D_\Psi(x_t, \widetilde{z}_t)]}_{\textit{Immediate costs}} + \underbrace{\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} l_{t,A_t}]}_{\textit{Skipping error}} \tag{20}$$

where $\mathcal{E}_t$ denotes the event that the feedback of round t's action is marked to "skipped", σ_t, B_t, x_t, $\widetilde{z}_t$ are variables determined by Algorithm 4.

Proof. Denote by i^* the actual offline optimal arm, let $\widetilde{y}$ be the l_2-projection of $\mathbf{1}_{i^*}$ onto $\Delta'^{[K]}$, it is easy to see (for sufficiently large T)

$$\widetilde{y} = \begin{cases} 1 - \frac{K-1}{T} & \text{if } i = i^* \\ \frac{1}{T} & \text{otherwise.} \end{cases}$$

Hence

$$\begin{aligned} \mathfrak{R}_T &= \mathbb{E}[\sum_{t=1}^{T} \langle l_t, x_t - \mathbf{1}_{i^*} \rangle] \\ &= \mathbb{E}[\sum_{t=1}^{T} \langle l_t, x_t - \widetilde{y} \rangle] + \mathbb{E}[\sum_{t=1}^{T} \langle l_t, \widetilde{y} - \mathbf{1}_{i^*} \rangle] \\ &\le \mathbb{E}[\sup_{y \in \Delta'^{[K]}} \sum_{t=1}^{T} \langle l_t, x_t - \widetilde{y} \rangle] + \mathbb{E}[\sum_{t=1}^{T} \|l_t\|_\infty \cdot \|\widetilde{y} - \mathbf{1}_{i^*}\|_1] \\ &\le E[\sup_{y \in \Delta'^{[K]}} \sum_{t=1}^{T} \langle l_t, x_t - \widetilde{y} \rangle] + 2KT, \end{aligned}$$

the first term can then be analysed normally using Theorem 4.2. □

Below, we will control the three terms in Eq. (20).

D.3. Bound for Immediate Costs

`Banker-SFLBINF` (Algorithm 4) is designed for scale-free MAB problem instance with general (possibly negative) losses, where the argument when we prove Lemma C.2 no longer works. When we are dealing with general losses, the importance sampling estimators $\widetilde{l}_t$ may contain negative components, and we cannot simply upper bound the Hessian $\nabla^2\Psi^*(\theta_t)$ by $\nabla^2\Psi^*(\nabla x_t)$. To solve this issue, in Algorithm 4, we instead use log-barrier $\Psi(x) = -\sum_{i=1}^{K} \ln(x_i)$ as the regularizer to utilize the following property:

Lemma D.2. *For any x_t in the interior of $\triangle^{[K]}$, $\sigma_t > 0$, $\widetilde{l}_t \in \mathbb{R}^K$, let Ψ be the log-barrier function, $\widetilde{z}_t = \nabla\Psi^*(\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t)$, if for all $i \in [K]$ we have $\widetilde{z}_{t,i} \le 2x_{t,i}$, then*

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) \le 2\sigma_t^{-1} \sum_{i=1}^{K} x_{t,i}^2 \widetilde{l}_{t,i}^2.$$

In particular, if A_t is an independent sample from $[K]$ according to x_t, $\widehat{l}_t \in \mathbb{R}^K$, and $\widetilde{l}_t = \frac{\widehat{l}_{t,A_t}}{x_{t,A_t}} \mathbf{1}_{A_t}$ $\widetilde{l}_t$ is an importance sampling estimator determined by $\widehat{l}_t$ and A_t, we have

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) \le 2\sigma_t^{-1} l_{t,A_t}^2.$$

Proof. Similar to the proof of Lemma C.2, we can write

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) = \frac{1}{2}\sigma_t^{-1} \left\|\widetilde{l}_t\right\|^2_{\nabla^2\Psi^*(w_t)} = \frac{1}{2}\sigma_t^{-1} \left\|\widetilde{l}_t\right\|^2_{\nabla^2\Psi(\nabla\Psi^*(w_t))^{-1}} \tag{21}$$

where $w_t = \nabla\Psi(x_t) - \frac{\theta}{\sigma_t}\widetilde{l}_t$ for some $\theta \in (0,1)$. When Ψ is the log-barrier function, we have $\nabla\Psi(x) = (-1/x_1, -1/x_2, \ldots, -1/x_K)^T$, $\nabla\Psi^*(\theta) = (-1/\theta_1, -1/\theta_2, \ldots, -1/\theta_K)^\mathsf{T}$. The monotonicity of each coordinate of $\nabla\Psi^*$ implies that for any $i \in [K]$ we have

$$\min\{\widetilde{z}_{t,i}, x_{t,i}\} \le \nabla\Psi^*(w_t)_i \le \max\{\widetilde{z}_{t,i}, x_{t,i}\}.$$

The condition $\widetilde{z}_{t,i} \leq 2x_{t,i}$ for all $i \in [K]$ implies that $\nabla\Psi^*(w_t)_i \leq 2x_{t,i}$ for all i. Plugging this upper bound and $\nabla^2\Psi(x) = \operatorname{diag}(x_1^{-2}, x_2^{-2}, \dots, x_K^{-2})$ into (21), we get

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) \leq 2\sigma_t^{-1} \left\|\widetilde{l}_t\right\|^2_{\nabla^2\Psi(x_t)^{-1}} = 2\sigma_t^{-1} \sum_{i=1}^{K} x_{t,i}^2 \widetilde{l}_{t,i}^2.$$

□

Intuitively, Lemma D.2 suggests that it remains safe to apply single-step OMD regret upper bounds when the intermediate result $\widetilde{z}_t$ does not get too large, which is automatically guaranteed when the actual suffered loss l_{t,A_t} is not too large compared with the action scale σ_t. This fact is formally stated as follows.

