Scale-Free Adversarial Multi-Armed Bandit with Arbitrary Feedback Delays

Jiatai Huang∗
IIS, Tsinghua University
hjt18@mails.tsinghua.edu.cn

Yan Dai∗
IIS, Tsinghua University
yan-dai20@mails.tsinghua.edu.cn

Longbo Huang
IIS, Tsinghua University
longbohuang@tsinghua.edu.cn

Abstract—We consider the Scale-Free Adversarial Multi-Armed Bandit (MAB) problem with unrestricted feedback delays. In contrast to the standard assumption that all losses are \([0, 1]\)-bounded, in our setting, losses can fall in a general bounded interval \([-L, L]\), unknown to the agent before-hand. Furthermore, the feedback of each arm pull can experience arbitrary delays. We propose an algorithm named SFBanker for this novel setting, which combines a recent banker online mirror descent technique and elaborately designed doubling tricks. We show that SFBanker achieves \(\Omega(\sqrt{K(D+T)L})\) total regret, which is near-optimal compared to the \(\Omega(\sqrt{KL})\) lower-bound ([Cesa-Bianchi et al., 2016]).

1. INTRODUCTION

Multi-Armed Bandit (MAB) is a classical sequential decision game of \(T\) rounds, carried out between an agent and an adversary. In each round \(t = 1, 2, \ldots, T\), the player chooses an arm \(i_t\) among a finite set of actions, namely the set \([K] = \{1, 2, \ldots, K\}\), and suffers from a loss \(\ell_{t,i_t}\) for that action. The suffered loss \(\ell_{t,i_t}\) is available to the player only after the decision \(i_t\) is made. The goal of the agent is to minimize the pseudo-regret \(R_T \triangleq \min_{i \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \ell_{t,i_t} - \sum_{t=1}^{T} \ell_{t,i} \right]\) of \(T\) rounds.

In the setting of delayed MAB, the agent can only observe the bandit feedback \(\ell_{t,i_t}\) after a delay \(d_t\), i.e., at the end of the \(t + d_t\)-th time step. In this paper, we focus on the oblivious adversarial setting, where both the loss vectors \(\ell_t \in \mathbb{L} \subseteq \mathbb{R}^K\) and the delays \(d_t\) can be arbitrarily chosen before the start of the game.

Under the assumption that \(\mathbb{L} = [0, 1]^K\), this problem is already well-studied with regret \(\mathcal{O}(\sqrt{KT} + \sqrt{D\log K})\), which is known to be optimal up to constants ([Zimmert and Seldin, 2020]). If \(\mathbb{L}\) is a subset of a general but known rectangle \([a, b]^K\), the problem can also be easily reduced to the standard \(\mathbb{L} = [0, 1]^K\) case by transforming each feedback \(\ell_{t,i_t}\) into \(\frac{\ell_{t,i_t} - a}{b - a}\).

However, in the scale-free setting, the agent only knows that \(\mathbb{L}\) is a subset of some unknown but bounded rectangle \([-L, L]^K\). It then turns out to be a challenging problem even without any delay [Putta and Agrawal, 2021].

A. Our Contribution

We present the first algorithm SFBanker for the scale-free delayed MAB problem, achieving \(\mathcal{O}(\sqrt{K(D+T)L})\) total regret without any knowledge of the loss vector set \(\mathbb{L}\) or restrictions on the delays \(d_t\). Here \(D \triangleq \sum_{t=1}^{T} d_t\) denote the total delay. The proposed algorithm is built upon the Banker-OMD framework [Huang and Huang, 2021], which provides a general way for decision making even in the presence of feedback delays. SFBanker uses a carefully designed step-size policy to deal with the unknown loss magnitudes.

We show that SFBanker-TINF, a variant of SFBanker, guarantees \(\mathcal{O}(\sqrt{K(D+T)\log(D+T)L})\) total regret when it is known that the loss vectors are all non-negative (i.e., \(\mathbb{L} \subseteq [0, \infty]^K\)). This result is nearly optimal, compared to the \(\Omega(\sqrt{KT} + \sqrt{D\log K} \cdot \log L)\) regret lower-bound (an immediate corollary of the \(\Omega(\sqrt{KT} + \sqrt{D\log K})\) lower-bound in [Zimmert and Seldin, 2020] for the standard \(\mathbb{L} = [0, 1]^K\) case).

In the general case that \(\mathbb{L}\) can be an arbitrary subset of \(\mathbb{R}^K\), we proposed another variant of SFBanker, namely SFBanker-LBINF. This algorithm enjoys a small-loss style regret bound which depends on the expectation of \(|L|\), a quantity depending on the total loss actually suffered. Specially, it also infers an \(\mathcal{O}(\sqrt{KT\log KL} + \log L)\) regret upper-bound for scale-free non-delayed adversarial MABs.

To summarize, we present an overview and comparison of our algorithms and related works in Table I.

B. Related Work

a) Delayed MAB with \(\mathbb{L} = [0, 1]^K\): This setting was first studied by [Cesa-Bianchi et al., 2016], under a uniform delay assumption, i.e., the delays are all equal to \(d\). In this setting, they gave a lower bound \(\Omega(\max\{\sqrt{KT}, \sqrt{D\log K}\})\), which further leads to an \(\Omega(\max\{\sqrt{KT}, \sqrt{D\log K}\})\) lower bound for the general delay case ([Zimmert and Seldin, 2020]). [Bistritz et al., 2019] and [Thune et al., 2019] independently studied the unrestricted, non-uniform delay setting and achieves an \(\mathcal{O}(\sqrt{KT})\) regret bound.

∗These authors contributed equally to this work

If the time-horizon length \(T\) is unknown before-hand, only a \(\Theta(\sqrt{\log T})\) extra factor is suffered (Appendix I).
O(\sqrt{KT}\log K + \sqrt{D\log K}) with prior knowledge on K and D. [Zimmert and Seldin, 2020] provided an algorithm with regret \( O(\sqrt{KT} + \sqrt{D\log K}) \), which requires neither the prior knowledge about \( D, T \) nor the action time knowledge about \( d_t \), which matches the lower bound in [Cesa-Bianchi et al., 2016] up to constants.

b) Scale-free Settings: Scale-free settings were first studied in full-information case (e.g., [De Rooij et al., 2014], [Orabona and Pal, 2018]). For stochastic MAB, [Hadiji and Stoltz, 2020] proposed an \( O(L_{\infty}\sqrt{KT}\log K) \) algorithm. [Putta and Agrawal, 2021] studied the adversarial case, yielding an \( O(\sqrt{KL_T\log K}) \) worst-case upper-bound.

II. NOTATIONS

We use \([N]\) to denote the integer set \( \{1, 2, \cdots, N\} \). Let \( f \) be any strictly convex function defined on a convex set \( \Omega \subseteq \mathbb{R}^K \). For \( x, y \in \Omega \), if \( \nabla f(x) \) exists, we denote the Bregman divergence induced by \( f \) as

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

We use \( f^*(y) \triangleq \sup_{x \in \mathbb{R}^K} \{ (y, x) - f(x) \} \) to denote the Fenchel conjugate of \( f \). We denote the \( K \)-1-dimensional probability simplex by \( \Delta_{[K-1]} = \{ x \in \mathbb{R}_+^K \mid x_1 + x_2 + \cdots + x_K = 1 \} \). Let \( \bar{f} \) denote the restriction of \( f \) on \( \Delta_{[K-1]} \), i.e.,

\[ \bar{f}(x) = \begin{cases} f(x), & x \in \Delta_{[K-1]} \\ \infty, & x \notin \Delta_{[K-1]} \end{cases}. \]

Let \( \mathcal{E} \) be a random event, we use \( \mathbb{1}_\mathcal{E} \) to denote the indicator of \( \mathcal{E} \), which equals 1 if \( \mathcal{E} \) happens, and 0 otherwise.

III. PROBLEM SETTINGS

We now present the scale-free MAB problem with feedback delays. Formally speaking, there are \( K \geq 2 \) available arms indexed from 1 to \( K \), and \( T \geq 1 \) time slots for the agent to make decisions sequentially. The loss matrix \( L \in \mathbb{R}^{T \times K} \) and delay sequence \( \{d_t\}_{t=1}^T \) are both fixed before the game starts but unknown to the agent (i.e., both \( L \) and \( \{d_t\}_{t=1}^T \) are both obliviously adversarially chosen).

At the beginning of each time slot \( t \), the agent needs to choose an action \( i_t \in [K] \). It will receive a feedback \( (t, \ell_{t,i_t}) \) for this action \( i_t \) at the end of the \( t + d_t \)-th time slot. The agent is allowed to make the decision \( i_t \) based on its history actions \( \{i_s\}_{s=1}^{t-1} \), all arrived feedback \( \{(s, \ell_{s,i_s}) \mid s + d_s < t\} \) and a private randomness. Note that the agent can infer \( d_t \) only after the feedback of \( i_t \) arrives, i.e., at the end of time slot \( t + d_t \).

The objective of the agent is to minimize the difference between its total loss and the minimum possible total loss sticking to a single action. Formally, the objective is to minimize the pseudo-regret (or just regret for simplicity) defined as follows:

**Definition 1** (Pseudo-regret). We define

\[ R_T \triangleq \min_{i \in [K]} \mathbb{E}_{\{F_t\}} \left[ \sum_{t=1}^T \ell_{t,i_t} - \sum_{t=1}^T \ell_{t,i} \right] \]

to be the pseudo-regret of an MAB algorithm.

For simplicity, denote the total delay as \( D \triangleq \sum_{t=1}^T d_t \) and the \( \infty \)-norm of all losses as \( L \triangleq \max_{1 \leq t \leq T} \| \ell_t \|_{\infty} \), which is finite but unknown to the agent. We use \( \{F_t\} \) to denote the natural filtration of \( \{i_1, i_2, \ldots, i_T\} \), i.e., \( F_t = \sigma(i_1, i_2, \ldots, i_t) \).

IV. ALGORITHMS

In this section we present two versions of our SFBanker algorithms for delayed scale-free adversarial MABs. The first algorithm SFBanker-TINF (IV-A) is applicable when the player has a prior knowledge that all losses are non-negative, i.e., the player is facing a non-negative-loss instance. The second algorithm SFBanker-LBINF (IV-B) applies to general cases with potentially negative losses and enjoys a small-loss style regret bound, which induces a universal regret upper-bound that is only \( O(\sqrt{\log T + \log L}) \) times worse than the former one.

**A. SFBanker-TINF for Non-Negative Losses**

First we consider the case where the loss vectors \( \ell_t \) are all non-negative, i.e., \( \ell_{t,i} \geq 0 \) for any \( 1 \leq i \leq K \). For this case, we develop a novel algorithm named SFBanker-TINF. It builds upon the Banker-OMD framework with GreedyPick volume combination planner ([Huang and Huang, 2021]), and uses the \( \Psi(x) = -2 \sum_{i=1}^K \sqrt{x_i} \), i.e., the \( \frac{2}{\gamma} \)-Tsallis entropy function ([Abernethy et al., 2013]) as the regularizer, which is commonly used in MAB algorithms, and leads to many optimal or nearly optimal results (e.g., [Zimmert and Seldin, 2019], [Zimmert and Seldin, 2020]). The pseudo-code is presented below.

