Improved Algorithms for Adversarial Bandits with Unbounded Losses

Mingyu Chen
Department of Electrical & Computer Engineering
Boston University
Boston, MA 02215
`mingyuc@bu`

Xuezhou Zhang
Faculty of Computing & Data Sciences
Boston University
Boston, MA 02215
`mingyuc@bu`

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 losses, `UMAB-NN` achieves the first fully adaptive and scale-free regret bound. Building up on that, we further develop `UMAB-G` that can learn from arbitrary unbounded losses. Complementing the algorithms is a new lower bound, demonstrating a fundamental trade-off between adaptivility and minimax optimality in scale-free MAB. Finally, we perform extensive empirical evaluations, showing that our algorithms consistently out-perform all existing algorithms that handle unbounded losses.

1 Introduction

Multi-armed Bandits (MAB) is a popular online learning framework for studying decision-making under uncertainty [Slivkins et al., 2019, Lattimore and 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 adversarial MAB (AMAB), where the losses are generated adversarially by the environment [Auer et al., 2002]. Most prior works on AMAB assume that the losses are uniformly bounded, e.g. in $[0, 1]$. With such knowledge, the algorithms can set their *learning rate* (in a general sense) properly. For instance, in the `EXP3` algorithm, the regret analysis 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 applications, however, such natural loss bounds do not exist. For example, in quantitative trading, the fluctuation of stock prices can differ wildly across time and across stocks. In online marketplaces, 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 regret 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 *adaptive* 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 and 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?

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

In the following, we present two algorithms, UMAB-NN and UMAB-G, for Non-Negative and General unbounded loss, respectively. Our contributions are 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. UMAB-NN achieves improved regret bounds upon the best previous results from Hadiji and Stoltz [2020], Putta and Agrawal [2022], Huang et al. [2023].
3. Complementing the new adaptive algorithms, We establish a novel lower bound for scale-free AMAB. We show that, surprisingly, the adaptive regret bound achieved in the unbounded **non-negative** loss setting, is unachievable in the unbounded **general** loss setting, revealing the fundamental asymmetry between positive and negative losses in AMAB.
4. 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 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 $\ell_t \in \mathbb{R}^n$ *simultaneously*. 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*:

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

2.1 Related Works

Scale-free algorithms: Scale-free algorithms are ones whose regret bound scales *linearly* with respect to ℓ_∞, while requiring no knowledge of ℓ_∞ *a priori*[1]. Scale-free regret bounds were first studied in the full information setting, such as experts problems [Freund and Schapire, 1997, De Rooij et al., 2014, Cesa-Bianchi et al., 2007] and online convex optimization [Mayo et al., 2022, Jacobsen and 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, Orabona and 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. On AMAB, Hadiji and Stoltz [2020] extends the method

[1]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.

Algorithm	Unbounded	Adaptive	Regret
[Hazan and Kale, 2011]	No	Yes	$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\|\ell_t\|_2^2}\right)$
Hadiji and Stoltz [2020]	Yes	No	$\widetilde{O}\left(\ell_\infty\sqrt{nT}\right)$
Putta and Agrawal [2022]	Yes	No	$\widetilde{O}\left(\ell_\infty\sqrt{nT}+\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_2^2}\right)$
Putta and Agrawal [2022]	Yes	Yes	$\widetilde{O}\left(\ell_\infty\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_1}+\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_2^2}\right)$
UMAB-G **Non-Adaptive**	Yes	No	$\widetilde{O}\left(\ell_\infty^-\sqrt{nT}+\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}\right)$
UMAB-G **Adaptive**	Yes	Yes	$\widetilde{O}\left(\ell_\infty\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}+\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2}\right)$

Table 1: Comparison between our results and previous works.