Lemma D.3. *For $l_t, x_t, A_t, \widetilde{z}_t$ discussed in Lemma D.2, if $l_{t,A_t} \geq -\frac{1}{2}\sigma_t$, then $\widetilde{z}_{t,A_t} \leq 2x_{t,A_t}$.*

Proof. It suffice to investigate the mirror map

$$\nabla\Psi(x) = (-1/x_1, -1/x_2, \dots, -1/x_K)^\mathsf{T}$$

and

$$\nabla\Psi^*(\theta) = (-1/\theta_1, -1/\theta_2, \dots, -1/\theta_K)^\mathsf{T}.$$

Now $\widetilde{l}_t$ is an importance sampling estimator, hence only the A_t-th coordinate can be non-zero and $\widetilde{z}_{t,i} = x_{t,i}$ for all $i \neq A_t$. As for the A_t-th coordinate, we have

$$\begin{aligned}\widetilde{z}_{t,A_t} &= \nabla\Psi^*\left(\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t\right)_{A_t} \\ &= -\left(-x_{t,A_t}^{-1} - \frac{l_{t,A_t}}{\sigma_t x_{t,A_t}}\right)^{-1} \\ &= x_{t,A_t} \cdot \left(1 + \frac{l_{t,A_t}}{\sigma_t}\right)^{-1},\end{aligned}$$

it is then easy to see that we have $\widetilde{z}_{t,A_t} \leq 2x_{t,A_t}$ whenever $l_{t,A_t} \geq -\frac{1}{2}\sigma_t$. □

Recall that in Algorithm 4 we slightly modified the skipping criteria (Line 14) to guarantee the effective *pre-importance-sampling* l_{t,A_t} fed into `Banker-OMD` no less than $-\frac{1}{2}\sigma_t$, thus by Lemma D.3, we can apply Lemma D.2 to all summands in the immediate costs term in Eq. (19). Specifically, we have for all $1 \leq t \leq T$,

$$\sigma_t D_\Psi(x_t, \widetilde{z}_t) \leq \mathbb{1}_{\neg\mathcal{E}_t} 2\sigma_t^{-1} l_{t,A_t}^2.$$

Define $\widetilde{D}_t$ to be a quantity similar to D_t used in the algorithm by

$$\widetilde{D}_t \triangleq \sum_{s \leq t : \neg\mathcal{E}_t} (\mathfrak{d}_s + 1) l_{s,A_s}^2.$$

Recall that the actual D_t used in Algorithm 4 is

$$D_t \triangleq \sum_{s \leq t : a_s = \text{“missing” at the beginning of time } t} (\mathfrak{d}_s + 1)\widehat{L}_s^2 + \sum_{s \in \{1,\dots,t\} : a_s = \text{“arrived” at } t} (\mathfrak{d}_s + 1) l_{s,A_s}^2$$

Compared to D_t, $\widetilde{D}_t$ does not take into account rounds that were not skipped at the beginning of time t but are skipped some time later. Also, for an unskipped round s, it contributes $(\mathfrak{d}_s + 1) l_{s,A_s}^2$ to $\widetilde{D}_t$, but $l_{s,A_s}^2 \widehat{L}_s^2$ to D_t. It is thus guaranteed that $\widetilde{D}_t \leq D_t$.

We further define

$$\widetilde{\sigma}_t = \left((\mathfrak{d}_t + 1) \sqrt{\frac{\ln(3 + \widetilde{D}_t / \widehat{L}_t^2)}{3 + \widetilde{D}_t}} \sqrt{K \ln T} \right)^{-1},$$

which is basically σ_t with D_t replaced by $\widetilde{D}_t$. Then $\sigma_t \geq \widetilde{\sigma}_t$, again by the monotonicity of $(3+x)/\ln(3+x/a)$. Then we can write

$$\begin{aligned}
\sum_{t=1}^T \sigma_t D_\Psi(x_t, \widetilde{z}_t) &\leq 2 \sum_{t=1}^T \mathbb{1}_{\neg \mathcal{E}_t} \sigma_t^{-1} l_{t,A_t}^2 \\
&= 2 \sum_{t \leq T : \neg \mathcal{E}_t} \sigma_t^{-1} l_{t,A_t}^2 \\
&\leq 2 \sum_{t \leq T : \neg \mathcal{E}_t} \widetilde{\sigma}_t^{-1} l_{t,A_t}^2 \\
&= 2\sqrt{K \ln T} \sum_{t \leq T : \neg \mathcal{E}_t} \sqrt{\ln(3 + \widetilde{D}_t / \widehat{L}_t^2)} \cdot \frac{(\mathfrak{d}_t + 1) l_{t,A_t}^2}{\sqrt{3 + \widetilde{D}_t}} \\
&\overset{(a)}{\leq} 2\sqrt{K \ln T} \cdot \mathcal{O}(\sqrt{\log(D+T)}) \sum_{t \leq T : \neg \mathcal{E}_t} \frac{(\mathfrak{d}_t + 1) l_{t,A_t}^2}{\sqrt{3 + \widetilde{D}_t}} \\
&\overset{(b)}{\leq} 2\sqrt{K \ln T} \cdot \mathcal{O}(\sqrt{\log(D+T)}) \cdot \mathcal{O}(\sqrt{1 + \widetilde{D}_T}) \qquad (22) \\
&\leq \mathcal{O}(L\sqrt{K(D+T) \log T \log(T+D)}) \\
&\leq \mathcal{O}(L\sqrt{K(D+T)} \log T)
\end{aligned}$$