**Algorithm 1** SFBanker-TINF

---

| Algorithm          | Loss Range | Prior Knowledge | Total Regret Upper-bound |
|--------------------|------------|-----------------|--------------------------|
| [Zimmert and Seldin, 2020] | \( L \subseteq [0,1)^K \) | None | \( O(\sqrt{KT} + \sqrt{D\log K}) \) |
| [Putta and Agrawal, 2021] | \( L \subseteq \mathbb{R}^K \) | \( D = 0 \) | \( O(\sqrt{KL_2 + \sqrt{KL_1}}) \) |
| SFBanker-TINF (ours) | \( L \subseteq [0, +\infty)^K \) | None | \( O \left( \sqrt{K(D + T)\log(D + T)L} \right) \) |
| SFBanker-LBINF (ours) | \( L \subseteq \mathbb{R}^K \) | \( T \) (non-essential) | \( \mathcal{O} \left( \sqrt{K(D + T)\log(D + T)L + \sqrt{DL}\log L} \right) \) |
Input: Number of arms $K$, $\frac{1}{T}$-Tsallis entropy regularizer $\Psi$, default investment choice $x_0 = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$, Time-horizon length $T$.

Output: Sequence of actions $i_1, i_2, \ldots, i_T \in [K]$.

1: $B_0 \leftarrow 0, \hat{L}_1 \leftarrow 1$

2: for $t = 1, 2, \ldots, T$ do

3: $a_t \leftarrow 0$ \quad $\triangleright$ Mark time step $t$ as not skipped and feedback not arrived

4: $v_t \leftarrow (\sum_{s=1}^{t} \mathbb{1}_{a_t = 0})$ \quad $\triangleright$ the current number of pulls whose feedback has not arrived

5: $D_t \leftarrow \sum_{s=1}^{t} \mathbb{1}_{a_t = 0} (\sigma_s + 1) \hat{L}_s^2$

6: $\sigma_t^{-1} \leftarrow (\hat{L}_t + 1) \frac{\ln(3 + D_t \hat{L}_t^2)}{3 + D_t} \quad \triangleright$ Decide the scale for $t$

7: $v_t \leftarrow \sigma_t, b_t \leftarrow \sigma_t$

8: for $s = 1, 2, \ldots, t - 1$ do

9: \textbf{if} $a_s = 1$ then

10: \textbf{end if}

11: $m_{t,s} \leftarrow \min\{v_s, b_t\}$

12: $u_s \leftarrow v_s - m_{t,s}, b_t \leftarrow b_t - m_{t,s}$

13: end do

14: $B_t \leftarrow B_{t-1} + b_t$

15: $x_t \leftarrow -\nabla \Psi^*\left(\sum_{s=1}^{t} \mathbb{1}_{a_s = 0} m_{s,t} \nabla \Psi(z_s) + \sum_{i=1}^{s} \mathbb{1}_{a_i = 0} \frac{b_i}{\hat{L}_i}\right)$

16: Sample $i_s$ according to $x_t$, pull the arm $i_s$

17: $\tilde{L}_{t+1} \leftarrow \hat{L}_t$

18: for each new feedback $(s, \ell_{s,i}, a_s)$ do

19: $\ell_{s,i} \leftarrow \begin{cases} \ell_{s,i} & i = i_s, \forall i \in [K] \\ 0 & i \neq i_s \end{cases}$

20: $z_s \leftarrow -\nabla \Psi^*\left(\nabla \Psi(x_s) - \frac{1}{\hat{L}_t}\right)$

21: $a_s \leftarrow 1$ \quad $\triangleright$ Mark time slot $s$ as received

22: if $|\ell_{s,i}| > \tilde{L}_s$ then

23: \textbf{end if}

24: $a_s \leftarrow 2$ \quad $\triangleright$ Mark time slot $s$ as skipped

25: $B_s \leftarrow B_t - b_s$ \quad $\triangleright$ Revert spent savings

26: for $u = 1, 2, \ldots, s - 1$ do

27: $v_u \leftarrow v_u - m_{t,u}, u_u \leftarrow u_u + m_{t,u}$

28: \textbf{end for}

29: if $2|\ell_{s,i}| \geq \tilde{L}_{s+1}$ then

30: $\tilde{L}_{s+1} \leftarrow 2|\ell_{s,i}|$

31: \textbf{end if}

32: end for

33: \textbf{end for}

Remark. From [Cesa-Bianchi et al., 2016], we know that any algorithm in the delayed adversarial MAB with $L = [0,1]^K$ has a regret lower-bound $\Omega(\sqrt{KT + D\log K})$. Therefore, in the scale-free setting, even if $L$ is known beforehand, any reasonable algorithm must suffer a total regret of $\Omega((\sqrt{KT + D\log K})L)$ (a directly corollary of the lower-bound in [Cesa-Bianchi et al., 2016] and [Zimmert and Seldin, 2020]). Thus, if we regard $K$ as a constant independent of $T$, Algorithm 1 is only $O(\sqrt{D + T})$ times worse than the theoretical lower-bound.

B. SFBanker-LBINF for General Losses

Similar to SFBanker-TINF, the proposed novel algorithm SFBanker-LBINF for the general-loss case also builds upon the Banker-OMD framework with GreedyPick volume combination planner ([Huang and Huang, 2021]).

In SFBanker-LBINF, We choose to use another regularizer $\Psi(x) = -\sum_{i=1}^{K} \ln x_i$, i.e., the log-barrier function ([Abernethy et al., 2015]). Compared to other regularizers, it tends to allocate more exploration probabilities to all arms ([Agarwal et al., 2017]) and leads to various adaptivity properties ([Wei and Luo, 2018]). It also plays a crucial role when deriving a small-loss style regret bound for our algorithm.

The pseudo-code is presented below.

Algorithm 2 SFBanker-LBINF

Input: Number of arms $K$, log-barrier regularizer $\Psi$, default investment choice $x_0 = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$, Time-horizon length $T$.

Output: Sequence of actions $i_1, i_2, \ldots, i_T \in [K]$.

1: $B_0 \leftarrow 0, \hat{L}_1 \leftarrow 1, \mathcal{D}_0 \leftarrow 1$

2: for $t = 1, 2, \ldots$ do

3: $a_t \leftarrow 0$

4: $\mathcal{D}_t \leftarrow \sum_{s=1}^{t} \mathbb{1}_{a_s = 0}$

5: $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} + \mathcal{D}_t$ \quad $\triangleright$ Experienced total delay up to time $t$

6: $D_t \leftarrow 0$

7: for $s = 1, 2, \ldots$ do

8: \textbf{if} $a_s = 1$ then

9: $D_t \leftarrow D_t + (\sigma_s + 1) \hat{L}_s^2$

10: \textbf{else} if $a_s = 0$ then

11: $D_t \leftarrow D_t + (\sigma_s + 1) \hat{L}_s^2$

12: \textbf{end if}

13: $\sigma_t \leftarrow (\hat{L}_t + 1) \frac{\ln(3 + D_t \hat{L}_t^2)}{3 + D_t} \quad \triangleright$ Decide the scale for $t$

14: if $\mathcal{D}_t \leq \frac{\hat{L}_t}{K}$ then

15: $\alpha_t \leftarrow \max\{\sigma_t, 2\hat{L}_t\}$

16: $v_t \leftarrow \sigma_t, b_t \leftarrow \sigma_t$

17: for $s = 1, 2, \ldots, t$ do

18: \textbf{if} $a_s = 1$ then

19: $m_{t,s} \leftarrow \min\{v_s, b_t\}$

20: $v_s \leftarrow v_s - m_{t,s}, b_t \leftarrow b_t - m_{t,s}$

21: \textbf{end if}

22: \textbf{end if}

23: \textbf{end for}

The performance of SFBanker-TINF is as follows.

Theorem 2 (Regret of Algorithm 1). When losses are non-negative, the regret of Algorithm 1 is:

$$R_T \leq O(\sqrt{KT(D + T)} \log(D + T) \cdot L).$$
where feedback that have not arrived at the beginning of time-step $d$ O\[\text{[Putta and Agrawal, 2021]}\]'s result.

**Theorem 3**

Theorem 3.  

**Algorithm 3** satisfies:

$$R_T \leq O \left( \sqrt{K \log(D + T) \log T} \mathbb{E}[\bar{L}^2_T] + \sqrt{DL \log L} + \sqrt{K D \log TL} \right)$$

where $\bar{L}^2_T \triangleq \sum_{t=1}^{T}(d_t + 1)\ell^2_{t,i}$, $d_t$ denotes the number of feedback that have not arrived at the beginning of time-step $t$. In particular,

$$R_T \leq O \left( \sqrt{K (D + T) \log (D + T) \log L} + \sqrt{DL \log L} \right).$$

**Remark.** When running on problem instances with $d_t \equiv 0$ (i.e., non-delayed case), SFBanker-LBINF enjoys $O(\sqrt{K \log T} \mathbb{E}[\bar{L}^2_T] + L \log L + \sqrt{K D \log TL}) \[\text{[Putta and Agrawal, 2021]}\]$ and $O(\sqrt{KT \log TL + L \log L})$ regret guarantee, which improves $\[\text{[Putta and Agrawal, 2021]}\]$'s result.

**V. ANALYSIS OF SFBanker-TINF**

We now present the analysis of Algorithm 3. For a detailed proof, please check Appendix A.

A. High-level Ideas

To bound the regret $R_T$, we can bound $\mathbb{E}[|\ell_t, x_t - y|]$ for any $y \in \Delta_{[K]}$, which equals $\mathbb{E}[|\ell_t, x_t - y|]$ by property of weighted importance samplers (Lemma 13 in appendix). We leverage the following lemma from the Banker-OMD framework in $\[\text{[Huang and Huang, 2021]}\]$:

**Lemma 4** (Theorem 3 in $\[\text{[Huang and Huang, 2021]}\]$). For any $y \in \Delta_{[K-1]}$, we have

$$\sum_{t=1}^{T}(\ell_t, x_t - y) \leq B_T \cdot D_\psi(y, x_0) + \sum_{t=1}^{T} \sigma_t D_\psi(x_t, \tilde{z}_t) - \sum_{t=1}^{T} v_t D_\psi(y, z_t)$$

where $\tilde{z}_t = \nabla^* (\nabla (x_t) - \frac{1}{\sigma_t} \ell_t)$ for any $1 \leq t \leq T$.

Since Bregman divergences are non-negative, we focus on the first two terms on the right hand side, which are called the total investment and immediate cost terms in the Banker-OMD language, as we will formally define in Lemma 7.

We first present a lemma for the immediate cost terms.

**Lemma 5.** For any $x_t$ in the interior of $\Delta_{[K-1]}$, $\sigma_t > 0$ and $\ell_t \in \mathbb{R}^K$ such that $\tilde{z}_t := \nabla^* (\nabla (x_t) - \frac{1}{\sigma_t} \ell_t)$ is properly defined, we have

$$\sigma_t D_\psi(x_t, \tilde{z}_t) = \frac{1}{2} \sigma_t^{-1} \left| \ell_t \right|^2_{\psi^\alpha(w_t)}$$

where $w_t = \nabla (x_t) - \frac{\theta}{\sigma_t} \tilde{z}_t$ for some $\theta \in (0, 1)$.

When $\psi$ is the $\frac{1}{2}$-Tsallis entropy function, and the losses we are dealing are all non-negative, due to the duality property between $\Psi$ and $\psi^*$ and the monotonicity of the diagonal elements of $\nabla^2 \Psi$, we have the following lemma:

**Lemma 6.** In Algorithm 3 for any $t \in [T]$, we have

$$\mathbb{E}[\sigma_t D_\psi(x_t, \tilde{z}_t) | F_{t-1}] \leq \frac{1}{\sigma_t} \sum_{i=1}^{K} \ell^2_{t,i} x^2_{t,i} \leq \frac{\sqrt{K}}{\sigma_t} \left| \ell_t \right|_{\infty}^2.$$
resulting in an $\tilde{O}(\sqrt{K(D+T)L})$ regret bound by the fact that $D_T \leq (D+T)L^2$.

