IMPROVED ALGORITHMS FOR ADVERSARIAL MULTI-ARMS BANDIT WITH UNBOUNDED LOSSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the Adversarial Multi-Armed Bandits (MAB) problem with unbounded losses, where the algorithms have no prior knowledge on the sizes of the losses. We present UMAB-NN and UMAB-G, two algorithms for non-negative and general unbounded loss respectively. For non-negative unbounded loss, UMAB-NN achieves the first adaptive and scale free regret bound without uniform exploration. Built up on that, we further develop UMAB-G that can learn from arbitrary unbounded loss. Our analysis reveals the asymmetry between positive and negative losses in the MAB problem and provide additional insights. We also accompany our theoretical findings with extensive empirical evaluations, showing that our algorithms consistently out-performs all existing algorithms that handles unbounded losses.

1 INTRODUCTION

Multi-armed bandit (MAB) presents a popular online learning framework for studying decision making under uncertainty (Slivkins et al., 2019; Lattimore & Szepesvári, 2020; Bubeck et al., 2012), with a wide range of applications such as advertisement (Schwartz et al., 2017), medical treatments (Villar et al., 2015), and recommendation systems (Mary et al., 2015). In this paper we focus on the adversarial MAB (AMAB), where the losses are allowed to be generated adversarially by the environment Auer et al. (2002). Most prior works on AMAB assume that the losses are naturally bounded, e.g. $\ell_t \in [0, 1], \forall t$. With such knowledge, the algorithms can set their *learning rate* (in a general sense) properly. For example, in its regret analysis, the EXP3 algorithm relies on the inequality $\exp(x) \leq 1 + x + (e-2)x^2$ to transform exponential terms into quadratic terms (Auer et al., 2002), which only holds true if the loss x can be upper bounded by 1. In many real-world applications, however, such natural loss bound does not always exist. For example, in quantitative trading, the fluctuation of stock prices can differ wildly across time. In online market places, the price can vary dramatically for different products. If one must give a uniform bound M for the losses across all actions and time, such a bound will likely be loose. In such cases, existing algorithms will have a regret that scales with M, which is suboptimal compared to a guarantee that depends on the actual size of the losses.

Motivated by the above limitation of existing algorithms, we wish to design AMAB algorithms that require no prior knowledge on the scale of the losses and *adaptively* achieves smaller regret when the losses are small in scale. In addition, instead of a regret bound that depends on the number of rounds and a (hidden) uniform bound of the losses, we wish to design *data-dependent* algorithms whose regret scales with the actual loss sequence, which is beneficial when the sequence of loss is sparse or when its scale varies across time (Wei & Luo, 2018; Bubeck et al., 2018). In other words, we would like to ask the following question:

Can we design an algorithm that achieves **optimal** and **adaptive** regret guarantee
without any prior knowledge on the losses?

Algorithm	Unbounded	Adaptive	Regret
(Hazan & Kale, 2011)	No	Yes	$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\Vert\ell_t\Vert_2^2}\right)$
(Hadiji & Stoltz, 2020)	Yes	No	$\widetilde{O}\left(\ell_\infty\sqrt{nT}\right)$
(Putta & Agrawal, 2022) Non-Adaptive	Yes	No	$\widetilde{O}\left(\ell_\infty\sqrt{nT}+\sqrt{n\sum_{t=1}^{T}\Vert\ell_t\Vert_2^2}\right)$
(Putta & Agrawal, 2022) Adaptive	Yes	Yes	$\widetilde{O}\left(\ell_\infty\sqrt{n\sum_{t=1}^{T}\Vert\ell_t\Vert_1}+\sqrt{n\sum_{t=1}^{T}\Vert\ell_t\Vert_2^2}\right)$
`UMAB-G` Non-Adaptive	Yes	No	$\widetilde{O}\left(\ell_\infty^-\sqrt{nT}+\sqrt{n\sum_{t=1}^{T}\Vert\ell_t\Vert_\infty^2}\right)$
`UMAB-G` Adaptive	Yes	Yes	$\widetilde{O}\left(\ell_\infty\sqrt{n\sum_{t=1}^{T}\Vert\ell_t\Vert_\infty}+\sqrt{n\sum_{t=1}^{T}\Vert\ell_t\Vert_\infty^2}\right)$

Table 1: Comparison between our results and previous works[1]

In the following, we present two algorithms, UMAB-NN and UMAB-G, for Non-Negative and General unbounded loss, respectively. Our main contributions can be summarized as follows.

1. We propose UMAB-NN, a **scale-free** AMAB algorithm that works for unbounded non-negative losses. The regret guarantee of UMAB-NN adapts to the infinity norm of the loss sequence while matching the worst-case lower bound of Auer et al. (2002).
2. Building upon UMAB-NN, we then propose UMAB-G which works for arbitrary unbounded losses that can be both possible and negative. We present two versions of the algorithm, distinguished by whether the exploration subroutine adapts to the observed losses. For the non-adaptive version, it achieves an optimal worst-case regret guarantee and partially adapts to the non-negative part of the loss sequence, improving upon the previous results of Hadiji & Stoltz (2020); Putta & Agrawal (2022); Huang et al. (2023). For the adaptive version, our algorithm achieves an improvement on the order of $\mathcal{O}(\sqrt{n})$ compared to Putta & Agrawal (2022), where n is the number of the actions.
3. Last but not least, we evaluate the performance of our algorithms on real world datasets. The results show that our algorithms consistently outperform existing methods in a variety of tasks with distinct loss patterns. We also construct synthetic simulations to illustrate the impact of our exploration strategy and draw comparisons between the two versions of our algorithm.

2 Problem Setup and Related Works

We start with some notations. Let $[n]$ denote the set $\{1,\ldots,n\}$ and $[T]$ denote the set $\{1,\ldots,T\}$. Let Δ_n be the probability simplex $\{\mathbf{p}\in\mathbb{R}^n : \sum_{k\in[n]} p_k = 1; p_k \geq 0, \forall k\in[n]\}$. Let $\mathbf{1}_n$ and $\mathbf{0}_n$ be all ones and all zeros n-dimensional vector respectively. Let $\mathbf{e}_k$ denotes the one-hot vector with 1 on the kth entry. For vectors $\mathbf{p}_t$ and ℓ_t, we use $p_{t,k}$ and $\ell_{t,k}$ to represent the kth entry of $\mathbf{p}_t$ and ℓ_t respectively. The L1, L2 and L-infinity norms of ℓ_t are denoted as $\|\ell_t\|_1 = \sum_{k\in[n]}|\ell_{t,k}|$, $\|\ell_t\|_2 = \sqrt{\sum_{k\in[n]}\ell_{t,k}^2}$, $\|\ell_t\|_\infty = \max_{k\in[n]}|\ell_{t,k}|$ respectively. We denote by $\ell_\infty = \max_{t\in[T]}\|\ell_t\|_\infty$ the uniform norm bound of the losses. Moreover, we denote by $\ell_\infty^- = \max_{t\in[T],k\in[n]}|\min(\ell_{t,k},0)|$ the magnitude of the most negative entry of the losses. Notice that $\ell_\infty^- \leq \ell_\infty$, and $\ell_\infty^- = 0$ if the loss sequence is non-negative. Both ℓ_∞ and ℓ_∞^- are unknown to the player through the game.

Adversarial Multi-armed Bandit: We consider the *oblivious adversarial* setting. In each round $t = 1,\ldots,T$, the player selects a distribution $\mathbf{p}_t$ over $[n]$ and the adversary selects a loss vector $\ell_t\in\mathbb{R}^n$ *simul-*

[1] For brevity we consider $n, \ell_\infty \ll T$ and omit the log terms. Detailed regret is provided later.

taneously. Then, the player samples action $k_t \sim \mathbf{p}_t$ and observes loss ℓ_{t,k_t}. We measure the performance of an algorithm in terms of its *pseudo-regret*:

$$\mathcal{R}_T := \mathbb{E}\Big[\sum\nolimits_{t=1}^{T} \ell_{t,k_t} - \min_{k\in[n]} \sum\nolimits_{t=1}^{T} \ell_{t,k}\Big] \tag{1}$$

2.1 Related Works

Scale-free algorithms are ones whose regret bound scales linearly with respect to ℓ_∞, while requiring no knowledge of ℓ_∞ a prior [2]. Scale-free regret bounds were first studied in the full information setting, such as experts problems (Freund & Schapire, 1997; De Rooij et al., 2014; Cesa-Bianchi et al., 2007) and online convex optimization (Mayo et al., 2022; Jacobsen & Cutkosky, 2023; Cutkosky, 2019). For experts problems, the `AdaHedge` algorithm from De Rooij et al. (2014) achieves the first scale-free regret bound. For online convex optimization, past algorithms can be categorized into two generic algorithmic frameworks: Mirror Descent (MD) and Follow The Regularizer Leader (FTRL). The scale-free regret from the MD family is achieved by `AdaGrad` proposed by Duchi et al. (2011). However, the regret bound of Duchi et al. (2011) is only non-trivial when the Bregman divergence associated with the regularizer can be well bounded. Later, the Orabona & Pál (2018) proposed the `AdaFTRL` algorithm which achieves the first scale-free regret bound in the FTRL family and generalizes Duchi et al. (2011)'s results to cases where the Bregman divergence associated with the regularizer is unbounded. For the AMAB problem, Hadiji & Stoltz (2020) extends the method of Duchi et al. (2011) and provides a scale-free regret bound of $\widetilde{O}\left(\ell_\infty\sqrt{nT}\right)$, which is optimal (up to log terms) in the worst case. However, such worst-case regret bounds can be overly pessimistic in general cases: a single outlier loss $\ell_{outlier}$ can result in an additional regret on the order of $O(\|\ell_{outlier}\|_\infty\sqrt{nT})$. To address it, Putta & Agrawal (2022) presents scale-free bounds that adapt to the individual size of losses across time. Unfortunately, the worst-case guarantee of Putta & Agrawal (2022) is $\widetilde{O}\left(\ell_\infty n\sqrt{T}\right)$, which scales linearly to the number of actions. Our paper closes this gap: our algorithms achieve an adaptive regret better than Putta & Agrawal (2022), as well as an optimal worst-case regret that matches with Hadiji & Stoltz (2020).

Adaptive algorithms refer to the algorithms that dynamically adjusts to the input data it encounters. Rather than scaling solely on T in the regret, an adaptive algorithm adapts to a "measure of hardness" of the sequence of losses. An adaptive regret algorithm performs better than the worst-case regret if the sequence of loss is "good". In the last two decades, adaptive algorithms have been widely studied in the settings of expert problems and online convex optimization (Hazan et al., 2007; Streeter & McMahan, 2010; Duchi et al., 2011; De Rooij et al., 2014; Orabona & Pál, 2015; 2018). For the MAB setting, several works derive adaptive regret bounds based on different "measure of hardness". For example, Allenberg et al. (2006); Foster et al. (2016); Pogodin & Lattimore (2020); Ito (2021) derive the first-order regret (a.k.a. *small-loss regret*), which depends on the cumulative loss $\min_{k\in[n]}\sum_{t\in[T]}|\ell_{t,k}|$, but under the assumption that $\ell_{t,k} \in [0,1], \forall t, k$. Hazan & Kale (2011); Bubeck et al. (2018); Wei & Luo (2018); Ito (2021) propose bounds that depend on the empirical variance of the losses, i.e., $\sum_{t\in[T]}\|\ell_t\|_2^2$. Path-length bounds are also studied (Wei & Luo, 2018; Bubeck et al., 2019; Zimmert & Seldin, 2021; Ito, 2021), which depends on the fluctuation of loss sequence $\sum_{t\in[T]}\|\ell_t - \ell_{t-1}\|_1$. We remark that *all* results above require the assumption that losses are bounded within $[0,1]$, which we remove in this paper.