where in step (a) we make use of the fact that $\widetilde{D}_t / \widehat{L}_t^2 \leq t + \sum_{s \leq t} \mathfrak{d}_s$ and then universally bound all $\widetilde{D}_t / \widehat{L}_t^2$ by $D + T$; in step (b) we utilize the fact that D_t is the cumulative sum of $1_{\neg \mathcal{E}_s}(\mathfrak{d}_s + 1) l_{s,A_s}^2$, which is just the numerator of each summand, hence we can apply the summation lemma Lemma B.1.

Let $\widetilde{\mathfrak{L}}_T^2 \triangleq \sum_{t=1}^T (\mathfrak{d}_t + 1) l_{t,A_t}^2$ denote the cumulative actually suffered square loss, weighted by the delay backlog size, we can see $\widetilde{D}_T \leq \widetilde{\mathfrak{L}}_T^2$ since $\widetilde{D}_T$ only takes unskipped time slots into account. Plugging $\widetilde{D}_T \leq \widetilde{\mathfrak{L}}_T^2$ into Eq. (22), then taking expectation on both sides, finally noticing that square root is a concave operation, we get

$$\mathbb{E}\left[\sum_{t=1}^T \sigma_t D_\Psi(x_t, \widetilde{z}_t) \right] \leq \mathcal{O}(\sqrt{K(1 + \mathbb{E}[\widetilde{\mathfrak{L}}_T^2])} \log T).$$

D.4. Bound for Total Investment Term

To bound the total investment term, we first identify the order of the factor $\sup_{y \in \Delta'^{[K]}} D_\Psi(y, x_0))$:

Lemma D.4. *When Ψ is the log-barrier function, we have $D_\Psi(y, x_0) \leq K \ln T + K = \mathcal{O}(K \log T)$ for all $y \in \Delta'^{[K]}$.*

Proof. Notice that $\Delta'^{[K]}$ is a compact convex subset of $\mathbb{R}_+^K$ (actually, it is a polyhedron), Ψ is a convex function over $\mathbb{R}_+^K$, the maximum value of Ψ must achieve on the boundary of $\Delta'^{[K]}$. Due to the symmetry of coordinates of x_0, it suffices to verify the bound for all vertices y from $\Delta'^{[K]}$, and now we have

$$\begin{aligned}
D_\Psi(y, x_0) = (K-1)&\left(-\ln \frac{1}{T} + \ln \frac{1}{K} + K\left(\frac{1}{T} - \frac{1}{K} \right) \right) \\
&+ \left(-\ln\left(1 - \frac{K-1}{T}\right) + \ln \frac{1}{K} + K\left(1 - \frac{K-1}{T} - \frac{1}{K}\right) \right) \\
&\leq (K-1)\ln T - \ln\left(1 - \frac{K-1}{T}\right) + K
\end{aligned}$$

$$\le K \ln T + K.$$

□

It then suffices to bound the leading coefficient B_T.

The choice of σ_t in Algorithm 4 (Line 7, Line 9) satisfies

$$\sigma_t \le \left[(\mathfrak{d}_t+1)\sqrt{\frac{\ln(3+D_t/\widehat{L}_t^2)}{3+D_t}}\sqrt{K\ln T}\right]^{-1} + \mathbb{1}_{\{\mathfrak{d}_t \le \sqrt{\frac{\mathfrak{D}_t}{K}}\}} \cdot 2\widehat{L}_t.$$

By Lemma 4.5, there exists some $t_0 \le T$ such that

$$B_T = B_{t_0} = \sigma_{t_0} + \sum_{s=1}^{t_0-1} \mathbb{I}[s+d_s \ge t_0]\sigma_s. \tag{23}$$

Let $t_1 < t_2 < \cdots < t_m$ be the time slots whose feedback has not arrived at time slot t_0, we will have $m = \mathcal{O}(\sqrt{D})$. Furthermore, at time slot t_i, we must have $\mathfrak{d}_{t_i} \ge i$ as the feedback of $t_1, t_2, \cdots, t_i$ are all absent, which can be used to bound the number of i's satisfying $\mathfrak{d}_{t_i} \le \sqrt{\frac{\mathfrak{D}_{t_i}}{K}}$. Therefore, B_T can be further bounded by