In the actual situation without such $\|\ell_t\|_\infty$ oracle, our rescue is to introduce a doubling trick and maintain an upper bound $\hat{L}$ of the maximum feedback $\ell_{t,i_t}$ we have ever seen. We use $\hat{L}$ as an estimate of the incoming $\|\ell_t\|_\infty$ in the ideal $\sigma_t$ choice, pretending that we have an $\|\ell_t\|_\infty$ oracle.

However, if it turns out that we underestimated the real losses, i.e., we receive a feedback $\ell_{t,i_t}$ exceeding our current upper bound $\hat{L}$, then we set the upper bound to $2\hat{L}_{t,i_t}$ and skip that time slot while bounding the regret of that time stamp by $\ell_{t,i_t}$, which does not exceed the new upper-bound.

After the skipping trick is applied, intuitively we gain access to a “weaker oracle” capable of returning an correct upper-bound for the current $\|\ell_t\|_\infty$ within $O(L)$, which is sufficient for achieving $O(\sqrt{(T+D)})$-style total regret.

To summarize the above idea, let $\xi_t$ be the event that $a_t > 2$ when the algorithm terminates, that in Algorithm 1 $a_t$ is the flag to denote whether the $t$-th time step’s feedback arrives and whether this time step has been skipped. Hence $\xi_t$ contains all sample paths that the $t$-th time step is not skipped by Algorithm 1. If we bound regret of each skipped step simply by the suffered losses (recall we are dealing with a non-negative problem instance) and apply Lemma 4 for remaining steps, we obtain the following lemma:

**Lemma 7.** For Algorithm 1 we have

\[
\mathcal{R}_T \leq \max_{y \in \Delta(K-1)} \mathbb{E}[B_T \cdot D_y(x_t, x_0)]
\]

\[
= \text{Total Investment Term} + \sum_{t=1}^{T} \mathbb{E}[\xi_t \sigma_t D_y(x_t, \tilde{z}_t)] + \sum_{t=1}^{T} \mathbb{E}[\mathbbm{1}_{\neg \xi_t} \hat{L}_{t,i_t}].
\]

(1)

In the remaining of this section, we give the key steps for bounding the three parts in (1).

**B. Total Investment Term**

To bound the total investment term $B_T D_y(x_t, x_0)$, we leverage the following lemma from Banker-OMD:

**Lemma 8 (Lemma 5 in [Huang and Huang, 2021]).** For any time index $1 \leq t \leq T$, there exists some $t_0 \leq t$ such that $B_t = B_{t_0} = \sigma_{t_0} + \sum_{s=1}^{t_0-1} \mathbb{1}[s+d_s \geq t] \sigma_s$.

Now apply Lemma 8 for $t = T$. Further let $t_1 \geq t_2 \geq \cdots \geq t_m$ be the time indices such that $s + d_s = t_0$, i.e., $t_1, t_2, \ldots, t_m$ have their feedback still on the way at the beginning of time $t_0$. By the fact that $\sum_{i=1}^{m} d_{i} \leq D$ and $d_t \geq t_{0} - t_i \geq i$, we have $m = O(\sqrt{D})$. Moreover, at time $t_i$, $\sigma_{t_i} \geq m - i$, as feedback of $t_{i+1}, \ldots, t_m$ are still on the way. Hence,

\[
B_T \leq \frac{1}{m+1} \sqrt{\frac{3 + D_{t_0}}{\ln(3 + D_{t_0}/L^2_{t_0})}} + \sum_{i=1}^{m} \frac{1}{i} \max_{1 \leq s \leq i} \sqrt{\frac{3 + D_t}{\ln(3 + D_t/L^2_t)}}.
\]

Using the fact that $\sum_{s=1}^{t} (d_s + 1) \leq D + T$ (Lemma 15 in appendix) and monotonicity of $\frac{1}{\ln x}$ (Lemma 17 in appendix), we can bound $B_T$ by

\[
B_T \leq O(\log D) \sqrt{\frac{(D+T)L^2}{\ln(D+T)}} = O(\sqrt{(D+T) \log DL}).
\]

Moreover, by the fact that $D_y(y_0) = O(\sqrt{K})$ (Lemma 15 in appendix), the total investment term is therefore bounded by $O(\sqrt{K(D+T) \log DL})$.

**C. Immediate Cost Term**

For any $1 \leq t \leq T$, we have

\[
E_{i_t \sim x_t} [\xi_t \sigma_t D_y(x_t, \tilde{z}_t)] \leq \max_{i_t \sim x_t} \left[ E_{i_t \sim x_t} \hat{L}^2_{t,i_t} \right] \sigma_t
\]

\[
= \max_{i_t \sim x_t} \left[ E_{i_t \sim x_t} \hat{L}^2_{t,i_t} \right] \sigma_t
\]

\[
\leq O(\sqrt{\hat{L}^2_{t,i_t}} \sigma_t)
\]

where the first inequality applies Lemma 6 and the last step is due to Cauchy-Schwartz inequality.

Define

\[
\hat{D}_t \equiv (d_t + 1) \hat{L}^2 + \sum_{s \in \{1, \ldots, t-1\}} (d_s + 1) \hat{L}^2_s,
\]

and let $\tilde{\sigma}_t = \sqrt{\frac{\ln(3+D_t)/L^2_t}{\hat{D}_t}}$. Compared to the quantity $D_t$ we use in Algorithm 1 $\hat{D}_t$ only contains previous time-steps that are not skipped. Denote by $T$ the random subset $\{t \in [T] : \xi_t\}$. We have

\[
\sum_{t=1}^{T} \mathbb{E}[\xi_t \sigma_t D_y(x_t, \tilde{z}_t)]
\]

\[
\leq \sqrt{K} \mathbb{E} \left[ \sum_{t=1}^{T} \frac{\hat{L}^2_t}{\tilde{\sigma}_t} \right] \leq \sqrt{K} \mathbb{E} \left[ \sum_{t \in T} \frac{\hat{L}^2_t}{\sigma_t} \right]
\]

\[
\leq \sqrt{K} \ln(3 + (D + T)) \cdot \mathbb{E} \left[ \sum_{t \in T} \frac{(d_t + 1) \hat{L}^2_t}{\sqrt{3 + \sum_{s \in T,s \leq t} (d_s + 1) \hat{L}^2_s}} \right].
\]

(2)

It turns out the last factor can be bounded using the following “square-root summation lemma”:

**Lemma 9 (Lemma 3.5 in [Auer et al., 2002].)** Let $x_1, x_2, \ldots, x_T$ be non-negative real numbers, then

\[
\sum_{t=1}^{T} \frac{x_t}{1 + \sum_{s=1}^{t} x_s} \leq 2 \sqrt{\frac{1 + \sum_{t=1}^{T} x_t}.}
\]

Lemma 9 states that the quantity to take expectation in (2) is $O\left(\sqrt{\sum_{t \in T} (d_t + 1) \hat{L}^2_t}\right) = O(\sqrt{(D+T)L})$. Therefore,

\[
\sum_{t=1}^{T} \mathbb{E}[\xi_t \sigma_t D_y(x_t, \tilde{z}_t) \leq O(\sqrt{K(D+T) \log(D+T)L}).
\]
D. Skipping Regret

Finally, let us look at the set of skipped time steps $\mathcal{T} \triangleq \{t \in [T] : \hat{z}_t^T \geq \sigma_t \}$. We have the following observation:

**Lemma 10.** Denote $\ell_{t,i}$ by $\hat{\ell}_t$. For any indices $s, t \in \mathcal{T}$, if $\hat{\ell}_t < 2\hat{\ell}_s$, we have $s + d_s \geq t$, i.e., the feedback of action $i$ arrives later than the decision is being made.

Suppose $|\mathcal{T}| = m$ and denote by $t_1, t_2, \cdots, t_m$ the elements in $\mathcal{T}$, which are in the arrival order of their feedback. Consider partitioning $\mathcal{T}$ into some non-empty subsets $\mathcal{T}_i \triangleq \{t_1, t_2, \cdots, t_m\}$, $\mathcal{T}_0 \triangleq \{t_{m+1}, t_{m+2}, \cdots, t_m\}$. Therefore, $|\mathcal{T}_i| \leq d_{t_{m+1}} + 1 \leq O(\sqrt{D})$.

By making these observations, one can bound the total losses incurred by time steps in $\mathcal{T}$. For each $\mathcal{T}_i$,

$$\sum_{t \in \mathcal{T}_i} \hat{\ell}_t \leq 2|\mathcal{T}_i|\hat{\ell}_{t_{m+1}} \leq 4 \cdot 2^{-(k-i)}|\mathcal{T}_i|\hat{\ell}_{t_{m+1}}$$

$$\leq 4 \cdot 2^{-(k-i)} \cdot O(\sqrt{D}) \cdot L$$

Thus the total loss incurred by $\mathcal{T}$ is bounded by $O(\sqrt{D})$. By putting everything into Lemma 7, we have $R_T \leq O(\sqrt{K(D + T)} \log D)$, as desired.

VI. ANALYSIS OF SFBanker-LBINF

Notice that Algorithm 2 has a quite similar framework to Algorithm 1. Therefore, here we only emphasize the main differences between the design and regret analysis between SFBanker-LBINF and SFBanker-TINF. A formal reasoning is presented in Appendix E.

A. High-level Ideas

When losses are no longer guaranteed to be non-negative, $\hat{z}_t$ in Lemma 5 is no longer component-wisely dominated by $x_t$. As a result, Lemma 3 no longer holds. Fortunately, we have the following lemma stating that, if the elements of $\hat{z}_t$ are not too large, we can continue to have a bound similar to Lemma 5.

**Lemma 11.** For any $x_t$ in the interior of $\Delta_{[K-1]}$, $\sigma_t > 0$, $\hat{\ell}_t \in \mathbb{R}^K$, let $\Psi$ be the log-barrier function, $\hat{z}_t = \nabla \Psi^*(\nabla\Psi(x_t) - \frac{1}{\sigma_t} \hat{\ell}_t)$, if for all $i \in [K]$ we have $\hat{z}_{t,i} \leq 2x_{t,i}$, then

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

*Here $\hat{S}$ denotes the completion of an event $S$.

If $\hat{\ell}_t$ is an importance sampling estimator used in OMD-based MAB algorithm, we have

$$\sigma_t D_\Psi(x_t, \hat{z}_t) \leq 2\sigma_t^{-1} \hat{\ell}_{t,i}^2.$$  

Intuitively, Lemma 11 suggests that it remains safe to apply single-step OMD regret upper-bounds when the intermediate result $\hat{z}_t$ does not get too large, which is automatically guaranteed when the actual suffered loss $\hat{\ell}_{t,i}$ is not too large compared with the action scale $\sigma_t$. This fact is formally stated as follows.

**Lemma 12.** If $\hat{\ell}_{t,i} \geq -\frac{1}{2} \sigma_t$, then $\hat{z}_{t,i} \leq 2x_{t,i}$.

Lemma 11 and Lemma 12 enable us to extend Algorithm 1 to the general case will potentially negative losses. First, we add another skipping criterion $\hat{\ell}_{t,i} < -\frac{1}{2} \sigma_t$ to Algorithm 2 at Line 5 so that Lemma 11 will be applicable to all time steps entering the OMD framework automatically.

Moreover, if we make an action scale $\sigma_t$ as large as at least $2L_t$, then bypassing the original $|\ell_{t,i}| \leq L_t$ criterion will be sufficient for Lemma 11 to apply to time $t$. At Line 16 we explicitly set $\sigma_t$ to at least $2L_t$ for some carefully selected time steps. We will show this policy can make both the total investment cost and skipped regret well-controlled.