[2] We note that an alternative and more strict interpretation of scale-free algorithms refers to ones that will not change the sequence of p_t's when losses are multiplied by a positive constant.

3 Algorithm and Analysis

We now present our two algorithms `UMAB-NN` and `UMAB-G`. `UMAB-NN` works for the case where losses are Non-Negative, i.e., $\ell_t \in \mathbb{R}_+^n$. Remarkably, `UMAB-NN` is a *strictly scale-free* algorithm: the algorithm will not change its sequence of action distributions if the sequence of losses is multiplied by a positive constant, which immediately implies scale-free regret. Our second algorithm, `UMAB-G`, builds upon the first algorithm to allow potentially negative losses, i.e., $\ell_t \in \mathbb{R}^n$. We provide two versions of the algorithm: `UMAB-G` with non-adaptive and adaptive exploration rates. For the non-adaptive version, our results achieve adaptability to the non-negative part of the loss, while ensuring the optimality for the worst case guarantee, which is new compared to previous works [3]. For the adaptive version, we improve the previous result (Putta & Agrawal, 2022) by $\mathcal{O}(\sqrt{n})$. A summary of the comparisons to prior works can be found in Table 1.

Both the algorithms we propose are based on the Follow-the-Regularized-Leader (FTRL) framework. Let us first consider the full information case, the traditional adaptive FTRL framework uses a regularizer Ψ and time-varying learning rates $\eta_1, \dots, \eta_{T+1}$, with certain regularity constraints (see, e.g., (Orabona & Pál, 2015)). The update rule takes the form of

$$\mathbf{p}_1 = \arg\min_{\mathbf{p}\in\Delta_n} \frac{1}{\eta_1}\Psi(\mathbf{p}), \qquad \mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n} \Big(\sum_{s=1}^{t-1}\langle \ell_s, \mathbf{p}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p})\Big), \tag{2}$$

where ℓ_s is the observed loss at round s and η_t is the adaptive learning rate depending on the losses $\ell_1, \dots, \ell_{t-1}$. In the bandit setting, we cannot observe the complete loss vector ℓ_t. Similar to prior works, we construct an unbiased loss estimator through the importance-weighted (IW) sampling method introduced by (Auer et al., 2002), i.e., construct $\hat{\ell}_t \in \mathbb{R}^n$ such that $\hat{\ell}_{t,k} = \frac{\mathbb{1}(k=k_t)}{p_{t,k}}\ell_{t,k}$, $\forall k \in [n]$, where $\mathbb{1}(k = k_t)$ denotes the indicator function. Notice that $\mathbb{E}[\hat{\ell}_t] = \sum_{k=1}^n p_{t,k}\frac{\mathbf{e}_k}{p_{t,k}}\ell_t = \ell_t$. Using $\hat{\ell}_t$, we are able to reduce the bandit setting to the full information case.

3.1 Non-negative loss

Let's start with the setting where the loss sequence is non-negative but can be arbitrarily large, i.e., $\ell_{t,k} \geq 0$ for every $t \in [T]$ and $k \in [n]$. `UMAB-NN` (Algorithm 1) is a natural adaptation of the classic FTRL algorithm with log-barrier regularizer. The log-barrier regularizer is defined as

$$\Psi(\mathbf{p}_t) = \sum_{k=1}^n \Big(\log\Big(\frac{1}{p_{t,k}}\Big) - \log\Big(\frac{1}{n}\Big)\Big).$$

Notice that $\Psi(\mathbf{p}) \geq 0$ for all $\mathbf{p} \in \Delta_n$. Such regularizers are commonly used for studying adaptive regret in the AMAB setting (Wei & Luo, 2018; Putta & Agrawal, 2022; Bubeck et al., 2019). In each round, `UMAB-NN` calculates an action distribution $\mathbf{p}_t$ through the update rule, then plays action k_t sampled from $\mathbf{p}_t$. After receiving loss $\ell_{t,k}$, `UMAB-NN` constructs the unbiased IW estimator $\hat{\ell}_t$ and updates the learning rate η_t. The novelty comes in our design of learning rate (line 5). Different from the learning rate in Orabona & Pál (2018), we use ℓ_{t,k_t}^2 instead of $\|\hat{\ell}_t\|_2^2$. This is because $\|\hat{\ell}_t\|_2^2$ is of order $1/p_{t,k_t}^2$. If one uses the one in Orabona & Pál (2018) instead, i.e. $\eta_{t+1} = O(\sqrt{n/\sum_{s=1}^t \|\hat{\ell}_s\|_2^2})$, the learning rate will be too small since $1/p_{t,k_t}^2$ cannot be bounded. Based on this observation, `UMAB-NN` adapts the learning rate to the sum of the square of the partial loss, i.e., $\eta_{t+1} = O(\sqrt{n/\sum_{s=1}^t \ell_{s,k_s}^2})$, which can be well bounded by $O(\ell_\infty\sqrt{n/T})$.

[3] We note that a recent work (Huang et al., 2023) proposes an algorithm that achieves adaptive regret for general unbounded loss. However, there exists a critical issue within their proof and algorithm, resulting in their regret being actually unbounded. We have communicated with the authors about the issue. The details are provided in Appendix B.2.

Algorithm 1: UMAB-NN: Unbounded AMAB for Non-Negative loss

```
Input: Log-barriers regularization $\Psi$, $\eta_1 = \infty$
1 for $t = 1, \dots, T$ do
2     Compute the action distribution $\mathbf{p}_t = \arg\min_{\mathbf{p} \in \Delta_n} \left( \sum_{s=1}^{t-1} \langle \hat{\ell}_s, \mathbf{p} \rangle + \frac{1}{\eta_t} \Psi(\mathbf{p}) \right)$
3     Sample and play action $k_t \sim \mathbf{p}_t$. Receive loss $\ell_{t,k_t}$
4     Construct IW estimator $\hat{\ell}_t$ such that $\hat{\ell}_{t,k} = \frac{\mathbb{1}(k=k_t)}{p_{t,k}} \ell_{t,k}, \ \forall k \in [n]$
5     Update learning rate $\eta_{t+1} = 2\sqrt{\frac{n}{\sum_{s=1}^{t} \ell_{s,k_s}^2}}$
```

We remark that Algorithm 1 is strictly scale-free. If all losses are multiplied by a constant c, then in line 2, both terms on the right hand side will be multiplied by c, resulting in the same p_t being picked by the algorithm. Our main result is the following regret bound for Algorithm 1.

Theorem 1 *For any $\ell_1, \dots, \ell_T \in \mathbb{R}_+^n$, the expected regret of Algorithm 1 is upper bounded by*

$$\mathcal{R}_T \le \tilde{\mathcal{O}}\left(\sqrt{n \sum\nolimits_t \|\ell_t\|_\infty^2}\right)$$

Notice that Theorem 1 is adaptive to the infinity norm of the losses. Furthermore, the worst case regret is bounded by $\tilde{\mathcal{O}}(\ell_\infty \sqrt{nT})$, which matches the lower bound established in (Auer et al., 2002). We remark that Theorem 1 is the first result that achieves both optimal adaptive rate and optimal minimax rate for unbounded non-negative losses. Next, we briefly highlight the key steps in proving Theorem 1, which also provide intuition for our further improvement in the next section.

Proof sketch of Theorem 1: Since $\hat{\ell}_t$ is an unbiased estimator of ℓ_t for every $t \in [T]$ and comparator $\mathbf{p}^\dagger \in \Delta_n$, we have $\mathbb{E}\left[\sum_{t=1}^T \ell_{t,k_t} - \sum_{t=1}^T \langle \ell_t, \mathbf{p}^\dagger \rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle\right]$. It suffices to focus on the regret of $\sum_{t=1}^T \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle$. We start with the standard analysis of a FTRL-type algorithm.

Lemma 1 *((Orabona, 2019) Lemma 7.1) For any $\hat{\ell}_1, \dots, \hat{\ell}_T \in \mathbb{R}^n$, using the update rule of (2) along with the non-increasing sequence of learning rates $\eta_1, \dots, \eta_{T+1}$, there is*

$$\sum_{t=1}^T \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \le \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \sum_{t=1}^T \Big(\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \Big)$$

for every comparator $\mathbf{p}^\dagger \in \Delta_n$, where function F_t is defined as $F_t(\mathbf{p}) = \sum_{s=1}^{t-1} \langle \hat{\ell}_s, \mathbf{p} \rangle + \frac{1}{\eta_t} \Psi(\mathbf{p})$.

For the sake of completeness, the proof of Lemma 1 is provided in the appendix. Lemma 1 decomposes the regret into two terms. The first term depends on the regularizer and the comparator. Intuitively, $\Psi(\mathbf{p}^\dagger)$ will appear to be infinity if $\mathbf{p}^\dagger$ is the best fixed action (some entries of $\mathbf{p}^\dagger$ are zeros). The problem can be easily solved by comparing with some close neighbor of the best action (Putta & Agrawal, 2022), i.e., mixing a uniform distribution with the best fixed action. Therefore, it suffices to focus on the terms $\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1})$. The following key lemma gives an upper bound using the notions of local norms.

Lemma 2 *For any $\hat{\ell}_1, \dots, \hat{\ell}_T \in \mathbb{R}^n$, using the update rule of (2), denote by $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}$, there is*

$$\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \le \frac{1}{2} \eta_t \|\hat{\ell}_t\|_{(\nabla^2 \Psi(\xi_t))^{-1}}^2, \tag{3}$$

where ξ_t is a point between $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$. Moreover, it suffices to set ξ_t as $\mathbf{p}_t$ when $\hat{\ell}_t \in \mathbb{R}_+^n$.

Algorithm 2: UMAB-G: Unbounded AMAB for General Loss

```
Initialize: Log-barriers regularization $\Psi$, learning rate $\eta_1 = 1/4$, exploration rate $\rho_1 = 1/2n^2$, clipping
            threshold $C_1 = -1$
for $t = 1, \ldots, T$ do
    Compute the action distribution: $\mathbf{p}_t = \arg\min_{\mathbf{p} \in \Delta_n} \left( \sum_{s=1}^{t-1} \langle \hat{\ell}'_s, \mathbf{p} \rangle + \frac{1}{\eta_t} \Psi(\mathbf{p}) \right)$
    Calculate $\mathbf{p}'_t$ by Algorithm 3 with exploration rate $\rho_t$. Play action $k_t \sim \mathbf{p}'_t$. Receive loss $\ell_{t,k_t}$.
    Construct loss estimator $\hat{\ell}'_t$ such that $\hat{\ell}'_{t,k} = \frac{\mathbb{1}(k=k_t)\ell'_{t,k}}{p'_{t,k}}$, $\forall k \in [n]$, where $\ell'_{t,k_t} = \max(2C_t, \ell_{t,k_t})$.
    Update clipping threshold $C_{t+1} = \min(C_t, \ell'_{t,k_t})$.
    Update learning rate: $\eta_{t+1} = \frac{1}{4}\sqrt{\frac{n}{nC_{t+1}^2 + \sum_{s=1}^{t} \ell'^2_{s,k_s}}}$.
    Update exploration rate:
        1. (Non-Adaptive): $\rho_{t+1} = 1/(2n^2 + \sqrt{nT})$.
        2. (Adaptive): $\rho_{t+1} = 1/(2n^2 + 2\sqrt{\sum_{s=1}^{t} |\langle \hat{\ell}_s, \mathbf{c}_t \rangle|})$.
```

Algorithm 3: Extra Exploration on Action Distribution

```
Input: Action distribution $\mathbf{p}_t$. Exploration rate $\rho_t \le 1/2n^2$
Output: Extra exploration distribution $\mathbf{p}'_t$
Define $k_t^\star \in \arg\max_{k' \in [n]} p_{t,k'}$. Construct a vector $\mathbf{c}_t \in \mathbb{R}^n$ such that for every $k \in [n]$, there is
    $$c_{t,k} = \begin{cases} 1, & \text{if } p_{t,k} < \rho_t \\ -\sum_{k' \in [n]/\{k\}} c_{t,k'} & \text{if } k = k_t^\star \\ 0, & \text{else} \end{cases}$$
Construct the extra exploration distribution $\mathbf{p}'_t = \mathbf{p}_t + \rho_t \mathbf{c}_t$.
```

Note that (3) holds for general losses and will be useful in the next section. When $\hat{\ell}_t \in \mathbb{R}_+^n$, we can further bound (3) by $\min\left(\frac{1}{2}\eta_t \ell_{t,k_t}^2, |\ell_{t,k_t}|\right)$ since $\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \le \langle \hat{\ell}_t, \mathbf{p}_t \rangle = |\ell_{t,k_t}|$, which implies

$$\sum\nolimits_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \le \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \sum\nolimits_{t=1}^{T} \min\left(\frac{1}{2}\eta_t \ell_{t,k_t}^2, |\ell_{t,k_t}|\right). \tag{4}$$

The right hand side of (4) takes a similar form as in scale-free online convex optimization (Orabona & Pál, 2018), but the upper bound depends on ℓ_{t,k_t} instead of $\|\ell_t\|_2$. Using a learning rate as in Algorithm 1, the second term on the right hand side of (4) can be bounded by $\mathcal{O}(\sqrt{n \sum_{t=1}^{T} \ell_{t,k_t}^2})$ based on (Orabona & Pál, 2018), which suffices to complete the proof.