of Duchi et al. [2011] and provides a scale-free regret bound of $\widetilde{\mathcal{O}}(\ell_\infty\sqrt{nT})$[2], which is optimal (up to log terms) in the worst case. However, such worst-case regret bounds can be overly pessimistic: a single outlier loss $\ell_{outlier}$ can result in an additional regret on the order of $\mathcal{O}(\|\ell_{outlier}\|_\infty\sqrt{nT})$. To address it, Putta and Agrawal [2022] presents scale-free bounds that adapt to the individual size of losses across time. Unfortunately, the worst-case guarantee of Putta and Agrawal [2022] is $\widetilde{\mathcal{O}}(\ell_\infty n\sqrt{T})$, which scales linearly to the number of actions. Our paper closes this gap: our algorithms achieve an adaptive regret better than Putta and Agrawal [2022], as well as an optimal worst-case regret that matches with Hadiji and Stoltz [2020].

Adaptive algorithms: Adaptive algorithms refer to algorithms which, instead of scaling solely on T in the regret, adapts to a "measure of hardness" of the sequence of losses. Such algorithm perform better than the worst-case regret if the sequence of loss is "benign". In the last two decades, such algorithms have been widely studied in the settings of expert problems and online convex optimization [Hazan et al., 2007, Streeter and McMahan, 2010, Duchi et al., 2011, De Rooij et al., 2014, Orabona and Pál, 2015, 2018]. For MAB, several works established adaptive regret bounds based on different "measure of hardness". For example, Allenberg et al. [2006], Foster et al. [2016], Pogodin and 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 and Kale [2011], Bubeck et al. [2018], Wei and 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 and Luo, 2018, Bubeck et al., 2019, Zimmert and 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* above results require the assumption that losses are bounded within $[0,1]$, which we remove in this paper.

3 Algorithm and Analysis

We now present our two algorithms `UMAB-NN` and `UMAB-G`. `UMAB-NN` works in the setting 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 main algorithm, `UMAB-G`, builds upon `UMAB-NN` but can work with potentially negative losses, i.e., $\ell_t\in\mathbb{R}^n$. In particular, We will present two versions of the algorithm: `UMAB-G` with non-adaptive and adaptive exploration rates. `UMAB-G` **Non-Adaptive** achieves the optimal worst-case guarantee $\widetilde{O}(\ell_\infty\sqrt{nT})$, while achieving partial adaptability to the non-negative part of the losses. `UMAB-G` **Adaptive** achieves an adaptive regret of $\widetilde{O}\left(\ell_\infty\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}\right)$, improving upon the best known adaptive regret of Putta and

[2]Throughout the paper we use $\mathcal{O}(\cdot)$ to suppress constant factors and $\widetilde{\mathcal{O}}(\cdot)$ to suppress factors that are poly-logarithmic in T, n, ℓ_∞.

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

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

Agrawal [2022] by an order of $\sqrt{n}$[3]. A summary of the comparisons to prior works can be found in Table 2.1[4].

Both algorithms we propose are based on the Follow-the-Regularized-Leader (FTRL) framework. Thus, let us prepare our discussion by briefly introducing FTRL. In the full-information setting, the classic 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 and Pál [2015]). The update rule takes the form of

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

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. The classic approach is to 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 $\hat{\ell}_{t,k} = \ell_{t,k}\mathbb{1}(k = k_t)/p_{t,k}$ for all $k \in [n]$, where $\mathbb{1}(k = k_t)$ denotes the indicator function. Notice that $\mathbb{E}[\hat{\ell}_t] = \sum_{k=1}^{n} \ell_t p_{t,k} \mathbf{e}_k / p_{t,k} = \ell_t$. Using $\hat{\ell}_t$, we are able to conceptually reduce the bandit setting to the full information setting.

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

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

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 and Luo [2018], Putta and 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 the loss ℓ_{t,k_t}, `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 and 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 and Pál [2018] instead, i.e. $\eta_{t+1} = \mathcal{O}(\sqrt{n/\sum_{s=1}^{t}\|\hat{\ell}_s\|_2^2})$, 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., $\eta_{t+1} = \mathcal{O}(\sqrt{n/\sum_{s=1}^{t}\ell_{s,k_s}^2})$, which can be bounded by $\mathcal{O}(\ell_\infty\sqrt{n/T})$.