B. Regret Decomposition

As the log-barrier regularizer will blow up at the boundary of $\Delta_{[K-1]}$, we need to compare our algorithm with a fixed action with positive probability mass on all the $K$ arms rather than one-a-hot vector. For example, say we want to upper-bound the total regret between Algorithm 2 and a static policy always choosing arm $k$. Then let $y = (1, 0, \ldots, 0)$, $y'$ be a $K$-dimensional vector will all coordinates equal to $1/T$, except that the $k$-th coordinate equals $1 - \Delta_{[K-1]}$. Both $y$ and $y'$ belong to $\Delta_{[K-1]}$, we can then write our expected regret with respect to arm 1 as

$$\mathbb{E} \left[ \sum_{t=1}^T (\ell_t x_t - y) \right] = \mathbb{E} \left[ \sum_{t=1}^T (\ell_t x_t - y') + (\ell_t y' - y) \right]$$

$$\leq \mathbb{E} \left[ \sum_{t=1}^T (\ell_t x_t - y') + \|\ell_t\|_\infty \cdot \|y' - y\|_1 \right]$$

$$\leq \mathbb{E} \left[ \sum_{t=1}^T (\ell_t x_t - y') \right] + O(KL).$$

Note that $D_\Psi(y', x)$ is properly defined for all $x$ in the interior of $\Delta_{[K-1]}$, so we can safely use OMD techniques to handle the sum.

Define $\Delta'_{[K-1]} = \{ x \in \Delta_{[K-1]} : x_i \geq 1/T, \forall i \in [K] \}$, we have the following upper-bound similar to (1).

**Lemma 13.** For Algorithm 2 we have

$$R_T \leq O(KL) + \max_{y \in \Delta'_{[K-1]}} \mathbb{E} \left[ B_T \cdot D_\Psi(y, x_0) \right]$$

$$+ \sum_{t=1}^T \mathbb{E} \left[ \|\ell_t\|_1 \cdot D_\Psi(x_t, \hat{z}_t) \right] + 2L \sum_{t=1}^T \mathbb{E} \|\ell_t\|_1. \quad (3)$$
In [3], the last term due to skipped time steps is different from the one in [1]. This is because when facing general losses, the regret incurred by an action \( \ell_{t,i} \) can be larger than \( \ell_{t,i} \), for the offline optimal action may suffer a negative loss value with magnitude up to \( L \).

In the remaining of this section, we emphasize on the changes needed to make comparing to Section V when building the three parts in [4].

C. Total Investment Term

When \( \Psi \) is the log-barrier function, we have \( \sup_{y \in \Delta_{[k-1]}} D \psi(y, x_0) \leq K \ln T \). The argument for bounding \( \tilde{B}_T \) in \( D, T, L \) can be adapted from [V-B] as follows. In Algorithm 2, we have \( \tilde{B}_T \leq 4(D + T)L^2 \) uniformly for all \( T \in [T] \). With \( t, m, t, \ldots, t_m \) defined in [V-B], we can bound \( B_T = \sum_{i=0}^{m} \sigma_i \) by

\[
\sum_{i=0}^{m} \sum_{t=0}^{1} \sigma_i \leq \frac{3 + \tilde{D}}{\ln(3 + \tilde{D}/L^2)} \tilde{D}_t + 1.
\]

\[
\leq O \left( \sqrt{D}L \right) + \frac{1}{\sqrt{K \ln T}} \sum_{i=0}^{m} \frac{1}{\ln(3 + \tilde{D}/L^2)} \tilde{D}_t + 1.
\]

\[
\leq O \left( \sqrt{D}L \right) + \Theta(\log m) \frac{1}{\sqrt{K \ln T}} \frac{3 + 4(D + T)L^2}{\log(3 + 4(D + T)L^2)}.
\]

\[
\leq O \left( \sqrt{D}L \right) + \frac{(D + T)L}{K \ln T}.
\]

where (a) comes from \( \sigma_i \geq i \) and \( \sqrt{D} \leq \sqrt{K} \).

Let \( \tilde{L}_2^2 \equiv \sum_{i=1}^{T} (\sigma_1 + \sigma_2 + \ldots + \sigma_M)^2 \), we can also obtain a similar bound for \( B_T \) in \( \tilde{L}_2^2 \). Let \( \bar{\sigma}_i = \max_{1 \leq t \leq T} \sigma_i \), we have \( \bar{\sigma}_i = O(\sqrt{D}) \) and \( \tilde{D}_T \leq \tilde{L}_2^2 + 4\bar{\sigma}_i^2L^2 \), which means \( \sqrt{\tilde{D}_T} \leq \sqrt{\tilde{L}_2^2 + 4\bar{\sigma}_i^2L^2} \), \( \forall t \in [T] \). Utilizing this \( \sqrt{\tilde{D}_T} \), we have

\[
B_T \leq O \left( \sqrt{D}L + \frac{\tilde{L}_2^2 \log D}{K \ln T} \right).
\]

D. Skipping Regret

An argument exactly identical to [V-D] will indicate that, the total number of time steps skipped due to the criterion \( \{ |t_{t+i} | > L_t \} \) is \( O(\sqrt{D} \log L) \).

It remains to control the total number of time steps skipped due to \( \{ |t_{t,i} | \leq \tilde{L}_t, \ell_{t,i} < -\frac{1}{2} \phi_t \} \). Note that this can happen only if \( \phi_t > \sqrt{\frac{D}{K}} \), hence once such skipping of time \( t \) occurs, we have \( \tilde{D}_t+1 \geq \tilde{D}_t+\sqrt{\frac{D}{K}} \). Therefore, consider the following discrete time analog of the dynamics \( \hat{x} = \sqrt{\frac{D}{K}} \hat{x} + \hat{f}_t = 1 \) and \( \hat{f}_t = \phi_{t-1} + \sqrt{\frac{D}{K}} \hat{f}_t \); then the number of these skipped steps is bounded by \( O(\sup \{ t \geq 1 : \hat{f}_t \leq D \}) \), which can be further bounded to \( O(\sqrt{K D}) \) (see Appendix E).

Therefore, the total regret incurred by skipping time steps is no more than \( O(\sqrt{D}L \log L + \sqrt{K \tilde{D}}})).

E. Immediate Cost Terms

Let \( \tilde{D}_t \equiv \sum_{i=1}^{t} (\sigma_1 + \sigma_2 + \ldots + \sigma_M)^2 \), and \( \bar{\sigma}_t = (\sigma_1 + 1)^{\frac{1}{2}} \sqrt{\frac{ln(3 + \bar{\sigma}_t^2)}{3 + \bar{\sigma}_t^2}} \). In Algorithm 2 our choice of \( \sigma_t \) guarantees \( \sigma_t \geq 2 \tilde{L}_t \), hence any time step \( t \) that passes the doubling trick skipping criterion \( |t_{t,i} | \leq \tilde{L}_t \) will automatically satisfy the condition to apply Lemma 1. Therefore, we have

\[
\sum_{t=1}^{T} \sum_{s \in T} \sigma_t \sigma_2 \leq \sum_{t=1}^{T} \sum_{s \in T} \sigma_t \sigma_2 \leq \frac{\tilde{L}_2^2}{\tilde{D}_T} \sum_{s \in T} \sigma_t \sigma_2 \leq O(\sqrt{K \log(D + T) \log L} \log L).
\]

We can bound \( \tilde{D}_t \) in \( T, D, L \) by \( \tilde{D}_t \leq O((D + T)L^2) \), or bound it in \( \tilde{L}_2^2 \) as we did in Section VI-C

\[
\tilde{D}_T \leq \tilde{L}_2^2 + |T| \cdot 4L^2 \leq \tilde{L}_2^2 + O(\sqrt{D} \log L \cdot L^2).
\]

Thus we can conclude that

\[
\sum_{t=1}^{T} \sum_{s \in T} \sigma_t \sigma_2 \leq \min \left\{ \frac{O(\sqrt{K(D + T) \log(D + T) \log L})}{O(\sqrt{K \log(D + T) \log L \tilde{L}_2^2} + D^2 \log L \log L)} \right\}
\]

Putting everything together gives Theorem 3.

VII. CONCLUSION

We design two algorithms SFBanker-TINF and SFBanker-LBINF for the novel delayed scale-free adversarial MAB problem. The proposed algorithm recovers \( O(\sqrt{T} + \sqrt{D}) \) style regret bounds in existing works on delayed adversarial MAB with standard losses (e.g., [Bistritz et al., 2019], [Thune et al., 2019], [Zimmert and Seldin, 2020]), and it achieves even slightly better regret bound on non-delayed instances comparing to the previous work on scale-free adversarial MAB ([Putta and Agrawal, 2021]).
REFERENCES

[Abernethy et al., 2015] Abernethy, J. D., Lee, C., and Tewari, A. (2015). Fighting bandits with a new kind of smoothness. Advances in Neural Information Processing Systems, 28:2197–2205.

[Agarwal et al., 2017] Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. (2017). Corralling a band of bandit algorithms. In Conference on Learning Theory, pages 12–38. PMLR.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. Journal of Computer and System Sciences, 64(1):48–75.

[Bistritz et al., 2019] Bistritz, I., Zhou, Z., Chen, X., Bamdos, N., and Blanchet, J. (2019). Exp3 learning in adversarial bandits with delayed feedback. Advances in neural information processing systems.

[Cesa-Bianchi et al., 2016] Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. (2016). Delay and cooperation in nonstochastic bandits. In Conference on Learning Theory, pages 605–622. PMLR.

[De Rooij et al., 2014] De Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. The Journal of Machine Learning Research, 15(1):1281–1316.

[Hadiji and Stoltz, 2020] Hadiji, H. and Stoltz, G. (2020). Adaptation to the range in k-armed bandits. arXiv preprint arXiv:2006.03378.

[Huang and Huang, 2021] Huang, J. and Huang, L. (2021). Banker online mirror descent. arXiv preprint arXiv:2106.08943.

[Orabona and Pál, 2018] Orabona, F. and Pál, D. (2018). Scale-free online learning. Theoretical Computer Science, 716:50–69.

[Putta and Agrawal, 2021] Putta, S. R. and Agrawal, S. (2021). Scale free adversarial multi armed bandits. arXiv preprint arXiv:2106.04700.

[Thune et al., 2019] Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. (2019). Nonstochastic multiaimed bandits with unrestricted delays. arXiv preprint arXiv:1906.00670.

[Wei and Luo, 2018] Wei, C.-Y. and Luo, H. (2018). More adaptive algorithms for adversarial bandits. In Conference On Learning Theory, pages 1263–1291. PMLR.

[Zimmert and Seldin, 2019] Zimmert, J. and Seldin, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 467–475. PMLR.