3.2 General loss

Next, we remove the non-negative assumption and study the general loss setting, i.e., $\ell_1, \ldots, \ell_T \in \mathbb{R}^n$. To begin with, we first explain why Algorithm 1 cannot work when the losses become negative. Recall Lemma 2, it requires bounding $\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1})$ by $\eta_t \|\hat{\ell}_t\|^2_{(\nabla^2 \Psi(\xi_t))^{-1}}/2$ for general losses. However, notice that

$$\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \sum_{k=1}^{n} \frac{\hat{\ell}^2_{t,k}}{\nabla^2_{k,k}\Psi(\xi_t)} = \sum_{k=1}^{n} \frac{\ell^2_{t,k}\mathbb{1}(k=k_t)}{p^2_{t,k}}\xi^2_{t,k} = \frac{\ell^2_{t,k_t}}{p^2_{t,k_t}}\xi^2_{t,k_t}, \tag{5}$$

where ξ_{t,k_t} is some value between p_{t,k_t} and p_{t+1,k_t}. Given p_{t+1,k_t} might significantly exceed p_{t,k_t}, the size of $\xi_{t,k_t}/p_{t,k_t}$ cannot be confined. In this case, $\ell^2_{t,k_t}\xi^2_{t,k_t}/p^2_{t,k_t}$ is potentially of order $O(1/p^2_{t,k_t})$, which is too large for the analysis. Additionally, $-\langle\hat{\ell}_t, \mathbf{p}_{t+1}\rangle$ could potentially be positive and cannot be well bounded due to the same reason. Thus, inequality (4) no longer holds under the condition of general loss. Inspired by such observations, it naturally follows to consider bounding the magnitude of $p_{t+1,k_t}/p_{t,k_t}$. Unfortunately, without imposing additional restrictions on the losses, using the update (2) directly cannot bound $p_{t+1,k_t}/p_{t,k_t}$. For example, given arbitrary $\mathbf{p}_t$, η_{t+1}, and k_t, we can always find a sufficiently small $\ell_{t,k_t} < 0$ that makes $p_{t+1,k_t} \geq 1/2$ through (2). In this case, if p_{t,k_t} is close to zero, $p_{t+1,k_t}/p_{t,k_t}$ could be extremely large.

To address this issue, we propose UMAB-G, as illustrated in Algorithm 2. The key ideas of UMAB-G include (1) using truncated loss to update the action distribution. Instead of directly taking $\hat{\ell}_t$ as the input loss, we clip it by a threshold C_t that depends on previous received losses $\hat{\ell}_1, \ldots, \hat{\ell}_{t-1}$. The truncation ensures that every input loss is "not too negative" for the update of action, and thus the magnitude of $p_{t+1,k_t}/p_{t,k_t}$ can be well bounded. (2) adding an extra exploration to ensure that the probability $p_{t,k}$ would not be overly small. For unbounded AMAB with general loss, we need to ensure that each arm has a certain probability to be pulled, so that we can perceive the change of loss norm in time to tune the learning rate. Instead of the commonly used scheme of mixing with a uniform distribution (Hadiji & Stoltz, 2020; Putta & Agrawal, 2022), we develop a data-dependent mixing strategy (Algorithm 3) that substantially reduces the error caused by the extra exploration. Specifically, similar to (Putta & Agrawal, 2022), we consider two exploration rate distinguished by whether the exploration rate is adaptive. The main result of Algorithm 2 is as follows.

Theorem 2 *For any $\ell_1, \ldots, \ell_T \in \mathbb{R}^n$, with the non-adaptive and adaptive exploration rate, the expected regret of Algorithm 2 is upper bounded by*

$$\textit{Non-Adaptive:} \quad \mathcal{R}_T \leq \tilde{\mathcal{O}}\Big(\ell_\infty n^2 + \sqrt{n\sum\nolimits_t \|\ell_t\|^2_\infty} + \ell^-_\infty\sqrt{nT}\Big), \tag{6}$$

$$\textit{Adaptive:} \quad \mathcal{R}_T \leq \tilde{\mathcal{O}}\Big(\ell_\infty n^2 + \sqrt{n\sum\nolimits_t \|\ell_t\|^2_\infty} + \ell_\infty\sqrt{n\sum\nolimits_t \|\ell_t\|_\infty} + \sqrt{n\sum\nolimits_t \|\ell_t\|_\infty}\Big) \tag{7}$$

Notice that the non-adaptive regret in Theorem 2 achieves "semi-adaptivity" to the loss sequence. If the loss sequence is non-negative, the right hand side of (6) is reduced to a form of the regret in Theorem 1. Moreover, the worst case bound of (6) is $\tilde{\mathcal{O}}(\ell_\infty\sqrt{nT})$ for large T, which is optimal up to log factors (Auer et al., 2002). For the adaptive exploration rate, our result improves upon the previous result (Putta & Agrawal, 2022) and achieves optimal dependency on n and T.

Proof sketch: Recall that $\hat{\ell}_t$ is the unbiased estimator and $\hat{\ell}'_t$ is the clipping biased estimator. By Algorithm 2 and the proof of Theorem 1, it suffices to bound the expectation of $\sum_{t=1}^{T}\langle\hat{\ell}_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle$. We first decompose the regret into three terms as follows.

$$\sum_{t=1}^{T}\langle\hat{\ell}_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle = \underbrace{\sum_{t=1}^{T}\langle\hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle}_{①} + \underbrace{\sum_{t=1}^{T}\langle\hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle}_{②} + \underbrace{\sum_{t=1}^{T}\langle\hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle}_{③}.$$

Here, term ① is the regret of the corresponding FTRL algorithm with truncated loss $\hat{\ell}'_1, \dots, \hat{\ell}'_T$. Term ② measures the error incurred by extra exploration, i.e., using $\mathbf{p}'_t$ instead of $\mathbf{p}_t$. Term ③ corresponds to the error of using the truncated loss $\hat{\ell}'_t$. In the rest of the proof, we bound these three terms respectively.

Bounding ①: By Lemma 1 and Lemma 2, we have

$$\sum_{t=1}^{T} \langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \leq \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\hat{\ell}'_t\|^2_{(\nabla^2 \Psi(\xi_t))^{-1}} = \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_t {\ell'}^2_{t,k_t} \frac{p^2_{t,k_t}}{{p'}^2_{t,k_t}} \frac{\xi^2_{t,k_t}}{p^2_{t,k_t}}.$$

The key step is to bound the magnitude of $p_{t,k_t}/{p'}_{t,k_t}$ and $p_{t+1,k_t}/p_{t,k_t}$ (since ξ_{t,k_t} is always between p_{t,k_t} and p_{t+1,k_t}) for $\ell_{t,k_t} \leq 0$. This in turn is guaranteed by our design of loss truncation and extra exploration. As shown in Lemma 4, Algorithm 2 ensures that both $p_{t,k_t}/p'_{t,k_t}$ and $p_{t+1,k_t}/p_{t,k_t}$ can be bounded by constants. With these two ratio bounded, we can immediately reduce the right-hand-side to the form of (4). Using a similar proof as in Section 3.1, we can bound ①.

Bounding ②: By the definition of $\mathbf{p}'_t$, we first note that $\sum_{t=1}^{T} \langle \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t \rangle = \sum_{t=1}^{T} \rho_t \langle \hat{\ell}'_t, \mathbf{c}_t \rangle$, where ρ_t is the exploration rate and $\mathbf{c}_t$ is an offset on p_t to prevent some entries in action distribution from being too small. The key of our extra exploration algorithm is to upper bound $\langle \hat{\ell}'_t, \mathbf{c}_t \rangle$ by $\mathcal{O}(\ell_\infty \sqrt{nT})$, in contrast to $\mathcal{O}(\ell_\infty n^{3/2} \sqrt{T})$ as in (Putta & Agrawal, 2022). This reduces the variance of our exploration rate, leading to an improved regret. The details are provided in Lemma 5.

Bounding ③: Notice that $\sum_{t=1}^{T} \langle \hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger \rangle \leq \sum_{t=1}^{T} \|\hat{\ell}_t - \hat{\ell}'_t\|_1 \|\mathbf{p}'_t - \mathbf{p}^\dagger\|_\infty \leq \sum_{t=1}^{T} \|\hat{\ell}_t - \hat{\ell}'_t\|_1$. The key idea of bounding ③ is to show that the number of distinct $(\hat{\ell}_t, \hat{\ell}'_t)$ pairs and $\|\hat{\ell}_t\|_\infty$ can be bounded by $\mathcal{O}(\log \ell_\infty)$ due to the double tricks, which is shown in Lemma 6.

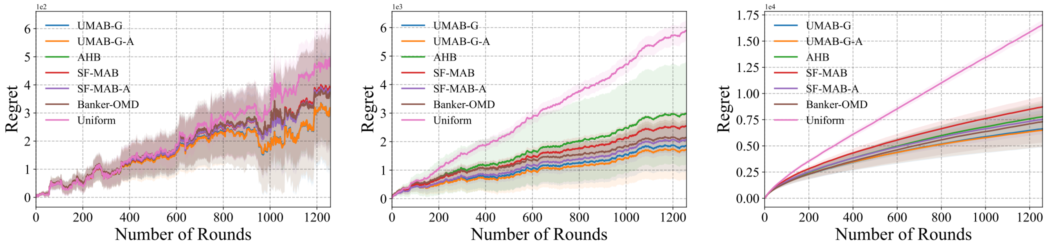
Summing the bounds for ①,②,③ gives Theorem 2.

4 Experiments

We now corroborate our theoretical improvements and testify the performance of our algorithms `UMAB-G` (Algorithm 2 with non-adaptive exploration) and `UMAB-G-A` (Algorithm 2 with adaptive exploration). We compare to **all** existing scale-free/unbounded AMAB algorithms, including `SF-MAB` (Putta & Agrawal, 2022), `SF-MAB-A` (Putta & Agrawal, 2022), `AHB` (Hadiji & Stoltz, 2020), and `banker-OMD` (Huang et al., 2023). The figures show the average performance and standard deviations across 500 trails.