$$\begin{aligned}
B_T &= \sum_{i=0}^m \sigma_{t_i} \\
&\le \sum_{i=0}^m \left\{ \left[(\mathfrak{d}_{t_i}+1)\sqrt{\frac{\ln(3+D_{t_i}/\widehat{L}_{t_i}^2)}{3+D_{t_i}}}\sqrt{K\ln T}\right]^{-1} + \mathbb{1}_{\{\mathfrak{d}_{t_i} \le \sqrt{\frac{\mathfrak{D}_{t_i}}{K}}\}} \cdot 2\widehat{L}_{t_i} \right\} \\
&\overset{(a)}{\le} \sum_{i=0}^m \left[(\mathfrak{d}_{t_i}+1)\sqrt{\frac{\ln(3+D_{t_i}/\widehat{L}_{t_i}^2)}{3+D_{t_i}}}\sqrt{K\ln T}\right]^{-1} + 4\left(\sqrt{\frac{D}{K}}+1\right)L \\
&= \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \sum_{i=0}^m \frac{1}{\mathfrak{d}_{t_i}+1}\sqrt{\frac{3+D_{t_i}}{\ln(3+D_{t_i}/\widehat{L}_{t_i}^2)}} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{2L}{\sqrt{K\ln T}} \sum_{i=0}^m \frac{1}{\mathfrak{d}_{t_i}+1}\sqrt{\frac{3+D_{t_i}/\widehat{L}_{t_i}^2}{\ln(3+D_{t_i}/\widehat{L}_{t_i}^2)}} \\
&\overset{(b)}{\le} \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{2L}{\sqrt{K\ln T}} \cdot \sqrt{\frac{3+D+T}{\ln(3+D+T)}} \sum_{i=0}^m \frac{1}{\mathfrak{d}_{t_i}+1} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{2L}{\sqrt{K\ln T}} \cdot \sqrt{\frac{3+D+T}{\ln(3+D+T)}} \cdot \mathcal{O}(\log D) \\
&= \mathcal{O}\left(\sqrt{\frac{D}{K}}L + \sqrt{\frac{(D+T)\log D}{K\log T}}L\right)
\end{aligned} \tag{24}$$

where in step (a), we simply bound $\widehat{L}_{t,i}$ by $2L$, and utilizing $\mathfrak{d}_{t_i} \ge i$ for all $i \ge 1$ to write $\sum_{i=0}^m \mathbb{1}_{\{\mathfrak{d}_{t_i} \le \sqrt{\frac{\mathfrak{D}_{t_i}}{K}}\}} \le \sum_{i=0}^m \mathbb{1}_{\{i \le \sqrt{\frac{\mathfrak{D}_{t_i}}{K}}\}} \le \sum_{0=1}^m \mathbb{1}_{\{i \le \sqrt{\frac{D}{K}}\}} \le \sqrt{\frac{D}{K}} + 1$; (b) uses the monotonicity of $(3+x)/ln(3+x)$ and the fact that $D_t/\widehat{L}_t^2 \le t + \sum_{s=1}^t \mathfrak{d}_s$.

In order to get a bound in $\widetilde{\mathfrak{L}}_T^2 = \sum_{t=1}^T (\mathfrak{d}_t+1) l_{t,A_t}^2$, similarly define $\widetilde{\mathfrak{L}}_t^2 \triangleq \sum_{s=1}^t (\mathfrak{d}_s+1) l_{s,i_s}^2$ and define $\widetilde{m} \triangleq \max_{1\le t\le T} \mathfrak{d}_t$. We must have $\widetilde{m} = \mathcal{O}(\sqrt{D})$ since $D \ge \binom{m+1}{2}$. Recall that in Algorithm 4 the value of D_t we pick is

$$\begin{aligned}
D_t &= \sum_{s\le t: a_s = \text{“missing” at } t} (\mathfrak{d}_s+1)\widehat{L}_s^2 + \sum_{s\le t: a_s = \text{“arrived” before } t} (\mathfrak{d}_s+1) l_{s,A_s}^2 \\
&\le \sum_{s=1}^{t} (\mathfrak{d}_s+1) l_{s,A_s}^2 + \widehat{L}_t^2 \sum_{s\le t: a_s = \text{“missing” at } t} (\mathfrak{d}_s+1) \\
&\overset{(a)}{\le} \widetilde{\mathfrak{L}}_t^2 + \widehat{L}_t^2 (\widetilde{m}+1)^2 \\
&\le \widetilde{\mathfrak{L}}_t^2 + 100D\widehat{L}_t^2
\end{aligned}$$

where (a) holds because in the sum $\sum_{s\le t: a_s = \text{“missing” at } t}(\mathfrak{d}_s+1)$, the number of summands and the value of each summand are both bounded by $\widetilde{m}+1$.

Leveraging this upper bound for D_t’s, we can continue from Eq. (24) to obtain an upper bound in $\widetilde{\mathfrak{L}}_T^2$:

$$\begin{aligned}
B_T &\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \sum_{i=0}^{m} \frac{1}{\mathfrak{d}_{t_i}+1} \sqrt{\frac{3+D_{t_i}}{\ln(3+D_{t_i}/\widehat{L}_{t_i}^2)}} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \sum_{i=0}^{m} \frac{\widehat{L}_{t_i}}{\mathfrak{d}_{t_i}+1} \sqrt{\frac{3+D_{t_i}/\widehat{L}_{t_i}^2}{\ln(3+D_{t_i}/\widehat{L}_{t_i}^2)}} \\
&\overset{(a)}{\le} \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \sum_{i=0}^{m} \frac{\widehat{L}_{t_i}}{\mathfrak{d}_{t_i}+1} \sqrt{\frac{3+\widetilde{\mathfrak{L}}_{t_i}^2/\widehat{L}_{t_i}^2+100D}{\ln(3+\widetilde{\mathfrak{L}}_{t_i}^2/\widehat{L}_{t_i}^2+100D)}} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \max_{1\le t\le T}\left\{\widehat{L}_t \sqrt{\frac{3+\widetilde{\mathfrak{L}}_t^2/\widehat{L}_t^2+100D}{\ln(3+\widetilde{\mathfrak{L}}_t^2/\widehat{L}_t^2+100D)}}\right\} \sum_{i=0}^{m} \frac{1}{\mathfrak{d}_{t_i}+1} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \max_{1\le t\le T}\left\{\widehat{L}_t \sqrt{3+\widetilde{\mathfrak{L}}_t^2/\widehat{L}_t^2+100D}\right\} \frac{1}{\sqrt{\ln(3+100D)}} \sum_{i=0}^{m} \frac{1}{\mathfrak{d}_{t_i}+1} \\
&= \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \max_{1\le t\le T}\left\{\widehat{L}_t \sqrt{3+\widetilde{\mathfrak{L}}_t^2/\widehat{L}_t^2+100D}\right\} \frac{\mathcal{O}(\log D)}{\sqrt{\ln(3+100D)}} \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \frac{1}{\sqrt{K\ln T}} \sqrt{\widetilde{\mathfrak{L}}_T^2 + (3+100D)\widehat{L}_T^2} \frac{\mathcal{O}(\log D)}{\sqrt{\ln(3+100D)}} \\
&= \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \mathcal{O}\left(\sqrt{\frac{\log D}{K\log T}} \sqrt{\widetilde{\mathfrak{L}}_T^2 + (3+100D)\widehat{L}_T^2}\right) \\
&\le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \mathcal{O}\left(\sqrt{\frac{\widetilde{\mathfrak{L}}_T^2 \log D}{K\log T}} + \sqrt{\frac{(1+D)\log D}{K\log T}}L\right)
\end{aligned}$$

where step (a) plugs in our D_t bound into $\widetilde{\mathfrak{L}}_t^2$. Taking expectation on both sides, we can see

$$\mathbb{E}[B_T] \le \mathcal{O}\left(\sqrt{\frac{D}{K}}L\right) + \mathcal{O}\left(\sqrt{\frac{\mathbb{E}[\widetilde{\mathfrak{L}}_T^2]\log D}{K\log T}} + \sqrt{\frac{(1+D)\log D}{K\log T}}L\right).$$

Combining the bounds for $\mathbb{E}[B_T]$ and $\sup D_\Psi(y, x_0)$, we can make the following conclusion:

Lemma D.5. *In Eq. (20), the total investment term when y is restricted on $\Delta'^{[K]}$ is bounded by*

$$\sup_{y\in\Delta'_{[K-1]}} \mathbb{E}\left[B_T \cdot D_\Psi(y, x_0)\right] = \mathcal{O}\left(\sqrt{KD}\log T L + \sqrt{K(D+T)\log D\log T}L\right),$$

$$\sup_{y\in\Delta'_{[K-1]}} \mathbb{E}\left[B_T \cdot D_\Psi(y,x_0)\right] = \mathcal{O}\left(\sqrt{KD}\log TL + \sqrt{K\log D\log T\,\mathbb{E}[\widetilde{\mathfrak{L}}_T^2]} + \sqrt{K(1+D)\log D\log T}L\right).$$

D.5. Bound for Skipping Error

It finally remains to bound $\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} l_{t,A_t}]$, the skipping error term in Eq. (20). We claim the following result:

Lemma D.6. *In Algorithm 4, the expected number of skipped time slots, namely the $\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_t}\right]$ term in Eq. (20), is bounded by $\mathcal{O}(\sqrt{D}\log L + \sqrt{KD})$. Furthermore, the skipping regret is bounded by $\mathcal{O}((\sqrt{D}\log L + \sqrt{KD})L)$.*

Proof. For any $1 \le t \le T$, define the following two events

$$U_t \triangleq \left\{|l_{t,A_t}| > \widehat{L}_t\right\},$$

$$V_t \triangleq \left\{|l_{t,A_t}| \le \widehat{L}_t, l_{t,A_t} < -\frac{1}{2}\sigma_t\right\}.$$

In other words, U_t happens if and only if round t is skipped by Algorithm 4 due to the skipping criterion inherited from Algorithm 3, and V_t happens if and only if round t is skipped *solely* due to the new skipping criterion $l_{t,A_t} < -\frac{1}{2}\sigma_t$. Hence our goal reduces to bound $\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{U_t}\right] + \sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{V_t}\right]$, where the first sum is bounded by $\mathcal{O}(\sqrt{D}\log L)$ according the argument in Appendix C.3. Therefore, it suffices to bound $\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{V_t}\right]$.

Recall that we maintain experienced total delay $\mathfrak{D}_t$ in Algorithm 4 and we have $\mathfrak{D}_0 = 1$, $\mathfrak{D}_T = D+1$. For any integer $i \ge 0$, define a stopping time

$$\tau_i \triangleq \inf\left\{t \ge 0 : \mathfrak{D}_t \ge \frac{D}{2^i}\right\}.$$

Clearly, we have $\tau_i \le T$ and $\tau_0 \ge \tau_1 \ge \cdots$ almost surely holds. The idea is to bound the sum of $\mathbb{1}_{V_t}$ during any two successive stopping times τ_i and τ_{i-1}. That is, to bound $\sum_{t=\tau_i}^{\tau_{i-1}-1} \mathbb{1}_{V_t}$ for each $i \ge 1$.