[Zimmert and Seldin, 2020] Zimmert, J. and Seldin, Y. (2020). An optimal algorithm for adversarial bandits with arbitrary delays. In International Conference on Artificial Intelligence and Statistics, pages 3285–3294. PMLR.
APPENDIX

Lemma 14. For any distribution \( x \in \triangle_{[K]} \) and loss vector \( \ell \in \mathbb{R}^K \), if we pulled an arm \( i \) according to \( x \), then the weighted importance sampler \( \hat{\ell}(j) \) gives an unbiased estimate of \( \ell(j) \), i.e.,

\[
E_{i \sim x} \left[ \hat{\ell}(j) \right] = \ell(j), \quad \forall 1 \leq j \leq K.
\]

Proof. As the adversary is oblivious, the expectation is only taken with respect to \( x \). Hence,

\[
E_{i \sim x} \left[ \hat{\ell}(j) \right] = \sum_{i=1}^{K} \frac{\ell(j)}{x_j} \mathbb{1}_{i=j} \cdot \text{Pr}\{i\} = \text{Pr}\{i = j\} \frac{\ell(j)}{x_j} = \ell(j),
\]

for any \( 1 \leq j \leq K \). \( \square \)

Lemma 15. Let \( \Psi(x) = -2 \sum_{i=1}^{K} \sqrt{x_i} \) be the \( \frac{1}{2} \)-Tsallis entropy function. Then for any \( y \in \triangle_{K-1} \) and \( x_0 = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}) \in \triangle_{[K-1]} \), we have

\[
D_{\Psi}(y, x_0) = O(\sqrt{K}).
\]

Proof. When \( \Psi \) is the \( \frac{1}{2} \)-Tsallis entropy function, we have \( \nabla \Psi(x) = (-1/\sqrt{x_1}, -1/\sqrt{x_2}, \ldots, -1/\sqrt{x_K})^T \). Hence, the Bregman divergence between \( y \) and \( x_0 \) with respect to \( \Psi \) is bounded by

\[
D_{\Psi}(y, x_0) = \Psi(y) - \Psi(x_0) - \langle \nabla \Psi(x_0), y - x \rangle = -2 \sum_{i=1}^{K} \sqrt{y_i} + 2 \sum_{i=1}^{K} \frac{1}{\sqrt{K}} + \sum_{i=1}^{K} \frac{y_i - 1}{\sqrt{K}}
\]

\[
\leq 2\sqrt{K} + \sqrt{K} \sum_{i=1}^{n} y_i - K \frac{\sqrt{K}}{K} = 2\sqrt{K} = O(\sqrt{K}).
\]

Hence, \( D_{\Psi}(y, x_0) = O(\sqrt{K}) \) for any \( y \in \triangle_{[K-1]} \). \( \square \)

Lemma 16. In Algorithm 1 and 2, the total observed delay \( \sum_{t=1}^{T} \varnothing_t \) is equal to \( D + T \). Moreover, we have \( \sum_{s=1}^{t} (\varnothing_s + 1) \leq D + T \) for any \( 1 \leq t \leq T \).

Proof. By definition of \( \varnothing_t \), we have

\[
\sum_{t=1}^{T} \varnothing_t = \sum_{t=1}^{T} \sum_{s=1}^{t-1} \mathbb{1}[s + d_s \geq t] = \sum_{s=1}^{T} \sum_{t=s+1}^{T} \mathbb{1}[s + d_s \geq t] = \sum_{s=1}^{T} d_s = D,
\]

as claimed. The second claim is simply due to the fact that all \( \varnothing_t \)'s are non-negative. \( \square \)

Lemma 17. The function \( f(x) = \frac{1}{\log x} \) is monotonically increasing when \( x \geq e \).

Proof. We have \( f'(x) = \frac{1}{\log x} - \frac{1}{(\log x)^2} \geq 0 \) when \( x \geq e \), so \( f(x) \) is monotonically increasing when \( x \geq e \). \( \square \)

A. Omitted Proofs in Section V-A

We first give formal proofs for Lemma 5, Lemma 6 and Lemma 7 presented in Section V-A.

Proof of Lemma 5. Due to the duality property of Bregman Divergences, we have

\[
\sigma_t D_{\Psi}(x_t, \tilde{z}_t) = \sigma_t D_{\Psi^*}(\nabla \Psi(\tilde{z}_t), \nabla \Psi(x_t))
\]

\[
= \sigma_t D_{\Psi^*} \left( \nabla \Psi(x_t) - \frac{1}{\sigma_t} \tilde{\ell}_t, \nabla \Psi(x_t) \right)
\]

\[
= \Psi^* \left( \nabla \Psi(x_t) - \frac{1}{\sigma_t} \tilde{\ell}_t \right) - \Psi^* \left( \nabla \Psi(x_t) \right) - \left\langle \nabla \Psi^* \left( \nabla \Psi(x_t) \right), \frac{1}{\sigma_t} \tilde{\ell}_t \right\rangle,
\]

which can be seen as a second-order Lagrange remainder of \( \Psi^* \) between \( \nabla \Psi(x_t) - \frac{1}{\sigma_t} \tilde{\ell}_t \) and \( \nabla \Psi(x_t) \). Hence according to the mean value theorem, we have

\[
\sigma_t D_{\Psi}(x_t, \tilde{z}_t) = \frac{1}{2} \sigma_t \left\| \nabla \Psi^* (w_t) \right\|^2 \|
\]

\[
= \frac{1}{2} \sigma_t \left\| \tilde{\ell}_t \right\|^2 \left\| \nabla \Psi^* (w_t) \right\|^2,
\]

where \( w_t \) is some \( K \)-dimensional vector lying between the line segment connecting \( \nabla \Psi(x_t) - \frac{1}{\sigma_t} \tilde{\ell}_t \) and \( \nabla \Psi(x_t) \). This is exactly what we want. \( \square \)
Proof of Lemma 8 When $\Psi$ is the $\frac{1}{2}$-Tsallis entropy function and $\tilde{\ell}_t$ is the importance sampling estimator in Algorithm 1, according to Lemma 5, we have

$$\sigma_t D_\Psi(x_t, z_t) = \frac{1}{2\sigma_t} \left\| \tilde{\ell}_t \right\|_{\nabla \Psi^*(w_t)}^2$$

$$= \frac{1}{2\sigma_t} \left\| \tilde{\ell}_t \right\|_{\nabla \Psi^*(\nabla \Psi(w_t))^{-1}}^2$$

$$= \frac{1}{2\sigma_t} \left( \frac{l_{t,i}}{x_{t,i}} \right)^2 \cdot \left( \frac{1}{2} (w_{t,i}^2) - \frac{3}{2} \right)^{-1}$$

where $w_t = \nabla \Psi^*(\nabla \Psi(x_t) - \frac{\theta}{\sigma t} \tilde{\ell}_t)$ for some $\theta \in (0, 1)$. In this particular case we have $-w_{t,i}^2 = -x_{t,i}^2 - \frac{1}{\sigma_t} \tilde{\ell}_{t,i}$, which leads to

$$\sigma_t D_\Psi(x_t, z_t) \leq \frac{1}{\sigma_t} \left( \frac{l_{t,i}}{x_{t,i}} \right)^2 \cdot x_{t,i}^2 \leq \frac{1}{\sigma_t} l_{t,i}^2.$$ 

Taking expectation with respect to $F_{t-1}$ to both sides, we get

$$\mathbb{E}[\sigma_t D_\Psi(x_t, z_t) \mid F_{t-1}] \leq \frac{1}{\sigma_t} \left( \frac{l_{t,i}}{x_{t,i}} \right)^2 \cdot x_{t,i}^2$$

$$\leq \frac{1}{\sigma_t} \sum_{i=1}^{K} l_{t,i}^2 x_{t,i}^2 \leq \frac{\sqrt{K} \left\| \ell_t \right\|_\infty}{\sigma_t}$$

where the last step is due to Cauchy-Schwartz inequality. □

Proof of Lemma 7 By definition of the pseudo-regret, we have

$$\mathcal{R}_T = \mathbb{E} \left[ \sum_{t=1}^{T} \ell_{t,i} \right] - \min_{i \in [K]} \mathbb{E} \left[ \sum_{t=1}^{T} \ell_{t,i} \right] \leq \mathbb{E} \left[ \sum_{t=1}^{T} \langle \ell_t, x_t \rangle \right] - \min_{y \in \Delta_{[K]-1}} \sum_{t=1}^{T} \langle \ell_t, y \rangle = \max_{y \in \Delta_{[K]-1}} \mathbb{E} \left[ \sum_{t=1}^{T} \langle \ell_t, x_t - y \rangle \right],$$

which further translates to $\max_{y \in \Delta_{[K]-1}} \mathbb{E} \left[ \sum_{t=1}^{T} \langle \ell_t, x_t - y \rangle \right]$ by Lemma 4.

Let $\mathcal{E}_t$ be the event that the feedback of time slot $t$ is never discarded, i.e., $a_t < 2$ when the algorithm terminates. Then, for the time slots such that $\mathbb{I}_{\mathcal{E}_t}$ holds, we extract them out to form a completely new game, where we can use Lemma 2 to bound them. For the remaining of them, note that $\langle \ell_t, x_t - y \rangle \leq \langle \ell_t, x_t \rangle = E_{t \sim x_t} [\ell_{t,i}]$ as all components of $\ell_t$ are non-negative. Therefore, we have

$$\mathcal{R}_T \leq \max_{y \in \Delta_{[K]-1}} \mathbb{E} \left[ B_T \cdot D_\Psi(y, x_0) + \sum_{i=1}^{T} \mathbb{I}_{\mathcal{E}_t} \sigma_t D_\Psi(x_t, z_t) \right] + \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}_{\mathcal{E}_t} \ell_{t,i}],$$

which is just Lemma 7 □

B. Total Investment Term

To bound the total investment term in eq. (1), we will make use of Lemma 8. In addition, the following simple but extremely useful lemma is needed.

Lemma 18. At any time $t \in [T]$, the number of incoming feedback is bounded by $O(\sqrt{D})$. In other words, the number of $1 \leq s \leq t$ satisfying $s + d_s \geq t$ is no more than $\Theta(\sqrt{D})$.

Proof. List all the time slots whose feedback has not arrived in decreasing order, namely $t = s_0 > s_1 > \cdots > s_m$ are all the time slots such that $s_i + d_{s_i} \geq t$ and $1 \leq s_i \leq t$. Then we have $d_{s_i} \geq t - s_i \geq t - s_i - 1 + 1 \cdots \geq t - s_0 + i = i$. As $\sum_{s=1}^{T} d_s = D$, we have $\sum_{i=0}^{m} i = \frac{1}{2} m(m + 1) \leq \sum_{i=0}^{m} d_i \leq D$. Hence, the number of incoming feedback is bounded by $m \leq O(\sqrt{D})$. □

Theorem 19. In Algorithm 1, the total investment term $B_T D_\Psi(y, x_0)$ is bounded by

$$B_T D_\Psi(y, x_0) = O \left( \sqrt{K (D + T) \log D} \right), \quad \forall y \in \Delta_{[K]-1}.$$

Proof. By Lemma 8 there exists some $t_0 \leq T$ such that

$$B_T = B_{t_0} = \sigma_t + \sum_{s=1}^{t_0-1} \mathbb{I}[s + d_s \geq t_0] \sigma_s.$$

(4)
Let \( t_1 < t_2 < \cdots < t_m \) be the time slots whose feedback has not arrived at time slot \( t_0 \). By Lemma 18 \( m = \mathcal{O}(\sqrt{T}) \). Furthermore, at time slot \( t_i \), we must have \( \sigma_{t_i} \geq i \) as the feedback of \( t_1, t_2, \ldots, t_i \) are all absent. Therefore, eq. (4) is further bounded by

\[
B_T = \sigma_{t_0} + \sum_{i=1}^{m} \sigma_{t_i} \leq \frac{1}{\sigma_{t_0}} \sqrt{\frac{3 + D_{t_0}}{\ln(3 + D_{t_0}/L_{t_0}^2)}} + \sum_{i=1}^{m} \frac{1}{i} \max_{1 \leq t \leq T} \sqrt{\frac{3 + D_t}{\ln(3 + D_t/L_t^2)}}.
\]