Applications to Stock Trading: In out first experiment, we consider an application to the stock market. Here we consider $n = 10$ stocks and $T = 1258$ rounds (daily price for 5-years). For every stock, its loss is the normalized price difference, i.e., the difference between two consecutive days for 100 shares. Stock prices are generally chaotic and the fluctuation can vary greatly among stocks and across time. The regret trajectories of the different algorithms are illustrated in Figure 1(a). Note that the regret of `UMAB-G` and `UMAB-G-A` is significantly smaller than that of other algorithms, especially when the number of rounds is large. This is because 1). Compared to (Putta & Agrawal, 2022), our algorithms tune the learning and exploration rate more carefully, resulting in a saving of $\mathcal{O}(\sqrt{n})$ term in theory and better empirical performance in practice. 2). Compared to (Huang et al., 2023), our exploration rate design ensures that the algorithms can perceive the changes in loss scale and adapt learning rate in time. 3). Compared to (Hadiji & Stoltz, 2020), our exploration design leads to smaller regret than mixing with uniform distribution.

Applications to Amazon Sales We further construct an experiment using Amazon sales data. Similar to the above, we consider $n = 10$ Amazon stores and $T = 1258$ rounds (weekly sales for 2-years). We assume that



(a) Stock Market Data (b) Amazon Sales Data (c) Meta Algorithm Selection

Figure 1: Real Data Experiments

in each round, each store randomly discloses the weekly sales of one of its departments. The loss is defined by the negative of the weekly sales. We generate 10 rounds of loss using one week's data. Notice that the loss we considered in this setting is completely negative. The simulation results are shown in Figure 1(b). As expected, our algorithms outperform all other competitors. Compared to the stock market example, the fluctuation of regret trajectories of Amazon sales data is more stable for all the algorithms. This is because changes in Amazon store sales are more gradual than those in stocks: since all the algorithms we consider in the experiment are based on the FTRL/OMD framework, such a loss sequence will induce a stable action distribution, thereby resulting in the smoothness of the regret curve.

Applications to Model Selection In the last setting, we explore an application to the model selection problem. We assume that we have access to $n = 10$ linear regression meta-algorithms (SGD with different learning rate). Similarly to the above, we set the number of rounds $T = 1258$. In each round t, the meta-algorithms output the training loss error based on a dataset of size t. Notice that since the size of the data set varies in each round, the optimal meta-algorithm will also change. In this scenario, the regret measures whether a model selection algorithm can promptly detect the change in the optimal meta-algorithm. Moreover, the prediction error can be very large when the data set is small. The results are shown in Figure 1(c). Again, the regrets of our algorithms are strictly smaller than all baselines. Compared to the first two experiments, the regret trajectories are smoother because of the stochastic nature of the loss sequence as t increases.

The above experiments demonstrate the effectiveness of our algorithms against several different loss sequence patterns. However, one observation is that the adaptive and non-adaptive version of our algorithm perform quite similarly in all three experiments, and there is no evidence to suggest the significance of extra exploration. In Appendix A, we perform an ablation study to illustrate the impact of extra exploration. We also construct a deeper comparison between the adaptive and non-adaptive version of our algorithm, and discuss their respective strength and weakness.

5 Conclusion

We proposed the first algorithms that achieve optimal adaptive and non-adaptive regrets in adversarial multi-armed bandit problem with unbounded losses. Real data experiments validate our theoretical findings and demonstrate the superior performance of our algorithms compared to all existing algorithms for unbounded losses. Future work include extending our algorithmic tools to more challenging settings such as contextual bandit and reinforcement learning.

REFERENCES

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Algorithmic Learning Theory: 17th International Conference, ALT 2006, Barcelona, Spain, October 7-10, 2006. Proceedings 17*, pp. 229–243. Springer, 2006.

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

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pp. 111–127. PMLR, 2018.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pp. 508–528. PMLR, 2019.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pp. 874–894. PMLR, 2019.

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

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Hédi Hadiji and Gilles Stoltz. Adaptation to the range in k-armed bandits. *arXiv preprint arXiv:2006.03378*, 2020.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, 20, 2007.

Jiatai Huang, Yan Dai, and Longbo Huang. Banker online mirror descent: A universal approach for delayed online bandit learning. *arXiv preprint arXiv:2301.10500*, 2023.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pp. 2552–2583. PMLR, 2021.

Andrew Jacobsen and Ashok Cutkosky. Unconstrained online learning with unbounded losses. *arXiv preprint arXiv:2306.04923*, 2023.

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

Jérémie Mary, Romaric Gaudel, and Philippe Preux. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data: First International Workshop, MOD 2015, Taormina, Sicily, Italy, July 21-23, 2015, Revised Selected Papers 1*, pp. 325–336. Springer, 2015.

Jack J Mayo, Hédi Hadiji, and Tim van Erven. Scale-free unconstrained online learning for curved losses. In *Conference on Learning Theory*, pp. 4464–4497. PMLR, 2022.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and Dávid Pál. Scale-free algorithms for online linear optimization. In *Algorithmic Learning Theory: 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings*, pp. 287–301. Springer, 2015.

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

Roman Pogodin and Tor Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Uncertainty in Artificial Intelligence*, pp. 894–904. PMLR, 2020.

Sudeep Raja Putta and Shipra Agrawal. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 910–930. PMLR, 2022.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

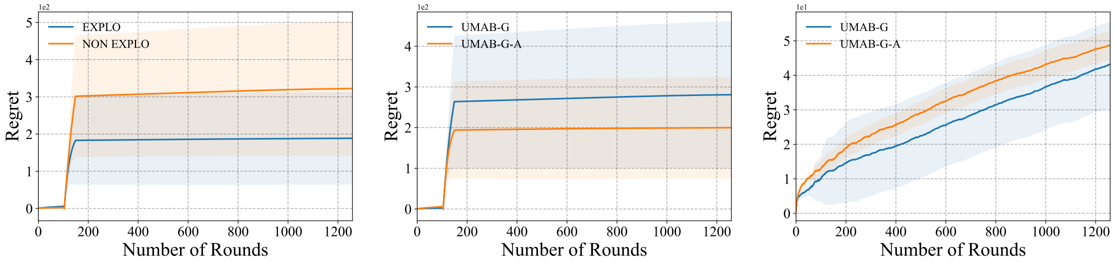
Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

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

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.



(a) Extra Exploration Benefits (b) Adaptive Better (c) Non-Adaptive Better

Figure 2: Impact of Extra Exploration with Non-Adaptive/Adaptive Rates

A Additional Experiments

In this section, we conduct additional experiments investigating the effect of extra exploration and the pros and cons of adaptive and non-adaptive exploration scheme.

A.1 Impact of extra exploration

First, we demonstrate the importance of extra exploration for unbounded loss. Consider a problem with two arms $n = 2$ and set $T = 1258$. We design the following loss sequence:

$$\ell_t = \begin{cases} [0, -0.5]^\top, & \text{if } 1 \leq t < 100 \\ [-10, 0]^\top, & \text{if } 100 \leq t < 150 \\ [-0.05, 0]^\top, & \text{if } 150 \leq t < 1258 \end{cases} \tag{8}$$

The intuition is to try deceive algorithms into taking the second arm as the "superior option" in the initial rounds which reduces the frequency of algorithms pulling the first arm, and thus hindering algorithms ability to detect the changes of the optimal arm. Especially, considering the loss can be unbounded, failing to detect the changes is costly. In this case, the regret trajectories are provided in Figure 2(a), where the comparison is between UMAB-G-A and our algorithm with no extra exploration. It suffices to note that the algorithm with extra exploration performs much better than the one without extra exploration. This is consistent with the intuition of our design: extra exploration ensures that each arm has a probability of being pulled, so that the algorithm can always perceive the changes in the losses and adjust its learning rate in relatively few rounds.

A.2 Comparison between UMAB-G and UMAB-G-A

Next, let's investigate the difference between our algorithms with non-adaptive and adaptive exploration rates. Intuitively, adaptive exploration rate is usually larger than the non-adaptive rate because it is of order $O(1/\sqrt{t})$ instead of $O(1/\sqrt{T})$ (assuming $\ell_\infty \ll T$). This makes adaptive exploration perform better in adversary cases, e.g. as shown in Figure 2(b), where we use the same loss sequence in (8). However, if the loss sequence is not adversary, e.g. there exists one arm that is always better than the others, non-adaptive exploration will be better since it loses less in extra exploration. An example is illustrated in Figure 2(c), where we use stochastic loss with expectation $[1, 0]^\top$. In summary, adaptive and non-adaptive have their own advantages under different loss sequences in practice.

B Additional Discussion about closely related works

B.1 Detailed comparison to Putta & Agrawal (2022)

In this subsection, we provide a detailed comparison between our work and Putta & Agrawal (2022) since it is the most closely related work to ours. Both works are based on FTRL-type algorithms design, and both consider non-adaptive and adaptive extra exploration. The key idea of Putta & Agrawal (2022) is to bound (3) by $\mathcal{O}(\ell^2_{t,k_t}/p_{t,k})$, resulting in an expectation regret $\mathcal{O}(\|\ell_t\|_2)$. In our work, we refine the analysis of (3), improving the bound to $\mathcal{O}(\ell^2_{t,k_t})$, where the expectation is bounded by $\mathcal{O}(\|\ell_t\|^2_\infty)$. Considering the worst case scenario where $\|\ell_t\|^2_2 = n\|\ell_t\|^2_\infty$, our algorithm saves $\sqrt{n}$ in the regret.

Furthermore, Putta & Agrawal (2022) choose a uniform distribution for extra exploration. This approach ensures an exploration error ($\textcircled{2}$ in this paper) of $\mathcal{O}(\ell_\infty\sqrt{nT})$ in non-adaptive case. However, for adaptive case, mixing a uniform distribution results in a large variance in the analysis of the exploration error. The proof idea of Putta & Agrawal (2022) can be summarized as (under our notations definition)

$$\langle \hat{\ell}'_t, \mathbf{c}_t \rangle \leq \|\hat{\ell}'_t\|_\infty \|\mathbf{c}_t\|_1 \leq \ell_\infty \sqrt{nT} \cdot n = \ell_\infty n^{3/2}\sqrt{T},$$

which is suboptimal in n. In this study, we design to a new exploration strategy, as described in Algorithm 3. By Lemma 10, we bound $\langle \hat{\ell}'_t, \mathbf{c}_t \rangle$ by $\mathcal{O}(\ell_\infty(\sqrt{nT} + n^2))$, which is optimal in n for large enough T. In summary, the algorithms presented in this article offer a $\mathcal{O}(\sqrt{n})$ improvement of the regret over Putta & Agrawal (2022), in both non-adaptive and adaptive settings, thanks to both our novel exploration strategy and tighter analysis.

B.2 The error in Banker-OMD (Huang et al., 2023)

Huang et al. (2023) shared a similar clipping (skipping) idea with us. In Lemma 4.2 of (Huang et al., 2023), the authors control the regret of the general case by the regret of the non-negative case directly (Theorem 4.2 of (Huang et al., 2023)). In this case, the authors bounded the clipping error (i.e., $\textcircled{3}$ in this paper) by

$$\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \leq \langle \hat{\ell}_t, \mathbf{p}_t \rangle \leq \ell_\infty.$$

However, notice that the above only holds true if $\hat{\ell}_t \geq 0$. When $\hat{\ell}_t < 0$, -$\langle \hat{\ell}_t, \mathbf{p}^\dagger \rangle$ is positive and on the order of $1/p_{t,k_t}$, which can be arbitrarily unbounded. In this case, their regret will always include a $\mathcal{O}(1/p_{t,k_t})$ term and thus be unbounded. We have confirmed this with the authors of Huang et al. (2023), and indeed they have made the mistake in their proof. So their current analysis for the general loss setting does not work.