Notice that for a $t \ge \tau_i$, the value of $\mathfrak{D}_t$ is at least $D/2^i$. If V_t happens, then at time t, Line 9 of Algorithm 4 cannot be executed (otherwise we will have $\sigma_t \ge 2\widehat{L}_t$ and $l_{t,A_t} < -\frac{1}{2}\sigma_t \le -\widehat{L}_t$, a contradiction), which means $\mathfrak{d}_t > \sqrt{\frac{\mathfrak{D}_t}{K}}$. Therefore conditioned on V_t and $t > \tau_i$, we have $\mathfrak{D}_t - \mathfrak{D}_{t-1} = \mathfrak{d}_t > \sqrt{\frac{\mathfrak{D}_t}{K}} \ge \sqrt{\frac{D}{2^iK}}$, and $\sum_{t=\tau_i}^{\tau_{i-1}-1} \mathbb{1}_{V_t}$ must be no more than $\frac{D}{2^{i-1}} \Big/ \sqrt{\frac{D}{2^iK}} = \sqrt{KD}2^{1-i/2}$. We can then conclude that

$$\begin{aligned}\sum_{t=1}^T \mathbb{1}_{V_t} &= \sum_{t=\tau_0}^T \mathbb{1}_{V_t} + \sum_{i=1}^\infty \sum_{t=\tau_i}^{\tau_{i-1}-1} \mathbb{1}_{V_t} \\ &\le (D+1) \Big/ \sqrt{\frac{D}{K}} + \sum_{i=1}^\infty \sqrt{KD}2^{-\frac{i}{2}+1} = \mathcal{O}(\sqrt{KD}).\end{aligned}$$

Therefore, we have $\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t}] = \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{U_t}] + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{V_t}] = \mathcal{O}(\sqrt{D}\log L + \sqrt{KD})$. Furthermore, the total skipping regret is therefore no more than $L\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t}] = \mathcal{O}(\sqrt{D}L\log L + \sqrt{KD}L)$. □

E. Technical Details for Banker-BOLO

E.1. Algorithm Design

In the language of Banker-OMD, Banker-BOLO uses $x_0 = \nabla\Psi^*(\mathbf{0})$ as the default investment option. As noticed by Abernethy et al. (2008), when Ψ is $\mathcal{O}(n)$-self-concordant and σ_t's are at least $8n$, this choice ensures (Ψ, x_0) to be $(\mathcal{O}(n\log T), \mathcal{O}(n^2))$-regular under the sampling scheme over Dikin ellipsoids (Line 8). According to Theorem 4.6, we then pick the action scale as $\sigma_t = \max\left\{\left(\sqrt{\frac{\ln T}{nt}} + \mathfrak{d}_t\sqrt{\frac{\ln(\mathfrak{D}_t+1)\ln T}{n\mathfrak{D}_t}}\right)^{-1}, 8n\right\}$ and achieve $\widetilde{\mathcal{O}}(n^{3/2}\sqrt{T} + n^2\sqrt{D})$ regret.

In the following section, we give more rigorous justification on the applicability of Banker-OMD to the linear bandit problem, the $(\mathcal{O}(n\log T), \mathcal{O}(n^2))$-regularity of the regularizer, and the sampling scheme combination used by Banker-BOLO.

Algorithm 5 `Banker-BOLO` for Delayed Adversarial Linear Bandits

```
Input: Number of dimension $n$; Time horizen length $T$; Legendre function $\Psi : \mathcal{C} \to \mathbb{R}$.
Output: A sequence of actions $A_1, A_2, \dots, A_T \in \mathcal{C}$.
$\mathfrak{D}_0 \leftarrow 0$.
for $t = 1, 2, \dots, T$ do
    Set $a_t \leftarrow$ missing, $\mathfrak{d}_t \leftarrow \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s=\text{missing}\}}$, and $\mathfrak{D}_t \leftarrow \mathfrak{D}_{t-1} + \mathfrak{d}_t$.
    Calculate $\sigma_t \leftarrow \max\left\{\left(\sqrt{\frac{\ln T}{nt}} + \mathfrak{d}_t\sqrt{\frac{\ln(\mathfrak{D}_t+1)\ln T}{n\mathfrak{D}_t}}\right)^{-1}, 8n\right\}$.
    Pick new action $x_t$ as indicated by Algorithm 2.
    Let $\{e_{t,1}, \dots, e_{t,n}\}$ and $\{\lambda_{t,1}, \dots, \lambda_{t,n}\}$ be the eigenvectors and eigenvalues of $\nabla^2\Psi(x_t)$.
    Sample $i_t$ uniformly from $[n]$ and $\varepsilon_t$ from a Rademacher random variable.
    Play $A_t$ defined by $A_t = x_t + \varepsilon_t \lambda_{t,i_t}^{-1/2} e_{t,i_t}$.          ▷ sample on the Dikin ellipsoid
    for upon each new feedback $(s, \widehat{l}_s)$ do
        $\widetilde{l}_s \leftarrow \widehat{l}_s \cdot n\varepsilon_s \lambda_{s,i_s}^{1/2} \cdot e_{s,i_s}$.          ▷ loss estimator matching with the sampling scheme
        Set $z_s \leftarrow \nabla\Psi^*(\nabla\Psi(x_s) - \frac{1}{\sigma_s}\widetilde{l}_s)$ and $a_s \leftarrow$ arrived.
```