Now, for any time slot \( t \), we have

\[
\sqrt{\frac{3 + D_t}{\ln(3 + D_t/L_t^2)}} \leq \hat{L}_t \sqrt{\frac{3 + D_t/L_t^2}{\ln(3 + D_t/L_t^2)}}.
\]

Using the fact that \( D_t/L_t^2 \leq \sum_{s=1}^{t} (\sigma_{s} + 1) \leq D + T \) by Lemma 10 and \( f(x) = \frac{x}{\ln x} \) is monotone when \( x \geq 3 \) by Lemma 17 we have

\[
\sqrt{\frac{3 + D_t}{\ln(3 + D_t/L_t^2)}} \leq \hat{L}_t \sqrt{\frac{3 + D + T}{\ln(3 + D + T)}}.
\]

Hence, by the fact that \( \hat{L}_t \leq \max_{1 \leq t \leq T} \ell_{t,i} \leq 2L \) for any \( 1 \leq t \leq T \), we have

\[
B_T \leq \frac{1}{m} 2L \sqrt{\frac{3 + D + T}{\ln(3 + D + T)} + 2L} \sqrt{\frac{3 + D + T}{\ln(3 + D + T)}} \sum_{i=1}^{m} \frac{1}{i} \leq 2L \sqrt{\frac{3 + D + T}{\ln(3 + D + T)}} \Theta(\log \sqrt{D}) = \mathcal{O}(\sqrt{(D + T)\log D}).
\]

Then by Lemma 15 we have \( B_T \cdot D_\Psi(y,x_0) = \mathcal{O}(\sqrt{K(D + T)\log D}) \) for any \( y \in \Delta_{[K\!-\!1]}. \)

C. Total Immediate Cost

**Theorem 20.** In Algorithm 7 the sum of all the immediate cost terms \( \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}_{E_t} \sigma_t D_\Psi(x_t, z_t)] \) is bounded by \( \mathcal{O}(\sqrt{K(D + T)\log(D + T)L}) \).

**Proof.** First apply Lemma 6 to each summmand. That is, for every \( 1 \leq t \leq T \), we have

\[
\mathbb{E}_{t \sim x_t} [\mathbb{1}_{E_t} \sigma_t D_\Psi(x_t, z_t)] \leq \mathbb{E}_{t \sim x_t} [\mathbb{1}_{E_t} \sigma_t^{-1} \hat{L}_t^{-1/2} \ell_{t,i} x_t^{-1/2}] \leq \mathbb{E}_{t \sim x_t} [\mathbb{1}_{E_t} \sigma_t^{-1} \hat{L}_t^{-1/2} x_t^{-1/2}],
\]

where the last step is because \( \mathbb{1}_{E_t} \) implies time slot \( t \) is not discarded, which further indicates that \( \ell_{t,i} \leq \hat{L}_t \). Then we can apply Cauchy-Schwarz inequality, yielding that

\[
\mathbb{E}_{t \sim x_t} [\mathbb{1}_{E_t} \sigma_t D_\Psi(x_t, z_t)] \leq \mathbb{E}_{t \sim x_t} \left[ \frac{K \hat{L}_t^2}{\sigma_t} \right] = \mathbb{E}_{t \sim x_t} \sum_{i=1}^{K} \sqrt{x_t i} \leq \mathbb{E}_{t \sim x_t} \left[ \frac{\sqrt{K \hat{L}_t^2}}{\sigma_t} \right].
\]

Now let’s define another \( \hat{D}_t \) by removing all the unskipped time slots, namely

\[
\hat{D}_t \triangleq \sum_{1 \leq i \leq E_t} (\sigma_{i} + 1) \ell_{t,i}^2.
\]

As we can rewrite \( D_t \) as:

\[
D_t = (\sigma_{t} + 1) \hat{L}_t^2 + \sum_{s < t E_t} (\sigma_{s} + 1) \hat{L}_s^2 + \sum_{s < t - E_t} (\sigma_{s} + 1) \hat{L}_s^2 \mathbb{1}_{\sigma_{s} < 2}
\]

we conclude that \( \hat{D}_t \leq D_t \). Define the corresponding \( \sigma_t \) as

\[
\bar{\sigma}_t = \left. \frac{\ln(3 + D_t/L_t^2)}{3 + D_t} \right)^{-1},
\]

we can then conclude that \( \sigma_t \geq \bar{\sigma}_t \) by Lemma 17 for any \( t \) satisfying \( E_t \). Hence,

\[
\sum_{t=1}^{T} \mathbb{E}_{t \sim x_t} [\mathbb{1}_{E_t} \sigma_t D_\Psi(x_t, z_t)] \leq \sqrt{K \mathbb{E}_{t \sim x_t} \left[ \sum_{t=1}^{T} \mathbb{1}_{E_t} \hat{L}_t^2 \right]} \leq \sqrt{K \mathbb{E}_{t \sim x_t} \left[ \sum_{t=1}^{T} \mathbb{1}_{E_t} \hat{L}_t^2 / \sigma_t \right]} = \sqrt{K \mathbb{E}_{t \sim x_t} \left[ \sum_{t=1}^{T} \mathbb{1}_{E_t} \hat{L}_t^2 (\sigma_t + 1) \sqrt{\frac{\ln(3 + D_t/L_t^2)}{3 + D_t}} \right]}.
\]
As $\hat{D}_t/\hat{L}_t^2 \leq \sum_{s=1}^t (\bar{d}_s + 1)$, we have $\sqrt{\ln(3 + \hat{D}_t/\hat{L}_t^2)} \leq \sqrt{\ln(3 + (D + T))}$ by Lemma 16. Denote by $T = \{1 \leq t \leq T : E_t\}$. Applying Lemma 22 further gives:

$$\sum_{t=1}^T \mathbb{E}_{t \sim c_t}[\mathbb{I}_{E_t} D_S(x_t, \hat{z}_t)] \leq \sqrt{K \ln(3 + (D + T))} \cdot \mathbb{E} \left[ \sum_{\ell \in T} \frac{(\bar{d}_\ell + 1)\hat{L}_\ell^2}{3 + \sum_{s \in T \cap \ell} (\bar{d}_s + 1)\hat{L}_s^2} \right]$$

$$\leq \sqrt{K \ln(3 + D + T)} \cdot L \left( \sum_{\ell \in T} \frac{(\bar{d}_\ell + 1)\hat{L}_\ell^2}{\sum_{\ell \in T} (\bar{d}_\ell + 1)} \right),$$

which further reduces to $O(\sqrt{K(D + T) \ln(D + T)L})$ by applying Lemma 16 again.

### D. Skipping Regret

Denote $\tilde{T}$ to be the complement of $T$, i.e., $\tilde{T} = \{1 \leq t \leq T : \neg E_t\}$. Suppose that $|\tilde{T}| = m$ and $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ where $t_1, t_2, \ldots, t_m$ are in the order of the arrival of their feedback. Throughout this section, for simplicity, denote $\ell_{t,i}$ by $\ell_i$, i.e., the feedback of time slot $t$. We first formalize the partition of $\mathcal{T}$ we described in the text.

**Definition 21.** A “good” partition of $\mathcal{T}$ is characterized by $m_0 = 0 < m_1 < m_2 < \cdots < m_k = m$, such that the following two conditions holds.

- (intra-subset loss upper-bound) for every $0 \leq i < k$, we have $\hat{\ell}_{t,i} \leq 2\hat{\ell}_{t,m_i+1}$ for any $m_i + 1 < j \leq m_{i+1}$;
- (inter-subset loss lower-bound) for every $0 \leq i < k - 1$, $\hat{\ell}_{m_i+1} > 2\hat{\ell}_{m_{i+1}}$.

In such a good partition, denote $\mathcal{T}_i = \{t_j \mid m_{i-1} < j \leq m_i\} \subseteq T$.

Intuitively, the splitting points are just the time slots where the loss feedback get doubled. We have the following lemma stating that this partition will not be too large:

**Lemma 22.** There exists a good partition of $\mathcal{T}$ with $k = O(\log L)$.

**Proof.** Consider the following greedy process. Initially, set $m_0 = 0$ and $i = 1$. Then iterate over all the time slots in $\mathcal{T}$ in the order of their feedback’s arrival. For the current $\ell_t$, if it exceeds twice the current bound $\hat{\ell}_{m_i+1}$, i.e., $\hat{\ell}_{t,i} > 2\hat{\ell}_{m_i+1}$, then put it into a new set $\mathcal{T}_{i+1}$ by setting $m_{i+1} = j - 1$ and $i \leftarrow i + 1$. Otherwise, just keep going (which will put $t_j$ into $\mathcal{T}_{i+1}$). At the end, we set $m_k = m$.

This algorithm indeed produces a good partition, noticing that: 1) $j$ lies in the current interval only if $\hat{\ell}_{t,i} \leq 2\hat{\ell}_{t,m_{i+1}}$, and 2) $j$ goes to a new interval only if $\hat{\ell}_{t,i} > 2\hat{\ell}_{t,m_{i+1}}$. Moreover, as $\hat{\ell}_{t,i} \leq L$ while we have $2\hat{\ell}_{m_{i+1}} < \ell_{m_{i+1}}$, such doubling process cannot last for more than $\log_2 L$ times. Hence, the size of the partition $k$ has the order of $O(\log L)$.

We then give a proof of the observation about arrival times that we made in the text.

**Proof of Lemma 22** If $s \geq t$, there is nothing to do. Otherwise, assume that $s + d_s < t$. As $t \in \mathcal{T}$ holds, $\hat{L}_t < \hat{\ell}_t \leq 2\hat{\ell}_s$. Moreover, by the doubling trick, we have $\hat{L}_{s+d_s} \geq 2\hat{\ell}_s$. However, this contradicts with the fact that $\hat{L}_t$ is monotonic, so we have $s + d_s \geq t$.

With this lemma, we can conclude that each set in this partition will also be small.

**Lemma 23.** For each $1 \leq i \leq k$, we can bound the size of the $i$-th set in the partition, $|\mathcal{T}_i|$, by $O(\sqrt{D})$.

**Proof.** Denote the first time slot of $\mathcal{T}_i$ by $s$, namely $s = t_{m_{i-1}+1}$. By definition of a good partition, $\hat{\ell}_t \leq 2\hat{\ell}_s$, for any $t \in \mathcal{T}_i$. Therefore, we can apply Lemma 10 and conclude that $t \leq s + d_s$ for any $t \in \mathcal{T}_i$. Hence, $t \leq s$ while $t + d_t \geq s + d_s$ (as each element in $\mathcal{T}_i$ is in the order of arrival). As each time slot can only produce one feedback, the $j$-th element of $\mathcal{T}_j$ should produce at least $j$ delay. Therefore, we can conclude from Lemma 18 that there are at most $\sqrt{D}$ such $j$’s, i.e., $|\mathcal{T}_i| = O(\sqrt{D})$.

Now, we can bound the total skipping regret, as follows.

**Theorem 24.** In Algorithm 7 the total skipping regret $\sum_{t=1}^T \mathbb{E}[\mathbb{I}_{E_t} \hat{\ell}_{t,i}]$ is bounded by $O(\sqrt{D}L)$.
Proof. Let us consider the skipping regret of a single interval, $T_i$. Still denote the first time slot of $T_i$ by $s$, namely $s = t_{m_i-1}+1$. By definition, $\ell_s \leq 2\ell_s$ for any $t \in T_i$. Hence, the total regret produced by $T_i$ is at most $2|T_i|\hat{\ell}_s$. Due to the fact that $\hat{\ell}_s$ is at least doubled for each interval while not exceeding $L$, we have

$$\sum_{t=1}^T \mathbb{E}[\mathbb{1}_{-\xi_t} \ell_{t,s}] \leq \sum_{i=1}^k \mathbb{E}\left[ \sum_{t \in T_i} \ell_t \right] \leq 2 \sum_{t=1}^k \mathbb{E}[|T_i|\hat{\ell}_t] \leq O(\sqrt{D}) \sum_{i=0}^{k-1} \frac{L}{2^i} = O(\sqrt{DL}),$$

as claimed, where (a) made use of Lemma 23.

\[\square\]

E. Proof of the Main Theorem

Combining all above will directly prove Theorem 2 as follows.