One may think that the issue can be solved by analyzing the regret using ℓ_t instead of $\hat{\ell}_t$, i.e.,

$$\mathbb{E}[\langle \ell_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle] = \mathbb{E}[\mathbb{1}_{\neg clip}\langle \ell_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle] + \mathbb{E}[\mathbb{1}_{\neg clip}\langle \ell_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle]$$

where $\mathbb{1}_{\neg clip}$ denotes the probability of the clipping happening. Using the proof of (Huang et al., 2023), it suffices to show the second term can be bounded by $\mathcal{O}(\ell_\infty \log \ell_\infty)$. It might be intuitive to think that the first term can also be bounded by using $\hat{\ell}'$ to estimate $\mathbb{1}_{\neg clip(t,k)}\ell_t$. However, we note that

$$\begin{aligned}\mathbb{E}[\langle \hat{\ell}_t', \mathbf{p}_t - \mathbf{p}^\dagger\rangle] &= \sum_k p_{t,k}\Big(\frac{\mathbb{1}_{\neg clip(t,k)}\ell_{t,k}}{p_{t,k}}p_{t,k} - \frac{\mathbb{1}_{\neg clip(t,k)}\ell_{t,k}}{p_{t,k}}\mathbb{1}(k = k^*)\Big)\\ &= \sum_k \mathbb{1}_{\neg clip(t,k)}\ell_{t,k}x_{t,k} - \sum_k \mathbb{1}_{\neg clip(t,k)}\ell_{t,k}\mathbb{1}(k = k^*)\\ &= \sum_k \mathbb{1}_{\neg clip(t,k)}\ell_{t,k}x_{t,k} - \mathbb{1}_{\neg clip(t,k^\star)}\ell_{t,k^\star}\\ &\neq \sum_k x_{t,k}\Big(\mathbb{1}_{\neg clip(t,k)}(\ell_{t,k} - \ell_{t,k^\star})\Big)\\ &= \mathbb{E}[\mathbb{1}_{clip(t)}\langle \ell_t, x_t - y\rangle],\end{aligned}$$

which implies that $\hat{\ell}_t'$ is not an unbiased estimator of $\mathbb{1}_{\neg clip(t,k)}\ell_t$, so this route does not work. Therefore, as far as we can see, there doesn't exist a clear way of fixing the proof in Huang et al. (2023) to make their results match ours.

In our paper, we avoid issue by adding extra exploration to upper bound $\|\hat{\ell}_t\|_\infty$. We suspect such explicit exploration is inevitable for no-regret learning under the unbounded losses (Bubeck et al., 2012). Besides this issue, our differences and improvements compared to (Huang et al., 2023) mainly include: (1). Our results reveal an asymmetry between positive and negative losses in the AMAB problem. Especially, there is no clipping in our algorithm UMAB-NN, which greatly simplifies the algorithms in (Huang et al., 2023). (2). The space complicity of our algorithms is $\mathcal{O}(n)$ because the algorithm only needs to maintain a constant number of $\mathbb{R}^n$ vectors. In contrast, the space complexity of (Huang et al., 2023) is $\mathcal{O}(T^2)$ due to the necessity of keeping a weight matrix of size $T \times T$.

C Proof of Theorem 1

C.1 Main proof and statement of technical lemmas

Recall (1), the expected regret is denoted by

$$\mathbb{E}\Big[\sum_{t=1}^T \ell_{t,k_t} - \min_{k\in[n]}\sum_{t=1}^T \ell_{t,k}\Big] = \mathbb{E}\Big[\sum_{t=1}^T\langle \ell_t, \mathbf{p}_t - \mathbf{p}^\star\rangle\Big] = \mathbb{E}\Big[\sum_{t=1}^T\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle\Big] + \sum_{t=1}^T\langle \ell_t, \mathbf{p}^\dagger - \mathbf{p}^\star\rangle,$$

where $\mathbf{p}^\star$ denote the best fixed strategy. Especially, we consider

$$\mathbf{p}^\dagger = \Big(1 - \frac{1}{T}\Big)\mathbf{p}^\star + \frac{\mathbf{1}_n}{nT}.$$

where $\mathbf{1}_n$ is the all-ones vector. It is obvious that $\mathbf{p}^\dagger \in \Delta_n$. In this case, there is

$$\langle \ell_t, \mathbf{p}^\dagger - \mathbf{p}^\star\rangle \le \langle \ell_t, \frac{\mathbf{1}_n}{nT} - \frac{1}{T}\mathbf{p}^\star\rangle \le \frac{1}{nT}\langle \ell_t, \mathbf{1}_n\rangle \le \frac{1}{nT}\|\ell_t\|_1 \le \frac{\ell_\infty}{T},$$

where the second inequality is due to $\ell_t \ge 0$ by assumption. Thus we have $\sum_{t=1}^T\langle \ell_t, \mathbf{p}^\dagger - \mathbf{p}^\star\rangle \le \ell_\infty$. It suffices to focus on $\sum_{t=1}^T\langle \ell_t, \mathbf{p}_t - \mathbf{p}^\star\rangle$. Recall (4), there is

$$\begin{aligned}\sum_{t=1}^T\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle &\le \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \sum_{t=1}^T \min\Big(\frac{1}{2}\eta_t\ell_{t,k_t}^2, |\ell_{t,k_t}|\Big)\\ &\le \frac{n\log(nT)}{\eta_{T+1}} + \sum_{t=1}^T \min\Big(\frac{1}{2}\eta_t\ell_{t,k_t}^2, |\ell_{t,k_t}|\Big).\end{aligned}$$

where the second inequality is because all entries of $\mathbf{p}^\dagger$ are no less than $1/nT$ by definition.

It remains to bound $\sum_{t=1}^T \min\left(\frac{1}{2}\eta_t \ell_{t,k_t}^2, |\ell_{t,k_t}|\right)$. The proof relies on a technical lemma from (Orabona & Pál, 2018).

Lemma 3 *((Orabona & Pál, 2018) Lemma 3) Let $a_1, \dots, a_T \geq 0$. Then*

$$\sum_{t=1}^T \min\Big(\frac{a_t^2}{\sqrt{\sum_{s=1}^{t-1} a_s^2}}, a_t\Big) \leq 3.5\sqrt{\sum_{t=1}^T a_t^2} + 3.5 \max_{t\in[T]} a_t$$

Using Lemma 3 and $\eta_0, \dots, \eta_T$ as in Algorithm 1, we have

$$\begin{aligned}
\sum_{t=1}^T \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle &\leq \frac{n\log(nT)}{\eta_{T+1}} + \sum_{t=1}^T \min\Big(\frac{1}{2}\eta_t \ell_{t,k_t}^2, |\ell_{t,k_t}|\Big) \\
&\leq \frac{1}{2}\sqrt{n\sum_{t=1}^T \ell_{t,k_t}^2 \log(nT)} + \sum_{t=1}^T \min\Big(\sqrt{\frac{n}{\sum_{s=1}^{t-1}\ell_{s,k_s}^2}}\ell_{t,k_t}^2, |\ell_{t,k_t}|\Big) \\
&\leq \frac{1}{2}\sqrt{n\sum_{t=1}^T \ell_{t,k_t}^2 \log(nT)} + \sqrt{n}\sum_{t=1}^T \min\Big(\frac{\ell_{s,k_t}^2}{\sqrt{\sum_{s=1}^{t-1}\ell_{s,k_s}^2}}, |\ell_{t,k_t}|\Big) \\
&\leq \frac{1}{2}\sqrt{n\sum_{t=1}^T \ell_{t,k_t}^2 \log(nT)} + 3.5\sqrt{n}\Big(\sqrt{\sum_{t=1}^T \ell_{t,k_t}^2} + \max_{t\in[T]}|\ell_{t,k_t}|\Big) \\
&\leq 4\sqrt{n\sum_{t=1}^T \ell_{t,k_t}^2 \log(nT)} + 3.5\sqrt{n}\ell_\infty \\
&\leq 4\sqrt{n\sum_{t=1}^T \|\ell_t\|_\infty^2 \log(nT)} + 3.5\sqrt{n}\ell_\infty.
\end{aligned}$$

Note that the right hand side of the above is deterministic. Thus

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^T \ell_{t,k_t} - \min_{k\in[n]}\sum_{t=1}^T \ell_{t,k}\Big] &\leq 4\sqrt{n\sum_{t=1}^T \|\ell_t\|_\infty^2 \log(nT)} + 3.5\sqrt{n}\ell_\infty + \ell_\infty \\
&\leq \tilde{\mathcal{O}}\Big(\sqrt{n\sum_{t=1}^T \|\ell_t\|_\infty^2}\Big)
\end{aligned}$$

completes the proof.

C.2 Proof of technical Lemmas

C.2.1 Proof of Lemma 1

For notations simplicity, we denote by

$$\Psi_t(\mathbf{p}) = \frac{1}{\eta_t}\Psi(\mathbf{p}).$$

We first note

$$\begin{aligned}\sum_{t=1}^{T}\langle\hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle &= -F_{T+1}(\mathbf{p}^\dagger) + \Psi_{T+1}(\mathbf{p}^\dagger) + \sum_{t=1}^{T}\langle\hat{\ell}_t, \mathbf{p}_t\rangle \\ &= -F_{T+1}(\mathbf{p}^\dagger) + \Psi_{T+1}(\mathbf{p}^\dagger) - F_1(\mathbf{p}_1) + F_{T+1}(\mathbf{p}_{T+1}) \\ &+ \sum_{t=1}^{T}\left(F_t(\mathbf{p}_t) - F_{t+1}(\mathbf{p}_{t+1})\right) + \sum_{t=1}^{T}\langle\hat{\ell}_t, \mathbf{p}_t\rangle \\ &= -F_{T+1}(\mathbf{p}^\dagger) + \Psi_{T+1}(\mathbf{p}^\dagger) - F_1(\mathbf{p}_1) + F_{T+1}(\mathbf{p}_{T+1}) \\ &+ \sum_{t=1}^{T}\Big(F_t(\mathbf{p}_t) + \langle\hat{\ell}_t, \mathbf{p}_t\rangle - F_{t+1}(\mathbf{p}_{t+1})\Big)\end{aligned}$$

By definition, there is

$$\begin{aligned}F_{T+1}(\mathbf{p}_{T+1}) - F_{T+1}(\mathbf{p}^\dagger) &= \min_{\mathbf{p}\in\Delta_n} F_{T+1}(\mathbf{p}) - F_{T+1}(\mathbf{p}^\dagger) \le 0 \\ \Psi_{T+1}(\mathbf{p}^\dagger) - F_1(\mathbf{p}_1) &= \Psi_{T+1}(\mathbf{p}^\dagger) - \min_{\mathbf{p}\in\Delta_n} \Psi_1(\mathbf{p}) \le \Psi_{T+1}(\mathbf{p}^\dagger).\end{aligned}$$

Thus, we obtain

$$\sum_{t=1}^{T}\langle\hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle \le \Psi_{T+1}(\mathbf{p}^\dagger) + \sum_{t=1}^{T}\Big(F_t(\mathbf{p}_t) + \langle\hat{\ell}_t, \mathbf{p}_t\rangle - F_{t+1}(\mathbf{p}_{t+1})\Big)$$

Furthermore, we note that

$$\begin{aligned}F_t(\mathbf{p}_t) + \langle\hat{\ell}_t, \mathbf{p}_t\rangle - F_{t+1}(\mathbf{p}_{t+1}) &= \sum_{s=1}^{t}\langle\hat{\ell}_s, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p}_t) - \frac{1}{\eta_{t+1}}\Psi(\mathbf{p}_t) \\ &\le \sum_{s=1}^{t}\langle\hat{\ell}_s, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p}_t) - \frac{1}{\eta_t}\Psi(\mathbf{p}_t) \\ &= \langle\hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}),\end{aligned}$$

where the first inequality is due to the assumption $\eta_{t+1} \le \eta_t$. Combining the above concludes the proof.