[3] Notice that the regret in Putta and Agrawal [2022] scales with $||\ell_t||_2^2$ which can be n times larger than $||\ell_t||_\infty^2$.

[4] We remark that a recent work Huang et al. [2023] also claims to achieve adaptive regret for unbounded loss, but their proof is wrong, which is confirmed by the authors. More details are provided in Appendix B.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 3.1. *For any $\ell_1, \ldots, \ell_T \in \mathbb{R}^n_+$, the expected regret of Algorithm 1 is upper bounded by*

$$\mathcal{R}_T \le \tilde{\mathcal{O}}\left(\sqrt{n \sum_{t=1}^{T} \|\ell_t\|_\infty^2}\right).$$

Notice that Theorem 3.1 is adaptive to *each individual loss* in the loss sequence. 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]. In this regard, `UMAB-NN` achieves both optimal adaptive rate and optimal minimax rate for unbounded non-negative losses. Next, we briefly highlight the key steps in proving Theorem 3.1, which also provide intuition for our further improvement in the next section.

Proof highlights of Theorem 3.1: 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

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

It suffices to focus on bounding $\sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle$. We start with the standard analysis of an FTRL-type algorithm.

Lemma 3.2. *(Orabona [2019] Lemma 7.1) For any $\hat{\ell}_1, \ldots, \hat{\ell}_T \in \mathbb{R}^n$, using the update rule of (2) along with the non-increasing sequence of learning rates $\eta_1, \ldots, \eta_{T+1}$, for every comparator $\mathbf{p}^\dagger \in \Delta_n$, we have*

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

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 3.2 is provided in the appendix. Lemma 3.2 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, 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 3.3. *For any $\hat{\ell}_1, \ldots, \hat{\ell}_T \in \mathbb{R}^n$, using the update rule of (2), denote by $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A} \mathbf{x}}$, there is*

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

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

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 $\min(\frac{1}{2} \eta_t \ell^2_{t,k_t}, |\ell_{t,k_t}|)$, since

$$\langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}_{t+1} \rangle + F_t(\mathbf{p}_t) - F_t(\mathbf{p}_{t+1}) \le \langle \hat{\ell}_t, \mathbf{p}_t \rangle = |\ell_{t,k_t}|, \tag{4}$$

which implies

$$\sum_{t=1}^{T} \langle \hat{\ell}_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \le \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \sum_{t=1}^{T} \min\Big(\frac{1}{2} \eta_t \ell^2_{t,k_t}, |\ell_{t,k_t}|\Big). \tag{5}$$

The right hand side of (5) takes a similar form as in scale-free online convex optimization [Orabona and 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 (5) can be bounded by $\mathcal{O}(\sqrt{n \sum_{t=1}^{T} \ell^2_{t,k_t}})$ based on Orabona and Pál [2018], which suffices to complete the proof.

Algorithm 2 UMAB-G: Unbounded AMAB for General Loss

Input: Log-barriers regularization Ψ, learning rate $\eta_1 = 1/4$, exploration rate $\rho_1 = 1/2n^2$, clipping threshold $C_1 = -1$
for $t = 1$ **to** 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 rate ρ_t.
  Sample and play action $k_t \sim \mathbf{p}'_t$. Receive loss ℓ_{t,k_t}.
  Construct $\hat{\ell}'_t$ such that $\hat{\ell}'_{t,k} = \frac{\mathbb{1}(k=k_t)\ell'_{t,k}}{p'_{t,k}}$ for all $\forall k \in [n]$, 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): $\rho_{t+1} = \frac{1}{2n^2 + \sqrt{nT}}$.
    2. (Adaptive): $\rho_{t+1} = \frac{1}{2n^2 + 2\sqrt{\sum_{s=1}^t |\langle \hat{\ell}_s, \mathbf{c}_t \rangle|}}$.
end for

Algorithm 3 Extra Exploration on Action Distribution

Input: Distribution $\mathbf{p}_t$. Exploration rate $\rho_t \le 1/2n^2$
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 extra exploration distribution $\mathbf{p}'_t = \mathbf{p}_t + \rho_t \mathbf{c}_t$.