Proof of Theorem 2 Plugging what we derived into Lemma 7 gives:

$$R_T \leq \max_{y \in \mathcal{A}[1:k]} B_T D\phi(y, \theta_0) + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{-\xi_t} \sigma_t D\phi(x_t, z_t)] + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{-\xi_t} \ell_{t,s}]$$

$$\leq O\left(\sqrt{K(D+T) \log DL}\right) + O\left(\sqrt{K(D+T) \log(D+T)L}\right) + O\left(\sqrt{DL}\right),$$

which completes the proof.

\[\square\]

F. Omitted Proofs in Section VI-A

Proof of Lemma 7 According to Lemma 5 we can write

$$\sigma_t D\phi(x_t, z_t) = \frac{1}{2} \sigma^{-1}_t \left\| \tilde{\ell}_t \right\|_{\nabla^2 \Psi^*(w_t)}^2 = \frac{1}{2} \sigma^{-1}_t \left\| \tilde{\ell}_t \right\|_{\nabla^2 \Psi(\nabla^* \Psi(w_t))^{-1}}^2 \tag{5}$$

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

$$\min\left\{ \tilde{z}_{t,i}, x_{t,i} \right\} \leq \nabla^* \Psi(w_t) = \max\left\{ \tilde{z}_{t,i}, x_{t,i} \right\}.$$ 

The condition $\tilde{z}_{t,i} \leq 2x_{t,i}$ for all $i \in \{1, \ldots, K\}$ implies that $\nabla^* \Psi(w_t) \leq 2x_{t,i}$ for all $i$. Plugging this upper-bound and $\nabla^2 \Psi(x) = \text{diag}(x_1^{-2}, x_2^{-2}, \ldots, x_K^{-2})$ into (5), we get

$$\sigma_t D\phi(x_t, z_t) \leq 2 \sigma^{-1}_t \left\| \tilde{\ell}_t \right\|_{\nabla^2 \Psi(x_t)} = 2 \sigma^{-1}_t \sum_{i=1}^K x_{t,i} \tilde{z}_{t,i}^2.$$ 

\[\square\]

Proof of Lemma 12 It suffice to investigate the mirror map

$$\nabla \Psi(x) = (-1/x_1, -1/x_2, \ldots, -1/x_K)^T$$

and

$$\nabla^* \Psi(\theta) = (-1/\theta_1, -1/\theta_2, \ldots, -1/\theta_K)^T.$$ 

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

$$\tilde{z}_{t,i_t} = \nabla^* \Psi \left( \nabla \Psi(x_t) - \frac{1}{\sigma_t} \tilde{\ell}_t \right)_{i_t}$$

$$= \left(-x_{t,i_t} - \frac{\ell_{t,i_t}}{\sigma_t x_{t,i_t}}\right)^{-1}$$

$$= x_{t,i_t} \cdot \left(1 + \frac{\ell_{t,i_t}}{\sigma_t}\right)^{-1},$$

it is then easy to see that we have $\tilde{z}_{t,i_t} \leq 2x_{t,i_t}$ whenever $\ell_{t,i_t} \geq -\frac{1}{2} \sigma_t$. 

\[\square\]
**G. Total Investment Term**

The way to bound the $\max_{y \in \Delta_{[k-1]}} \mathbb{E} [B_T \cdot D_\Psi (y, x_0)]$ term in (6) is similar to what we have done in Appendix B. For our current purpose, we first give a bound for $\max_{y \in \Delta_{[k-1]}} D_\Psi (y, x_0)$.

**Lemma 25.** When $\Psi$ is the log-barrier function, we have $D_\Psi (y, x_0) \leq K \ln T + K$ for all $y \in \Delta_{[k-1]}$.

**Proof.** Notice that $\Delta_{[k-1]}$ is a compact convex subset of $\mathbb{R}_+^K$ (actually, it is a polyhedron), $\Psi$ is a convex function over $\mathbb{R}_+^K$, the maximum value of $\Psi$ must achieve on the boundary of $\Delta_{[k-1]}$. Due to the symmetry of coordinates of $x_0$, it suffices to verify the bound for all vertices $y$ from $\Delta_{[k-1]}$, and now we have

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

$$\leq (K - 1) \ln T - \ln \left( 1 - \frac{K - 1}{T} \right) + K$$

$$\leq K \ln T + K.$$

\(\Box\)

We also have the following bound for $B_T$.

**Lemma 26.** For Algorithm 2 we have

$$B_T \leq O \left( \sqrt{\frac{D}{K} L} + \frac{(D + T) \log D}{K \log T} \right).$$

Let $\tilde{L}_T^2 \triangleq \sum_{t=1}^T (\delta_t + 1) I_{t, t'}$, we also have a bound for $B_T$ with the $\sqrt{D + TL}$ factor replaced by $\sqrt{\tilde{L}_T^2}$:

$$B_T \leq O \left( \sqrt{\frac{D}{K} L} + \sqrt{\tilde{L}_T^2 \frac{D}{K \log T}} \right).$$

**Proof.** The choice of $\sigma_t$ in Algorithm 2 (Line 14, Line 16) satisfies

$$\sigma_t \leq \left( \delta_t + 1 \right) \frac{\ln (3 + D_t / \tilde{L}_t^2)}{3 + D_t} \sqrt{K \ln T} \sum_{s=1}^{t-1} \mathbb{I}[s + d_s \geq t_0] \sigma_s.$$

By Lemma 8 there exists some $t_0 \leq T$ such that

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

Let $t_1 < t_2 < \cdots < t_m$ be the time slots whose feedback has not arrived at time slot $t_0$. By Lemma 18, $m = O(\sqrt{D})$. Furthermore, at time slot $t_i$, we must have $\delta_t \geq i$ 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 $\delta_i \leq \sqrt{\frac{D_{t_i}}{K}}$. Therefore, $B_T$ can be further bounded by

$$B_T = \sum_{i=0}^m \sigma_{t_i} \leq \sum_{i=0}^m \left( \delta_{t_i} + 1 \right) \frac{\ln (3 + D_{t_i} / \tilde{L}_{t_i}^2)}{3 + D_{t_i}} \sqrt{K \ln T} \sum_{s=1}^{t_i-1} \mathbb{I}[\delta_t \leq \sqrt{\frac{D_{t_i}}{K}}] \cdot 2\tilde{L}_{t_i}$$

By (a)

$$\leq \sum_{i=0}^m \left( \delta_{t_i} + 1 \right) \frac{\ln (3 + D_{t_i} / \tilde{L}_{t_i}^2)}{3 + D_{t_i}} \sqrt{K \ln T} \sum_{s=1}^{t_i-1} \mathbb{I}[\delta_t \leq \sqrt{\frac{D_{t_i}}{K}}] \cdot 2\tilde{L}_{t_i}$$

$$= O \left( \sqrt{\frac{D}{K} L} \right) + \frac{1}{\sqrt{K \ln T}} \sum_{i=0}^m \frac{1}{\delta_{t_i} + 1} \sqrt{\frac{3 + D_{t_i}}{\ln (3 + D_{t_i} / \tilde{L}_{t_i}^2)}}$$

(6)
where (a) is due to $\delta_i \geq i$ for all $i \geq 1$, (b) uses both Lemma [17] and the fact that $D_t/\hat{L}_t^2 \leq D + T$ given by Lemma [16].

In order to get a bound in $\hat{L}_T^2$, similarly define $\hat{L}_t^2 \triangleq \sum_{s=1}^t (\delta_s + 1)/\hat{L}_{s,i}^2$ and define $\hat{m} = \max_{1 \leq t \leq T} \delta_t$. By Lemma [18] we have $\hat{m} = \mathcal{O}(\sqrt{D})$. On the other hand, $D_t$ in Algorithm [2] (see Line 9) satisfies

$$D_t = \sum_{s \in \{1, \ldots, t\}: s + d_s < t} (\delta_s + 1)/\hat{L}_{s,i}^2 + \sum_{s \in \{1, \ldots, t\}: s + d_s \geq t} (\delta_s + 1)\hat{L}_t^2 \leq \hat{L}_t^2 + \hat{L}_t^2(\hat{m} + 1)^2 \leq \hat{L}_t^2 + 100D\hat{L}_t^2$$

where (a) is because in the sum $\sum_{s \in \{1, \ldots, t\}: s + d_s \geq (\delta_s + 1)$, the number of summands and the value of each summand are both bounded by $\hat{m} + 1$.

Leveraging this upper-bound for $D_t$'s, we can continue from eq. [6] to have an upper-bound in $\hat{L}_T^2$:

$$B_T \leq \mathcal{O}\left(\sqrt{\frac{D}{K}}L + \frac{1}{\sqrt{K \ln T}} \sum_{i=0}^{m} \frac{1}{\delta_i + 1} \sqrt{\frac{3 + D_{i}}{\ln(3 + D_{i}/\hat{L}_t^2)}}\right) \leq \mathcal{O}\left(\sqrt{\frac{D}{K}}L + \frac{1}{\sqrt{K \ln T}} \sum_{i=0}^{m} \frac{\hat{L}_t}{\delta_i + 1} \sqrt{\frac{3 + D_{i}/\hat{L}_t^2}{\ln(3 + D_{i}/\hat{L}_t^2)}}\right)$$

$$\leq \mathcal{O}\left(\sqrt{\frac{D}{K}}L + \frac{1}{\sqrt{K \ln T}} \max_{1 \leq t \leq T} \left\{ \hat{L}_t \sqrt{\frac{3 + \hat{L}_t^2/\hat{L}_t^2 + 100D}{\ln(3 + \hat{L}_t^2/\hat{L}_t^2 + 100D)}} \right\} \sum_{i=0}^{m} \frac{1}{\delta_i + 1} \right)$$

$$= \mathcal{O}\left(\sqrt{\frac{D}{K}}L + \frac{1}{\sqrt{K \ln T}} \max_{1 \leq t \leq T} \left\{ \hat{L}_t \sqrt{3 + \hat{L}_t^2/\hat{L}_t^2 + 100D} \right\} \frac{\log D}{\ln(3 + 100D)} \sum_{i=0}^{m} \frac{1}{\delta_i + 1} \right)$$

where step (a) plugs in our $D_t$ bound into $\hat{L}_T^2$.\qed
Taking expectation for the above $B_T$ upper-bound and noticing that the square-root is a concave function, we have $\mathbb{E}[B_T] \leq O\left(\sqrt{\frac{\mathbb{E}D}{K \log T}} + \sqrt{\frac{\mathbb{E}^2 D}{K^2 \log T}}\right)$. Combining the two bounds for $\mathbb{E}[B_T]$ and $D_\phi(y, x_0)$ gives the claimed bound for the total investment term.