C.2.2 Proof of Lemma 2

We first prove inequality (3). By Taylor's expansion,

$$F_t(\mathbf{p}_{t+1}) - F_t(\mathbf{p}_t) = \langle\nabla F_t(\mathbf{p}_t), \mathbf{p}_{t+1} - \mathbf{p}_t\rangle + \frac{1}{2}\|\mathbf{p}_{t+1} - \mathbf{p}_t\|^2_{\nabla^2 F_t(\xi_t)}.$$

where $\xi_t = \alpha \mathbf{p}_t + (1-\alpha)\mathbf{p}_{t+1}$ for some $\alpha \in [0,1]$. By definition,

$$\mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n} F_t(\mathbf{p}).$$

By KKT conditions, there exists some $\lambda_t \in \mathbb{R}$ such that

$$\mathbf{p}_t = \arg\min_{\mathbf{p}\in\mathbb{R}} \Big(F_t(\mathbf{p}) + \lambda_t(1 - \sum_{k=1}^{n} p_{t,k})\Big).$$

By the optimality of $\mathbf{p}_t$, we have

$$\nabla F_t(\mathbf{p}_t) + \lambda_t \mathbf{1}_n = 0,$$

which implies

$$\langle \nabla F_t(\mathbf{p}), \mathbf{p}_{t+1} - \mathbf{p}_t\rangle = \langle -\lambda_t \mathbf{1}_n, \mathbf{p}_{t+1} - \mathbf{p}_t\rangle = 0.$$

Thus, there is

$$F_t(\mathbf{p}_{t+1}) - F_t(\mathbf{p}_t) = \frac{1}{2}\|\mathbf{p}_{t+1} - \mathbf{p}_t\|^2_{\nabla^2 F_t(\xi_t)}.$$

Using the above,

$$\begin{aligned}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) &= \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle - \frac{1}{2}\|\mathbf{p}_{t+1} - \mathbf{p}_t\|^2_{\nabla^2 F_t(\xi_t)} \\ &\le \max_{\mathbf{p}\in\mathbb{R}} \Big(\langle \hat{\ell}_t, \mathbf{p}\rangle - \frac{1}{2}\|\mathbf{p}\|^2_{\nabla^2 F_t(\xi_t)}\Big) \\ &\le \frac{1}{2}\|\hat{\ell}_t\|^2_{(\nabla^2 F_t(\xi_t))^{-1}} = \frac{1}{2}\eta_t \|\hat{\ell}_t\|^2_{(\nabla^2 \Psi(\xi_t))^{-1}},\end{aligned}$$

where the second inequality is because $\nabla^2\Psi(\xi_t)$ is a diagonal matrix and the second equality is due to $\nabla^2 F_t(\xi_t) = \nabla^2\Psi(\xi_t)/\eta_t$. Thus the proof of (3) is complete.

Now we prove

$$\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \le \frac{1}{2}\eta_t \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}$$

if $\hat{\ell}_t \in \mathbb{R}^n_+$. Recall

$$\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \sum_{k=1}^{n} \frac{\hat{\ell}^2_{t,k}}{\nabla^2_{k,k}\Psi(\xi_t)} = \sum_{k=1}^{n} \frac{\ell^2_{t,k}\mathbb{1}(k=k_t)}{p^2_{t,k}}\xi^2_{t,k} = \frac{\ell^2_{t,k_t}}{p^2_{t,k_t}}\xi^2_{t,k_t}$$

and ξ_t is between $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$, we prove case by case.

1. ($p_{t,k_t} - p_{t+1,k_t} < 0$): In this case, we have

$$\begin{aligned}\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) &\le \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1}\rangle \\ &= \hat{\ell}_{t,k_t}(p_{t,k_t} - p_{t+1,k_t}) \\ &\le 0 \le \frac{1}{2}\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}.\end{aligned}$$

   The first inequality is due to $\mathbf{p}_t$ minimizing F_t.

2. ($p_{t,k_t} - p_{t+1,k_t} \ge 0$): In this case, we have $\xi_{t,k_t} \le p_{t,k_t}$, and thus

$$\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} \le \ell^2_{t,k_t} = \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\mathbf{p}_t))^{-1}}$$

   completes the proof.

C.2.3 Proof of Lemma 3

The proof refers to Lemma 3 in (Orabona & Pál, 2018). Without loss of generality, we can assume $a_t > 0$, otherwise we can remove all $a_t = 0$ without affecting either side of the inequality. Let $M_t = \max_{s\in[t]} a_s$ and $M_0 = 0$. We aim to prove for any $\alpha > 1$

$$\min\Big(\frac{a_t^2}{\sqrt{\sum_{s=1}^{t-1} a_s^2}}, a_t\Big) \le 2\sqrt{1+\alpha^2}\Big(\sqrt{\sum_{s=1}^{t} a_s^2} - \sqrt{\sum_{s=1}^{t-1} a_s^2}\Big) + \frac{\alpha}{\alpha-1}(M_t - M_{t-1}).$$

from which Lemma 3 follows by summing over $t = 1, \dots, T$ and choosing $\alpha = \sqrt{2}$. The proof is based on case analysis.

1. $(a_t^2 \le \alpha^2 \sum_{s=1}^{t-1} a_s^2)$

$$\begin{aligned}\min\Big(\frac{a_t^2}{\sqrt{\sum_{s=1}^{t-1} a_s^2}}, a_t\Big) &\le \frac{a_t^2}{\sqrt{\sum_{s=1}^{t-1} a_s^2}} = \frac{a_t^2}{\sqrt{\frac{1}{1+\alpha^2}(\alpha^2 \sum_{s=1}^{t-1} a_s^2 + \sum_{s=1}^{t-1} a_s^2)}} \\ &\le \frac{a_t^2(1+\alpha^2)}{\sqrt{a_t^2 + \sum_{s=1}^{t-1} a_s^2}} \le 2\sqrt{1+\alpha^2}\Big(\sqrt{\sum_{s=1}^{t} a_s^2} - \sqrt{\sum_{s=1}^{t-1} a_s^2}\Big)\end{aligned}$$

   where the last inequality is by $x^2/\sqrt{x^2+y^2} \le 2(\sqrt{x^2+y^2} - \sqrt{y^2})$.

2. $(a_t^2 > \alpha^2 \sum_{s=1}^{t-1} a_s^2)$

$$\min\Big(\frac{a_t^2}{\sqrt{\sum_{s=1}^{t-1} a_s^2}}, a_t\Big) \le a_t = \frac{\alpha a_t - a_t}{\alpha - 1} \le \frac{\alpha}{\alpha-1}\Big(a_t - \sqrt{\sum_{s=1}^{t-1} a_s^2}\Big) \le \frac{\alpha}{\alpha-1}(M_t - M_{t-1}),$$

   where we use $a_t = M_t$ and $M_{t-1} \le \sqrt{\sum_{s=1}^{t-1} a_s^2}$.

D Proof of Theorem 2

D.1 Main proof and statement of technical lemmas

We begin by presenting the lemma statements that were left out in Section 3.2.

Lemma 4 *Given any action sequence $k_1, \dots, k_T$, if $\ell_{t,k_t} \le 0$. there is $p_{t,k_t} \le 2p'_{t,k_t}$ and $p_{t+1,k_t} \le 6p_{t,k_t}$ for every $t \in [T]$.*

Lemma 5 *With the non-adaptive and adaptive exploration rates as in Algorithm 3, we have*

$$\begin{aligned}&\textit{Non-Adaptive:} && \mathbb{E}\Big[\sum\nolimits_t \langle \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle\Big] \le 2\sqrt{n\sum\nolimits_t \|\ell_t\|_\infty^2},\\ &\textit{Adaptive} && \mathbb{E}\Big[\sum\nolimits_t \langle \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle\Big] \le 2n^2\ell_\infty + 2\sqrt{1 + 4n\sum\nolimits_t \|\ell_t\|_\infty} + 2\ell_\infty\sqrt{n\sum\nolimits_t \|\ell_t\|_\infty}.\end{aligned}$$

Lemma 6 *Given any action sequence $k_1, \dots, k_T$, with the non-adaptive and adaptive exploration rates as in Algorithm 3, we have*

$$\textit{Non-Adaptive:} \qquad \mathbb{E}\Big[\sum\nolimits_t \langle \hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger \rangle\Big] \leq \ell^-_\infty(2n^2 + \sqrt{nT}) \log_2(1+\ell_\infty),$$

$$\textit{Adaptive:} \qquad \mathbb{E}\Big[\sum\nolimits_t \langle \hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger \rangle\Big] \leq \ell^-_\infty\Big(2n^2 + 3\sqrt{n\sum\nolimits_t \|\ell_t\|_\infty}\Big) \log_2(1+\ell_\infty).$$

Now we give the detailed proof of Theorem 2. By Lemma 2, we have

$$\langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \leq \min\Big(\frac{1}{2}\eta_t {\ell'}^2_{t,k_t}, |\ell'_{t,k_t}|\Big).$$

if $\ell_{t,k_t} \geq 0$. Alternatively, when $\ell_{t,k_t} < 0$, by Lemma 2 and 4 and inequality (5), we have

$$\begin{aligned}
\langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) &\leq \frac{1}{2}\eta_t \|\hat{\ell}'_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \frac{1}{2}\eta_t \frac{{\ell'}^2_{t,k_t}}{{p'}^2_{t,k_t}} \xi^2_{t,k_t} \\
&\leq \frac{1}{2}\eta_t {\ell'}^2_{t,k_t} \frac{p^2_{t,k_t}}{{p'}^2_{t,k_t}} \frac{\max(p^2_{t,k_t}, p^2_{t+1,k_t})}{p^2_{t,k_t}} \\
&\leq 72\eta_t {\ell'}^2_{t,k_t}.
\end{aligned}$$

Moreover, we further note by Lemma 4,

$$\begin{aligned}
\langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) &\leq \langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle \\
&\leq \left|\frac{{\ell'}_{t,k_t}}{p_{t,k_t}}\right| \left|\frac{p_{t,k_t}}{{p'}_{t,k_t}}\right| |p_{t,k_t} - p_{t+1,k_t}| \\
&\leq \left|\frac{{\ell'}_{t,k_t}}{p_{t,k_t}}\right| \left|\frac{p_{t,k_t}}{{p'}_{t,k_t}}\right| |5p_{t,k_t}| \leq 10|{\ell'}_{t,k_t}|.
\end{aligned}$$

Combining the above we have

$$\langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \leq 18 \min\Big(4\eta_t {\ell'}^2_{t,k_t}, |\ell'_{t,k_t}|\Big)$$

for any $\ell_t \in \mathbb{R}^n$. Using a similar proof as in Theorem 1, we have

$$\begin{aligned}
\sum_{t=1}^T \langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle &\leq \frac{n\log(nT)}{\eta_{T+1}} + 18\min\Big(4\eta_t {\ell'}^2_{t,k_t}, |\ell'_{t,k}|\Big) \\
&\leq 4\sqrt{2n^2\ell^2_\infty + n\sum_{t=1}^T {\ell'}^2_{t,k_t}} \log(nT) + 18\sqrt{n}\min\Big(\frac{{\ell'}^2_{t,k_t}}{\sqrt{\sum_{s=1}^{t-1} {\ell'}^2_{s,k_s}}}, |\ell'_{t,k}|\Big) \\
&\leq 4\sqrt{2n^2\ell^2_\infty + n\sum_{t=1}^T {\ell'}^2_{t,k_t}} \log(nT) + 63\sqrt{n}\Big(\sqrt{\sum_{t=1}^T {\ell'}^2_{t,k_t}} + \max_{t\in[T]} |\ell'_{t,k}|\Big) \\
&\leq 67\sqrt{2n^2\ell^2_\infty + n\sum_{t=1}^T {\ell'}^2_{t,k_t}} \log(nT) + 63\sqrt{n}\max_{t\in[T]} |\ell'_{t,k}| \\
&\leq 67\sqrt{2n^2\ell^2_\infty + n\sum_{t=1}^T \|\ell_t\|^2_\infty} \log(nT) + 63\sqrt{n}\ell_\infty.
\end{aligned}$$

The last inequality is because $|{\ell'}_{t,k_t}| \le |\ell_{t,k_t}|$. In short, we can bound

$$\mathbb{E}\Big[\sum_{t=1}^{T}\langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle\Big] \le \tilde{\mathcal{O}}\Big(\sqrt{n^2\ell_\infty^2 + n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}\Big). \tag{9}$$

Now we summarize all the results.