E.2. Regret Analysis

When we use a regularizer Ψ that is a Legendre function natively defined on the convex action set $\mathcal{C}$, we have the following alternatives for Lemma B.4 and Lemma 4.1:

Lemma E.1. *Let $\mathcal{C}$ be a convex subset of $\mathbb{R}^n$ with non-empty interior, then for any $\sigma > 0$, $x, y \in int(\mathcal{C})$, $l \in \mathbb{R}^n$ and Legendre function $\Psi : \mathcal{C} \to \mathbb{R}$, we have*

$$\langle l, x - y\rangle \le \sigma D_\Psi(y, x) - \sigma D_\Psi(y, z) + \sigma D_\Psi(x, z)$$

where

$$z = \operatorname*{arg\,min}_{x' \in \triangle^{[K]}} \langle l, x'\rangle + D_\Psi(x', x), \tag{25}$$

or equivalently,

$$z = \nabla\Psi^*(\nabla\Psi(x) - \frac{1}{\sigma}l).$$

Lemma E.2. *For any $m \ge 1$, $z_1, \dots, z_m \in int(\mathcal{C})$, $\sigma_1, \dots, \sigma_m > 0$ and Legendre function $\Psi : \mathcal{C} \to \mathbb{R}$, let $\sigma = \sum_{i=1}^m \sigma_i$ and*

$$x = \nabla\Psi^*(\sum_{i=1}^m \frac{\sigma_i}{\sigma}\nabla\Psi(z_i)),$$

we have

$$\sigma D_\Psi(y, x) \le \sum_{i=1}^m \sigma_i D_\Psi(y, z_i)$$

for any $y \in int(\mathcal{C})$.

Lemma E.1 and Lemma E.2 can be proved in almost the same way as Lemma B.4 and Lemma E.2. In fact, in the case of Ψ natively defined on $\mathcal{C}$, the update rule Eq. (25) already guarantees $l = \sigma(\nabla\Psi(x) - \nabla\Psi(y))$, we no longer need $\overline{\Psi}$, a "constrained version" of Ψ as we did in the MAB case.

Follow the same reasoning in Section 4, we can see that `Banker-BOLO` has the following regret bound:

$$\sum_{t=1}^T \langle \widetilde{l}_t, x_t - y\rangle \le (\sum_{t=1}^T b_t) D_\Psi(y, \nabla\Psi^*(\mathbf{0})) + \sum_{i=1}^T \sigma_t D_\Psi(x_t, z_t) \tag{26}$$

for any $y \in int(\mathcal{C})$. Theorem 4.6 continues to apply to $\sum_{t=1}^T b_t$ and gives an $\mathcal{O}(\sqrt{T} + \sqrt{D\log D})$ bound for this total investment coefficient. In order to derive the claimed regret bound in Theorem 7.1, it remains to justify the following properties

- $\widetilde{l}_t$ is an unbiased estimate for the true loss vector l_t, i.e., $\mathbb{E}[\widetilde{l}_t \mid \mathcal{F}_{t-1}] = l_t$;
- $\mathbb{E}[\sigma_t D_\Psi(x_t, z_t)]$ can be bounded by $\mathcal{O}(n^2/\sigma_t)$;
- $D_\Psi(y, \nabla\Psi^*(\mathbf{0}))$ can be uniformly bounded by $\mathcal{O}(n \log T)$.

These properties are, therefore, all unrelated to the presence of feedback delays, and all have been proved in (Abernethy et al., 2008). For the sake of completeness, we put the most important technical lemmas here.

Definition E.3. A self-concordant function $\Psi : \mathcal{C} \to \mathbb{R}$ is a C^3 convex function such that

$$|D^3\Psi(x)[h,h,h]| \le 2(D^2\Psi(x)[h,h])^{3/2}.$$

Here, the third-order differential is defined as

$$D^3\Psi(x)[h,h,h] \triangleq \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3} \,|_{t_1=t_2=t_3=0}\, \Psi(x + t_1 h_1 + t_2 h_2 + t_3 h_3).$$

It is further called ϑ-self-concordant if

$$|D\Psi(x)[h]| \le \vartheta^{1/2}[D^2\Psi(x)[h,h]]^{1/2}.$$

Definition E.4. If Ψ is a self-concordant barrier over $\mathcal{C}$, for $x \in int(\mathcal{C})$ and $r > 0$, define the open Dikin ellipsold of radius r centered at x as the set

$$W_r(x) \triangleq \left\{ y \in \mathcal{C} : \|y - x\|_{\nabla^2\Psi(x)^{-1}} \le r \right\}.$$

We have the following properties of Dikin ellipsoids (refer to (Nemirovski, 2004, Page 23) for proofs):

Lemma E.5. *For any $x \in int(\mathcal{C})$, we have*

1. $W_1(x) \subseteq \mathcal{C}$;
2. For any $\|h\|_{\nabla^2\Psi(x)^{-1}} < 1$, we have

$$(1 - \|h\|_{\nabla^2\Psi(x)^{-1}})^2 \nabla\Psi^2(x) \preccurlyeq \nabla^2\Psi(x+h) \preccurlyeq (1 - \|h\|_{\nabla^2\Psi(x)^{-1}})^{-2} \nabla\Psi^2(x).$$