3.2 General loss

Next, we move on to study the general loss setting, i.e., $\ell_1, \ldots, \ell_T \in \mathbb{R}^n$. Let us first explain why Algorithm 1 cannot work when the losses become negative. Recall that Lemma 3.3 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 $\eta_t \|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}}/2$. However, notice that

$$\|\hat{\ell}_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \sum_{k=1}^n \frac{\hat{\ell}^2_{t,k}}{\nabla^2_{k,k}\Psi(\xi_t)} = \sum_{k=1}^n \frac{\ell^2_{t,k}\mathbb{1}(k=k_t)}{p^2_{t,k}}\xi^2_{t,k} = \frac{\ell^2_{t,k_t}}{p^2_{t,k_t}}\xi^2_{t,k_t}, \tag{6}$$

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 $\mathcal{O}(1/p^2_{t,k_t})$, 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, which implies that (4) will not go through. Thus, inequality (5) 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 $p_{t+1,k_t} \ge 1/2$ 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 (Algorithm 2). There are two key algorithmic innovations in UMAB-G. First, we use truncated losses 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. Second, we add extra exploration to ensure that the probability $p_{t,k}$ would not be overly small. Intuitively, for AMAB with general losses, 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 and Stoltz [2020], Putta and Agrawal [2022], we develop a data-dependent mixing strategy (Algorithm 3) that reduces the regret caused by the extra exploration. Below, we present the results for two exploration schemes distinguished by whether the exploration rate is adaptive. The theoretical guarantee of Algorithm 2 is presented below.

Theorem 3.4. *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*

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

$$\textit{Adaptive:} \quad \mathcal{R}_T \leq \tilde{\mathcal{O}}\Big(\ell_\infty n^2 + \sqrt{n \sum\nolimits_t \|\ell_t\|_\infty^2} + \ell_\infty \sqrt{n \sum\nolimits_t \|\ell_t\|_\infty} + \sqrt{n \sum\nolimits_t \|\ell_t\|_\infty}\Big) \tag{8}$$

Notice that the non-adaptive regret in (7) achieves "semi-adaptivity" to the loss sequence. If the loss sequence is non-negative, the right hand side of (7) reduces the form in Theorem 3.1. Moreover, the worst case bound of (7) 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 of Putta and Agrawal [2022] by a factor of $\sqrt{n}$ and achieves the optimal dependency on n and T.

Theorem 3.4 indicates that the two exploration schemes each have their own strengths and weaknesses under different loss sequence. Specifically, adaptive exploration rate is better when $\sum_{t=1}^T \|\ell_t\|_\infty \leq T$, whereas non-adaptive exploration rate is better when $\sum_{t=1}^T \|\ell_t\|_\infty > T$. This is later verified by our experiments.

Proof highlights of Theorem 3.4 Recall Algorithm 2 that $\hat{\ell}_t$ is the unbiased estimator and $\hat{\ell}'_t$ is the clipped biased estimator. By Algorithm 2 and the proof of Theorem 3.1, 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 $\hat{\ell}'_1, \ldots, \hat{\ell}'_T$. 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 $\hat{\ell}'_t$ instead of ℓ_t. In the rest of the proof, we bound these three terms respectively.