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T}\ell_{t,k_t} - \min_{k\in[n]}\sum_{t=1}^{T}\ell_{t,k}\Big]\\
=&\mathbb{E}\Big[\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}'_t - \mathbf{p}^\star\rangle\Big]\\
\le&\mathbb{E}\Big[\sum_{t=1}^{T}\langle \hat{\ell}_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle\Big] + \ell_\infty\\
\le&\mathbb{E}\Big[\sum_{t=1}^{T}\langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger\rangle\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}\langle \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}\langle \hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle\Big] + \ell_\infty.
\end{aligned}$$

Based on Lemma 5 and 6 and inequality (9), we have

1. (Non-Adaptive):

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T}\ell_{t,k_t} - \min_{k\in[n]}\sum_{t=1}^{T}\ell_{t,k}\Big]\\
\le&\tilde{\mathcal{O}}\Big(\sqrt{n^2\ell_\infty^2 + n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}\Big) + \tilde{\mathcal{O}}\Big(\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}\Big) + \tilde{\mathcal{O}}\Big(\ell_\infty^-(n^2 + \sqrt{nT})\Big)\\
=&\tilde{\mathcal{O}}\Big(\ell_\infty n^2 + \sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2} + \ell_\infty^-\sqrt{nT}\Big).
\end{aligned}$$

2. (Adaptive):

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T}\ell_{t,k_t} - \min_{k\in[n]}\sum_{t=1}^{T}\ell_{t,k}\Big]\\
\le&\tilde{\mathcal{O}}\Big(\sqrt{n^2\ell_\infty^2 + n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}\Big) + \tilde{\mathcal{O}}\Big(\ell_\infty\Big(n^2 + \sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}\Big) + \sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}\Big)\\
&+ \tilde{\mathcal{O}}\Big(\ell_\infty^-\Big(n^2 + \sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}\Big)\Big)\\
=&\tilde{\mathcal{O}}\Big(\ell_\infty n^2 + \sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2} + \ell_\infty\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty} + \sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}\Big).
\end{aligned}$$

D.2 Proof of technical Lemmas

D.2.1 Proof of Lemma 4

The first inequality $p_{t,k_t} \leq 2p'_{t,k_t}$ can be easily verified. Recall $\mathbf{p}'_t = \mathbf{p}_t + \rho_t \mathbf{c}_t$ and $k^\star_t \in \arg\max_{k' \in [n]} p_{t,k'}$ as in Algorithm 3, it suffices to focus on the case $k_t = k^\star_t$, otherwise $p_{t,k_t} \leq p'_{t,k_t}$. When $k_t = k^\star_t$, we note that

$$p'_{t,k_t} = p_{t,k_t} + \rho_t c_{t,k_t} \geq p_{t,k_t} - \frac{1}{2n^2} n = p_{t,k_t} - \frac{1}{2n}.$$

The first inequality is due to $\rho_t \leq 1/2n^2$ and $c_{t,k_t} \geq -n$ by definition. Moreover, there is

$$p_{t,k_t} \in \arg\max_{k' \in [n]} p_{t,k'} \geq \frac{1}{n}.$$

Thus

$$p_{t,k_t} \leq p_{t,k_t} + p_{t,k_t} - \frac{1}{n} = 2\Big(p_{t,k_t} - \frac{1}{2n}\Big) = 2p'_{t,k_t}$$

completes the proof.

The proof of the second inequality relies on the following two technical lemmas.

Lemma 7 *Given any $L \in \mathbb{R}^n$ and $k \in [n]$, consider*

$$\mathbf{x} = \arg\min_{\mathbf{p} \in \Delta_n} \Big(\langle L, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p})\Big)$$
$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{p} \in \Delta_n} \Big(\langle L + \frac{l}{x_k}\mathbf{e}_k, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p})\Big)$$

where x_k is the kth entry of $\mathbf{x}$. If

$$-\frac{1}{2\eta} \leq l \leq 0,$$

then

$$\tilde{x}_k \leq 2x_k.$$

Lemma 8 *Given any $L \in \mathbb{R}^n$, consider*

$$\mathbf{x} = \arg\min_{\mathbf{p} \in \Delta_n} \Big(\langle L, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p})\Big)$$
$$\mathbf{x}' = \arg\min_{\mathbf{p} \in \Delta_n} \Big(\langle L, \mathbf{p} \rangle + \frac{1}{\eta'}\Psi(\mathbf{p})\Big),$$

if

$$\eta' \leq \eta \leq C\eta',$$

for some $C > 0$, then

$$x'_k \leq Cx_k,\ \forall k \in [n].$$

Now we use Lemma 7 and 8 to bound the magnitude of $p_{t+1,k_t}/p_{t,k_t}$. Recall the update rule of action distribution

$$\mathbf{p}_t = \arg\min_{\mathbf{p}\in\Delta_n} \Big(\langle\sum_{s=1}^{t-1}\hat{\ell}'_s, \mathbf{p}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p})\Big),$$

$$\mathbf{p}_{t+1} = \arg\min_{\mathbf{p}\in\Delta_n} \Big(\langle\hat{\ell}'_t + \sum_{s=1}^{t-1}\hat{\ell}'_s, \mathbf{p}\rangle + \frac{1}{\eta_{t+1}}\Psi(\mathbf{p})\Big).$$

Define the intermediate distribution

$$\tilde{\mathbf{p}}_t = \arg\min_{\mathbf{p}\in\Delta_n} \Big(\langle\hat{\ell}'_t + \sum_{s=1}^{t-1}\hat{\ell}'_s, \mathbf{p}\rangle + \frac{1}{\eta_t}\Psi(\mathbf{p})\Big).$$

Notice that $\hat{\ell}'_t = \ell'_{t,k_t}\mathbf{1}_{k_t}/p_{t,k_t}$. Denote by $L = \sum_{s=1}^{t-1}\hat{\ell}'_s$, by Lemma 7, $\tilde{p}_{t,k_t}/p_{t,k_t} \le 2$ if $-1/2\eta_t \le \ell'_{t,k_t} \le 0$. Moreover, by Lemma 8, $p_{t+1,k_t}/p_{t,k_t} \le 3$ if $\eta_{t+1} \le \eta_t \le 3\eta_{t+1}$. Combining these two results leads to $p_{t+1,k_t}/p_{t,k_t} \le 6$, which completes the proof. Therefore, it remains to show that the two conditions hold.

We first prove $-1/2\eta_t \le \ell'_{t,k_t} \le 0$. Recall

$$\eta_t = \frac{1}{4}\sqrt{\frac{n}{nC_t^2 + \sum_{s=1}^{t-1}{\ell'}^2_{s,k_s}}}.$$

We have

$${\ell'}^2_{t,k_t} \le 4C_t^2 \le 4\Big(\frac{nC_t^2 + \sum_{s=1}^{t-1}{\ell'}^2_{s,k_s}}{n}\Big) \le \frac{1}{4\eta_t^2},$$

where the first inequality is by the assumption $\ell_{t,k_t} \le 0$, which implies $\ell'_{t,k_t} \le 0$, and the clipping rule (line 5 of Algorithm 2).

Then we show $\eta_{t+1} \le \eta_t \le 3\eta_{t+1}$. Since $\eta_{t+1} \le \eta_t$ is trivial, it suffices to prove $\eta_t \le 3\eta_{t+1}$. Notice that

$$C_{t+1}^2 = \min\Big(C_t^2, {\ell'}^2_{t,k_t}\Big) \le \min\Big(C_t^2, 4C_t^2\Big) = 4C_t^2.$$

Thus,

$$\begin{aligned}\eta_t &= \frac{1}{4}\sqrt{\frac{n}{nC_t^2 + \sum_{s=1}^{t-1}{\ell'}^2_{s,k_s}}}\\ &= \frac{3}{4}\sqrt{\frac{n}{9nC_t^2 + 9\sum_{s=1}^{t-1}{\ell'}^2_{s,k_s}}}\\ &\le \frac{3}{4}\sqrt{\frac{n}{4nC_t^2 + 4nC_t^2 + \sum_{s=1}^{t-1}{\ell'}^2_{s,k_s}}}\\ &\le \frac{3}{4}\sqrt{\frac{n}{nC_{t+1}^2 + {\ell'}^2_{t,k_t} + \sum_{s=1}^{t-1}{\ell'}^2_{s,k_s}}}\\ &= 3\eta_{t+1}.\end{aligned}$$

completes the proof.

D.2.2 Proof of Lemma 5

Non-Adaptive exploration:

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T}\langle\hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle\Big] &= \mathbb{E}\Big[\sum_{t=1}^{T}\rho_t\langle\hat{\ell}'_t, \mathbf{c}_t\rangle\Big] \\
&\le \mathbb{E}\Big[\sum_{t=1}^{T}\rho_t\langle|\hat{\ell}_t|, |\mathbf{c}_t|\rangle\Big] \\
&= \sum_{t=1}^{T}\rho_t\langle|\ell_t|, |\mathbf{c}_t|\rangle \\
&\le 2n\sum_{t=1}^{T}\frac{\|\ell_t\|_\infty}{n^2+\sqrt{nT}} \\
&\le 2\sqrt{n}\frac{\sum_{t=1}^{T}\|\ell_t\|_\infty}{\sqrt{T}} \\
&\le 2\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}.
\end{aligned}$$

The first inequality is due to that $\hat{\ell}'_t$ is the truncation of $\hat{\ell}_t$, thus $|\hat{\ell}'_t| \le |\hat{\ell}_t|$. The last inequality is by Cauchy–Schwartz inequality.

Adaptive exploration: We first introduce two auxiliary lemmas.