Lemma E.5 asserts that any Dikin ellipsold of radius 1 centered in the interior of $\mathcal{C}$ will be contained in $\mathcal{C}$. Recall that in Algorithm 5, the new action A_t is sampled from the surface of a unit-radius Dikin ellipsold centered at x_t (Line 8), so A_t is guaranteed to be a valid action in $\mathcal{C}$. Lemma E.5 also states that, within a unit-radius Dikin ellipsold inside $\mathcal{C}$, the Hessians of Ψ are "almost proportional" to the Hessian at the center of the ellipsold. This fact plays a crucial role in bounding immediate costs $\mathbb{E}[\sigma_t D_\Psi(x_t, z_t)]$. Prior that, we need another lemma.

Lemma E.6. *For any $1 \le t \le T$, we have*

$$z_t \in W_{\frac{4n}{\sigma_t}}(x_t).$$

Lemma E.6 is Lemma 6 of (Abernethy et al., 2008); refer to the original paper for a proof. It claims that the single-step OMD image z_t is located within a Dikin ellipsold centered at x_t. Combining Lemma E.5 and Lemma E.6, we can derive an upper bound for immediate costs.

Lemma E.7. *In `Banker-BOLO`, if $\sigma_t \ge 8n$, we have*

$$\mathbb{E}[\sigma_t D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}] \le \frac{2n^2}{\sigma_t}.$$

Proof. Similar to the proof of Lemma B.5, we have

$$\sigma_t D_\Psi(x_t, z_t) = \frac{\|\widetilde{l}_t\|^2_{\nabla^2\Psi^*(\theta_t)}}{2\sigma_t} = \frac{\|\widetilde{l}_t\|^2_{\nabla^2\Psi(w_t)^{-1}}}{2\sigma_t}$$

where θ_t is some element inside the line segment connecting $\nabla\Psi(x_t) - \frac{1}{\sigma_t}\widetilde{l}_t$ and $\nabla\Psi(x_t)$, $w_t = \nabla\Psi^*(\theta_t)$. According to Lemma E.6, w_t is located in $W_{\frac{4n}{\sigma_t}}(x_t) \subseteq W_{1/2}(x_t)$. We can thus apply the second result in Lemma E.5 to get

$$\sigma_t D_\Psi(x_t, z_t) = \frac{\|\widetilde{l}_t\|^2_{\nabla^2\Psi(w_t)^{-1}}}{2\sigma_t} \le \frac{2\|\widetilde{l}_t\|^2_{\nabla^2\Psi(x_t)^{-1}}}{\sigma_t}$$

hence

$$\begin{aligned}\mathbb{E}[\sigma_t D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}] &\leq \frac{2}{\sigma_t}\,\mathbb{E}[\|\widetilde{l}_t\|^2_{\nabla^2\Psi(w_t)^{-1}} \mid \mathcal{F}_{t-1}] \\ &\leq \frac{2}{\sigma_t}\,\mathbb{E}[\|n\lambda^{1/2}_{t,i_t}\cdot e_{t,i_t}\|^2_{\nabla^2\Psi(w_t)^{-1}} \mid \mathcal{F}_{t-1}] \\ &= \frac{2n^2}{\sigma_t}.\end{aligned}$$

□

In order to derive an upper bound for the investment term $D_\Psi(y, \nabla\Psi^*(\mathbf{0}))$, we first introduce the following property of ϑ-self-concordant barriers. The proof can be found in (Nesterov and Nemirovskii, 1994, Page 34).

Lemma E.8. *Define*

$$\pi_y(x) \triangleq \inf\{t \geq 0 : y + t^{-1}(x-y) \in \mathcal{C}\}.$$

If Ψ is a ϑ-self-concordant barrier on $\mathcal{C}$, then for any $x, y \in int(\mathcal{C})$, we have

$$\Psi(y) - \Psi(x) \leq \vartheta \ln \frac{1}{1-\pi_x(y)}.$$

Corollary E.9. *Define $\mathcal{C}_T \triangleq \{y \in \mathcal{C} : \pi_{\nabla\Psi^*(\mathbf{0})}(y) \leq 1 - 1/T\}$, if Ψ is a ϑ-self-concordant barrier on $\mathcal{C}$, we have*

$$\sup_{y\in\mathcal{C}_T} D_\Psi(y, \nabla\Psi^*(\mathbf{0})) \leq \vartheta \ln T.$$

This uniform upper bound for $D_\Psi(y, \nabla\Psi^*(\mathbf{0}))$ over $\mathcal{C}_T$ is enough for us to design linear bandits algorithms. The reason is, for any $y \in \mathcal{C}\setminus\mathcal{C}_T$, the definition of $\mathcal{C}_T$ guarantees that there exists a $y' \in \mathcal{C}_T$ such that

$$\sum_{t=1}^T |\langle l_t, y - y'\rangle| = \mathcal{O}(1)$$

uniformly holds. Therefore there is a only constant difference between $\mathfrak{R}_T$ and $\sup_{y\in\mathcal{C}_T}\mathbb{E}[\sum_{t=1}^T\langle l_t, x_t - y\rangle]$.