**Theorem 27.** In Algorithm 2, the total investment term when $y$ is restricted on $\Delta'_{[K-1]}$ is bounded by

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

$$\max_{y \in \Delta'_{[K-1]}} \mathbb{E}[B_T \cdot D_\phi(y, x_0)] = O\left(\sqrt{KD} \log TL + \sqrt{K \log TL} \sqrt{\hat{DL}_T^2} + \sqrt{K D \log TL}\right).$$

**H. Skipping Regret**

**Theorem 28.** In Algorithm 2, the expected number of skipped time slots, namely the $\sum_{t=1}^T \mathbb{E}[\mathbb{I}_V]$ term in eq. (3), is bounded by $O(\sqrt{D} \log L + \sqrt{KD})$. Furthermore, the skipping regret is bounded by $O((\sqrt{D} \log L + \sqrt{KD})L)$.

**Proof.** For any $1 \leq t \leq T$, define two events

$$U_t \triangleq \left\{ |\ell_{t,i,1}| > \hat{L}_t \right\},$$

$$V_t \triangleq \left\{ |\ell_{t,1}| \leq \hat{L}_t, \ell_{t,i} < -\frac{1}{2}\sigma_t \right\}.$$

In other words, $U_t$ happens if and only if time step $t$ is skipped by Algorithm 2 due to the skipping criterion inherited from Algorithm 1 and $V_t$ happens if and only if time step $t$ is skipped solely due to the new skipping criterion $\ell_{t,i} < -\frac{1}{2}\sigma_t$. Hence our goal reduces to bound $\sum_{t=1}^T \mathbb{E}[U_t] + \sum_{t=1}^T \mathbb{E}[V_t]$, where the first sum is already bounded by $O(\sqrt{D} \log L)$ according to Lemma 22 and Lemma 23. Therefore, it suffices to bound $\sum_{t=1}^T \mathbb{E}[V_t]$. Recall that we maintain experienced total delay $D_t$ in Algorithm 2 and we have $D_0 = 1$, $D_T = D + 1$. For any integer $i \geq 0$, define a stopping time

$$\tau_i \triangleq \inf \left\{ t \geq 0 : D_t \geq \frac{D}{2^i} \right\}.$$

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

Notice that for a $t \geq \tau_i$, the value of $D_t$ is at least $D/2^i$. If $V_t$ happens, then at time $t$, Line 16 of Algorithm 2 cannot be executed (otherwise we will have $\sigma_t \geq 2L_t$ and $\ell_{t,i} < -\frac{1}{2}\sigma_t \leq -\hat{L}_t$, a contradiction), which means $\sigma_t > \frac{D}{2^i \sqrt{T}}$. Therefore conditioned on $V_t$ and $t > \tau_i$, we have $D_t - D_{t-1} = \hat{d}_t > \sqrt{\frac{D}{2^i \sqrt{T}}} \geq \sqrt{\frac{D}{2^i \sqrt{T}}} \geq \sqrt{\frac{D}{2^i \sqrt{T}}}$, and $\sum_{t=\tau_{i-1}}^{\tau_i-1} \mathbb{I}_V$ must be no more than $\frac{D}{2^i \sqrt{T}}/\sqrt{\frac{D}{2^i \sqrt{T}}} = \sqrt{KD} 2^{i-1}/2$. We can then conclude that

$$\sum_{t=1}^T \mathbb{I}_V = \sum_{t=\tau_0}^T \mathbb{I}_V + \sum_{i=1}^{\infty} \sum_{t=\tau_{i-1}}^{\tau_i-1} \mathbb{I}_V \leq (D + 1) \sqrt{\frac{D}{2^i \sqrt{T}}} + \sum_{i=1}^{\infty} \sqrt{KD} 2^{i-1}/2 = O(\sqrt{KD}).$$

Therefore, we have $\sum_{t=1}^T \mathbb{E}[\mathbb{I}_V] = \sum_{t=\tau_0}^T \mathbb{E}[\mathbb{I}_V] + \sum_{i=1}^{\infty} \sum_{t=\tau_{i-1}}^{\tau_i-1} \mathbb{E}[\mathbb{I}_V] = O(\sqrt{D} \log L + \sqrt{KD})$. Furthermore, the total skipping regret is therefore $L \sum_{t=1}^T \mathbb{E}[\mathbb{I}_V] = O(\sqrt{DL} \log L + \sqrt{KDL})$. \hfill $\square$

**I. Total Immediate Cost**

**Theorem 29.** In Algorithm 2, the sum of all the immediate cost terms $\sum_{t=1}^T \mathbb{E}[\mathbb{I}_E, \sigma_t D_\phi(x_t, \hat{z}_t)]$ is bounded by $O(\sqrt{K(D+T)} \log (D+T) \log TL)$. Let $\hat{L}_T^2 \triangleq \sum_{t=1}^T (\hat{d}_t + 1)\ell_{t,i,1}^2$, we also have a bound with the $\sqrt{D+T}L$ factor replaced by $\hat{L}_T$:

$$\sum_{t=1}^T \mathbb{E}[\mathbb{I}_E, \sigma_t D_\phi(x_t, \hat{z}_t)] \leq O\left(\sqrt{K \log (D+T)} \log T \mathbb{E}[\hat{L}_T^2]\right).$$

**Proof.** First apply Lemma 11 and Lemma 13 to each summand. That is, for every $1 \leq t \leq T$, we have

$$\mathbb{I}_E, \sigma_t D_\phi(x_t, \hat{z}_t) \leq \mathbb{I}_E, 2\sigma_t^{-1}\ell_{t,i,1}^2.$$
Now let's define another $D_t$ by removing all the unskipped time slots, namely

$$
\tilde{D}_t \triangleq (\vartheta_t + 1) \tilde{L}_t^2 + \sum_{s < t} (\vartheta_s + 1) \tilde{L}_s^2.
$$

As we can rewrite $D_t$ as:

$$
D_t = (\vartheta_t + 1) \tilde{L}_t^2 + \sum_{s < t} (\vartheta_s + 1) \tilde{L}_s^2 + \sum_{s < t} (\vartheta_s + 1) L_t^2 \mathbb{1}_{a_s < 2 \text{ at that moment}},
$$

we conclude that $\tilde{D}_t \leq D_t$. Define the corresponding $\sigma_t$ as

$$
\bar{\sigma}_t = \left( (\vartheta_t + 1) \sqrt{\frac{\ln(3 + \tilde{D}_t/\tilde{L}_t^2)}{3 + \tilde{D}_t}} \right)^{-1},
$$

we can then conclude that $\sigma_t \geq \bar{\sigma}_t$ by Lemma [17] for any $t$ satisfying $\mathcal{E}_t$. Hence,

$$
\sum_{t=1}^T \mathbb{1}_{\mathcal{E}_t} \sigma_t D_{\varphi}(x_t, \tilde{z}_t) \leq 2 \sum_{t=1}^T \mathbb{1}_{\mathcal{E}_t} \frac{\ell_{t,i}^2}{\sigma_t} \leq 2 \sum_{t=1}^T \sum_{i=1}^{\mathcal{L}_d} \mathbb{1}_{\mathcal{E}_t} \ell_{t,i}^2 \leq 2 \sqrt{K \ln T} \left[ \sum_{t=1}^T \sum_{i=1}^{\mathcal{L}_d} \mathbb{1}_{\mathcal{E}_t} (\vartheta_t + 1) \left( \frac{1}{3} + \frac{\sum_{s \in \mathcal{T}, i} (\vartheta_s + 1) \ell_{s,i}^2}{3 + \tilde{D}_t} \right) \right].
$$

As $\tilde{D}_t/\tilde{L}_t^2 \leq \sum_{s=1}^t (\vartheta_s + 1)$, we have $\sqrt{\ln(3 + \tilde{D}_t/\tilde{L}_t^2)} \leq \sqrt{\ln(3 + (D + T))}$ by Lemma [16]. Denote by $\mathcal{T} = \{1 \leq t \leq T : \mathcal{E}_t \}$, Applying Lemma [9] further gives:

$$
\sum_{t=1}^T \mathbb{1}_{\mathcal{E}_t} \sigma_t D_{\varphi}(x_t, \tilde{z}_t) \leq 2 \sqrt{K \ln(3 + (D + T)) \ln T} \sum_{t=1}^T \mathbb{1}_{\mathcal{E}_t} (\vartheta_t + 1) \left( \frac{1}{3} + \frac{\sum_{s \in \mathcal{T}, i} (\vartheta_s + 1) \ell_{s,i}^2}{3 + \tilde{D}_t} \right) \leq 2 \sqrt{K \ln(3 + (D + T)) \ln T} \cdot O \left( \sqrt{\sum_{t=1}^T \mathbb{1}_{\mathcal{E}_t} \ell_{t,i}^2} \right).
$$

This directly leads to an $O(\sqrt{K(D + T) \log(D + T)L})$ upper-bound. We can also take the expectation of (7) and use Jensen’s inequality to conclude that

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

completing our proof. \(\square\)

J. Proof of the Main Theorem
Combining all above will directly prove Theorem [3] as follows.

Proof of Theorem [3] Plugging what we derived into Lemma [13] gives:

$$
\mathcal{R}_T \leq O(KL) + \max_{y \in \mathcal{A}_{[K-1]}} \mathcal{B}_T D_{\varphi}(y, x_0) + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} \sigma_t D_{\varphi}(x_t, \tilde{z}_t)] + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_t} \sigma_t D_{\varphi}(x_t, \tilde{z}_t)] \leq O \left( \sqrt{K D \log T L} + \sqrt{K D \log T \mathbb{E}[\tilde{L}_T^2]} + \sqrt{K D \log T L} \right) + O \left( \sqrt{K \log(D + T) \log T \mathbb{E}[\tilde{L}_T^2]} \right) \leq O \left( \sqrt{K \log(D + T) \log T \mathbb{E}[\tilde{L}_T^2]} + \sqrt{D L \log T L} \right),
$$

which completes the proof, where the $O(KL)$ term is regarded as a constant without dependency on $T$ and $D$. Plugging in $\tilde{L}_T^2 = O((D + T)L^2)$, we also get

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

\(\square\)
In Algorithm 2, we require the information of time horizon $T$. However, it turns out that this requirement is not critical and is only used to trade off between the total investment term and the immediate costs.

In fact, keeping it unknown will not affect our $O(\sqrt{K(D+T)L})$ regret bound. In that case, we only need to change $\sigma_t$ into

$$\sigma_t = \left(\delta_t + 1\right)\sqrt{\ln(3 + D_t/\tilde{L}^2_t)/3 + D_t}\sqrt{K}^{-1}, \tag{8}$$

i.e., simply removing the $\sqrt{\ln T}$ term in Algorithm 2.

After this modification, everything is unchanged, except for:

- The total investment $B_T$ becomes $O\left(\sqrt{DL} + \sqrt{(D+T)\log D L}\right)$, which will make the total investment term $\Theta(\sqrt{\log T})$ times worse than original.
- The total immediate costs get rid of the $\sqrt{\ln T}$ term, so it becomes $\Theta(\sqrt{\log T})$ times better than original. Now, the two terms of it inside the big-Oh notation become dominated by total investment cost and skipping regret, respectively.

Therefore, we can conclude that:

**Theorem 30.** For SFBanker-LBINF algorithm, if the time horizon length $T$ is unknown, we can still get the following regret bounds by changing Line 14 of Algorithm 2 into eq. (8).

$$R_T \leq O\left(\sqrt{K\log(D+T)\mathbb{E}[\tilde{L}^2_T]}\log T + \sqrt{DL\log L} + \sqrt{KD\log TL}\right),$$

$$R_T \leq O\left(\sqrt{K(D+T)\log(D+T)\log T} + \sqrt{DL\log L}\right).$$

where $\tilde{L}^2_T = \sum_{t=1}^T (\delta_t + 1)\ell^2_{t,i}$, and $\delta_t$ denotes the number of feedback that have not arrived at the beginning of time-step $t$, just as Theorem 2.