Lemma 9 *Let $a_1, \ldots, a_T \ge 0$. Then*

$$\sum_{t=1}^{T}\frac{a_t}{\sqrt{2\sum_{s=1}^{t-1}a_s+1}} \le 2\sqrt{\sum_{t=1}^{T}a_t+1} + \max_{t\in[T]}(a_t).$$

Lemma 10 *Given any action sequence $k_1, \ldots, k_T$, with the adaptive exploration rate as in Algorithm 2, there is*

$$|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle| \le \ell_\infty\Big(2n^2 + \sqrt{2\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|}\Big).$$

The detailed proof of Lemma 9 and 10 would be provided later. Now we can prove Lemma 5.

$$\begin{aligned}
\sum_{t=1}^{T}\langle\hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle &\leq \sum_{t=1}^{T}\rho_t|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle| \\
&\leq \sum_{t=1}^{T}\frac{|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|}{\sqrt{1+2\sum_{s=1}^{t-1}|\langle\hat{\ell}'_s, \mathbf{c}_s\rangle|}} \\
&\leq 2\sqrt{1+2\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle| + \max_{t\in[T]}\Big(|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|\Big)} \\
&\leq 2\sqrt{1+2\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle| + \ell_\infty\Big(2n^2 + \sqrt{2\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|}\Big)}.
\end{aligned}$$

where the second inequality is due to $\rho_t = 1/(2n^2 + \sqrt{2\sum_{s=1}^{t-1}|\langle\hat{\ell}'_s, \mathbf{c}_s\rangle|}) \leq 1/(\sqrt{1+2\sum_{s=1}^{t-1}|\langle\hat{\ell}'_s, \mathbf{c}_s\rangle|}$, the third inequality is by Lemma 9 with $a_t = |\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|$. The last inequality is by Lemma 10. Taking expectation on the both sides, there is

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T}\langle\hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}_t\rangle\Big] &\leq \mathbb{E}\Big[2\sqrt{1+2\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|}\Big] + \mathbb{E}\Big[\ell_\infty\Big(2n^2 + \sqrt{2\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|}\Big)\Big] \\
&\leq 2n^2\ell_\infty + 2\sqrt{1+2\mathbb{E}\Big[\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|\Big]} + \ell_\infty\sqrt{2\mathbb{E}\Big[\sum_{t=1}^{T}|\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|\Big]} \\
&\leq 2n^2\ell_\infty + 2\sqrt{1+2\sum_{t=1}^{T}\langle\mathbb{E}\Big[|\hat{\ell}'_t|\Big], |\mathbf{c}_t|\rangle} + \ell_\infty\sqrt{2\sum_{t=1}^{T}\langle\mathbb{E}\Big[|\hat{\ell}'_t|\Big], |\mathbf{c}_t|\rangle} \\
&\leq 2n^2\ell_\infty + 2\sqrt{1+2\sum_{t=1}^{T}\langle|\ell_t|, |\mathbf{c}_t|\rangle} + \ell_\infty\sqrt{2\sum_{t=1}^{T}\langle|\ell_t|, |\mathbf{c}_t|\rangle} \\
&\leq 2n^2\ell_\infty + 2\sqrt{1+4n\sum_{t=1}^{T}\|\ell_t\|_\infty} + 2\ell_\infty\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}.
\end{aligned}$$

The second inequality is by using Jensen's inequality. The fourth inequality is because $\mathbb{E}\Big[|\hat{\ell}'_t|\Big] = |\ell'_t|$ and the magnitude of the truncation loss is not more than that of the original loss, i.e., $|\ell'_t| \leq |\ell_t|$. The last inequality is due to $\langle|\ell_t|, |\mathbf{c}_t|\rangle \leq \|\ell_t\|_\infty\|\mathbf{c}_t\|_1 \leq 2n\|\ell_t\|_\infty$. The whole proof is completed.

D.2.3 Proof of Lemma 6

Recall

$$\sum_{t=1}^{T}\langle\hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle \leq \sum_{t=1}^{T}\|\hat{\ell}_t - \hat{\ell}'_t\|_1 \leq \sum_{t=1}^{T}\|\hat{\ell}_t\|_1\mathbb{1}(\hat{\ell}_t \neq \hat{\ell}'_t).$$

where the last inequality is due to $\|\hat{\ell}_t - \hat{\ell}'_t\|_1 \leq \|\hat{\ell}_t\|_1$ by the clipping property. We note that the clipping occurs only if $\hat{\ell}_t \leq 0$ and $\hat{\ell}_{t,k_t} \leq \ell_{t,k_t}/\rho_t$ for every $t \in [T]$ by extra exploration. Thus,

$$\sum_{t=1}^{T} \|\hat{\ell}_t\|_1 \mathbb{1}(\hat{\ell}_t \neq \hat{\ell}'_t) \leq \sum_{t=1}^{T} \frac{|\min(\ell_{t,k_t}, 0)|}{\rho_t} \mathbb{1}(\hat{\ell}_t \neq \hat{\ell}'_t) \leq \frac{\ell_\infty^-}{\rho_{T+1}} \sum_{t=1}^{T} \mathbb{1}(\hat{\ell}_t \neq \hat{\ell}'_t).$$

It suffices to prove $\sum_{t=1}^{T} \mathbb{1}(\hat{\ell}_t \neq \hat{\ell}'_t) \leq \log_2(1+\ell_\infty)$. Notice that $\hat{\ell}_t \neq \hat{\ell}'_t$ will happen if and only if

$$\ell_{t,k_t} \leq 2C_t.$$

In this case, we have

$$C_{t+1} = 2C_t.$$

Now we need to get an upper bound on the size of C_T. In Algorithm 2, C_t will be updated (i.e., $C_t \neq C_{t+1}$) if and only if the received loss $\ell_{t,k_t} < C_t$. When C_t is updated, we can note that $C_{t+1} \geq \ell_{t,k_t}$ holds, which also means $|C_{t+1}| \leq |\ell_{t,k_t}|$. Thus, we have

$$|C_T| \leq \max_{t \in [T]}(1, |\ell_{t,k_t}|) \leq 1 + \ell_\infty.$$

Since $|C_t|$ is non-decreasing with t, it suffices to say that $\ell_{t,k_t} \neq \ell'_{t,k_t}(k_{1:t-1})$ will happen at most $\log_2(1+\ell_\infty)$ times. This completes the proof.

D.2.4 Proof of Lemma 7

We first note that for every $\alpha \in \mathbb{R}$,

$$\arg\min_{\mathbf{p} \in \Delta_n} \left(\langle L, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p}) \right) = \arg\min_{\mathbf{p} \in \Delta_n} \left(\langle L + \alpha \mathbf{1}_n, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p}) \right)$$

Thus, without loss of generality, we can assume that $L = [L_1, \ldots, L_n]^\top$ satisfies

$$\sum_{k=1}^{n} \frac{1}{\eta L_k} = 1; \; L_k \geq 0, \; \forall k \in [n].$$

Notice that under such conditions, there is

$$\arg\min_{\mathbf{p} \in \Delta_n} \left(\langle L, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p}) \right) = \arg\min_{\mathbf{p} \in \mathbb{R}^n} \left(\langle L, \mathbf{p} \rangle + \frac{1}{\eta}\Psi(\mathbf{p}) \right)$$

by KKT conditions.

Now we start the proof. By the optimality of $\mathbf{x}$, there is

$$\eta L_k + \frac{1}{x_k} = 0, \; \forall k \in [n].$$

Then we have

$$\frac{l}{x_k} \geq -\frac{1}{2\eta x_k} = -\frac{L_k}{2},$$

thus

$$L_k + \frac{l}{x_k} \geq \frac{L_1}{2}.$$

By the optimality of $\mathbf{x}'$, there exists Lagrangian multiplier λ' such that

$$\eta L_k - \eta \frac{l}{x_k} + \lambda' - \frac{1}{x'_k} = 0,$$
$$\eta L_{k'} + \lambda' - \frac{1}{x'_{k'}} = 0,\ \forall k' \in [n] \backslash \{k\}.$$

and satisfies

$$\sum_{k' \in [n] \backslash \{k\}} \frac{1}{\eta L_{k'} + \lambda'} + \frac{1}{\eta L_k + \eta \frac{l}{x_k} + \lambda'} = 1.$$

Using the above, we note that

$$x'_k = \frac{1}{\eta L_k + \eta \frac{l}{x_k} + \lambda'} \leq \frac{1}{\eta \frac{L_k}{2} + \lambda'}.$$

It suffices to prove that $\lambda' \geq 0$. Define function

$$f(\lambda') = \sum_{k' \in [n] \backslash \{k\}} \frac{1}{\eta L_{k'} + \lambda'} + \frac{1}{\eta L_k + \eta \frac{l}{x_k} + \lambda'},$$

we note that

$$\sum_{k' \in [n] \backslash \{k\}} \frac{1}{\eta L_{k'}} + \frac{1}{\eta L_k + \eta \frac{l}{x_k}} \geq \sum_{k=1}^{n} \frac{1}{\eta L_k} = 1,$$

due to $l \leq 0$, which implies $f(0) \geq 1$, Since f decreases with λ', it suffices to conclude $\lambda' \geq 0$. Thus,

$$x'_k \leq \frac{1}{\eta \frac{L_k}{2} + \lambda'} \leq \frac{2}{\eta L_k} = 2x_k.$$

completes the proof.

D.2.5 Proof of Lemma 8

Similar to the proof of Lemma 7, it suffices to choose $L = [L_1, \ldots, L_n]^\top$ such that

$$\eta L_k - \frac{1}{x_k} = 0,\ \forall k \in [n].$$

By the optimality of $\mathbf{x}'$, there exists Lagrangian multiplier λ' such that

$$\eta' L_k + \lambda' - \frac{1}{x'_k} = 0,\ \forall k \in [n],$$
$$\sum_{k=1}^{n} \frac{1}{\eta' L_k + \lambda'} = 1.$$

Similar to the above, it suffices to show that $\lambda' \geq 0$ considering $\eta' \leq \eta$. Thus,

$$x'_k = \frac{1}{\eta' L_k + \lambda'} \leq \frac{1}{\eta' L_k} \leq \frac{C}{\eta L_k} = C x_k.$$

This completes the proof.

D.2.6 Proof of Lemma 9

We denote by

$$h_t = \min\Big(\max_{s\in[t-1]}(a_s), a_t\Big),\ b_t = a_t - h_t.$$

It suffices to say that

$$\sum_{t=1}^{T} b_t = \max_{t\in[T]}(a_t).$$

The proof can be completed as follows.

$$\begin{aligned}
\sum_{t=1}^{T} \frac{a_t}{\sqrt{2\sum_{s=1}^{t-1} a_s + 1}} &\le \sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=1}^{t-1} a_s + \max_{s\in[t-1]}(a_s) + 1}} \\
&= \sum_{t=1}^{T} \frac{h_t + b_t}{\sqrt{\sum_{s=1}^{t-1} a_s + \max_{s\in[t-1]}(a_s) + 1}} \\
&\le \sum_{t=1}^{T} \frac{h_t}{\sqrt{\sum_{s=1}^{t} h_s + 1}} + \sum_{t=1}^{T} b_t \\
&\le 2\sqrt{\sum_{t=1}^{T} h_t + 1} + \max_{t\in[T]}(a_t) \\
&\le 2\sqrt{\sum_{t=1}^{T} a_t + 1} + \max_{t\in[T]}(a_t)
\end{aligned}$$

D.2.7 Proof of Lemma 10

$$\begin{aligned}
|\langle \hat{\ell}'_t, \mathbf{c}_t\rangle| &\le \sum_{k=1}^{n} \frac{|\ell_{t,k}|\mathbb{1}(k = k_t)}{p_{t,k} + \rho_t c_{t,k}} |c_{t,k}| \\
&\le \ell_\infty \frac{\mathbb{1}(k_t^\star = k_t)}{p_{t,k_t^\star} + \rho_t c_{t,k_t^\star}} |c_{t,k_t^\star}| + \ell_\infty \sum_{k\in[n]\backslash\{k_t^\star\}} \frac{\mathbb{1}(k = k_t)}{p_{t,k} + \rho_t c_{t,k}} |c_{t,k}| \\
&\le \ell_\infty \frac{\mathbb{1}(k_t^\star = k_t)}{1/n - 1/2n} n + \ell_\infty \sum_{k\in[n]\backslash\{k_t^\star\}} \frac{\mathbb{1}(k = k_t)}{\rho_t} \\
&\le \ell_\infty \max(2n^2, 1/\rho_t) \\
&\le \ell_\infty \max(2n^2, 1/\rho_{T+1}) = \ell_\infty\Big(2n^2 + \sqrt{2\sum_{t=1}^{T} |\langle \hat{\ell}'_t, \mathbf{c}_t\rangle|}\Big),
\end{aligned}$$

where the first inequality is by the definition of $\hat{\ell}'_t$ and $\mathbf{p}'_t$, the second inequality is by the definition of ℓ_∞. The third inequality is due to 1). $\mathbf{p}_{t,k_t^\star}$ is one of the largest entries in $\mathbf{p}_t$, which implies $\mathbf{p}_{t,k_t^\star} \ge 1/n$. 2). $c_{t,k_t^\star} \ge -n$ and $\rho_t \le 1/2n^2$ for all $t \in [T]$. 3). $p_{t,k} + \rho_t c_{t,k} \ge \rho_t$ for all $k \in [n]\backslash\{k_t^\star\}$ by Algorithm 3. The last inequality is because ρ_t is nonincreasing.