Bounding 1: By Lemma 3.2 and Lemma 3.3, we have

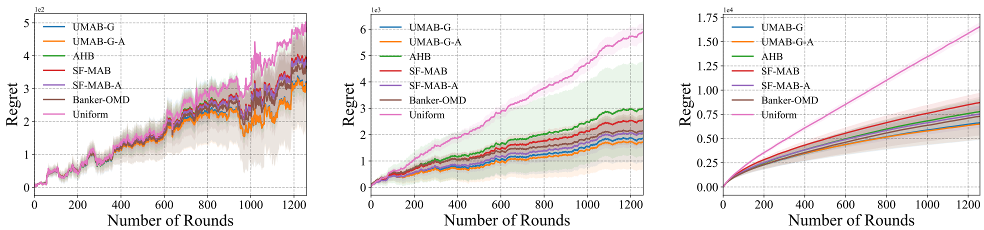
$$\sum_{t=1}^T \langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle \leq \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^T \eta_t \|\hat{\ell}'_t\|^2_{(\nabla^2\Psi(\xi_t))^{-1}} = \frac{\Psi(\mathbf{p}^\dagger)}{\eta_{T+1}} + \frac{1}{2}\sum_{t=1}^T \eta_t {\ell'}^2_{t,k_t}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

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, which is illustrated in the following lemma.

Lemma 3.5. *Given any action sequence $k_1, \ldots, k_T$, if $\ell_{t,k_t} \leq 0$. there is $p_{t,k_t} \leq 2p'_{t,k_t}$ and $p_{t+1,k_t} \leq 6p_{t,k_t}$ for every $t \in [T]$.*

Lemma 3.5 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 (5). 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 and Agrawal [2022]. This reduces the variance of our exploration rate, leading to an improved regret. The details are provided in Lemma 3.6 as follows.



(a) Stock Market Data (b) Amazon Sales Data (c) Model Selection

Figure 1: Real Data Experiments.

Lemma 3.6. *With the non-adaptive and adaptive exploration rates as in Algorithm 3, we have*

$$\textit{Non-Adaptive:} \qquad \mathbb{E}\Big[\,2\,\Big] \le 2\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2},$$

$$\textit{Adaptive:} \qquad \mathbb{E}\Big[\,2\,\Big] \le 2n^2\ell_\infty + 2\sqrt{1+4n\sum_{t=1}^{T}\|\ell_t\|_\infty} + 2\ell_\infty\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}$$

Bounding 3: Notice that

$$\sum_{t=1}^{T}\langle \hat{\ell}_t - \hat{\ell}'_t, \mathbf{p}'_t - \mathbf{p}^\dagger\rangle \le \sum_{t=1}^{T}\|\hat{\ell}_t - \hat{\ell}'_t\|_1\|\mathbf{p}'_t - \mathbf{p}^\dagger\|_\infty \le \sum_{t=1}^{T}\|\hat{\ell}_t - \hat{\ell}'_t\|_1.$$

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 3.7.

Lemma 3.7. *Given any action sequence $k_1, \dots, k_T$, with the non-adaptive and adaptive exploration rates as in Algorithm 3, we have*

$$\textit{Non-Adaptive:} \qquad \mathbb{E}\Big[\,3\,\Big] \le \ell_\infty^-\left(2n^2 + \sqrt{nT}\right)\log_2(1+\ell_\infty),$$

$$\textit{Adaptive:} \qquad \mathbb{E}\Big[\,3\,\Big] \le \ell_\infty^-\left(2n^2 + 3\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty}\right)\log_2(1+\ell_\infty).$$

Summing the bounds for 1,2,3 gives Theorem 3.4.

4 Lower bound for Scale-free Adversarial Multi-armed Bandit

Recall that `UMAB-NN` simutaneously achieves the optimal minimax regret rate $\widetilde{O}(\ell_\infty\sqrt{nT})$ and fully adaptive regret rate $\tilde{\mathcal{O}}(\sqrt{n\sum_{t=1}^{T}\|\ell_t\|_\infty^2})$, while the adaptive regret rate of `UMAB-G` is worse by a factor of $\sqrt{\ell_\infty}$ in comparison. Then, a natural question is, whether there exists an algorithm that can achieve full adaptability and minimax optimality in the unbounded general loss setting.

Unfortunately and perhaps surprisingly, we give a negative answer to this question. In the following, we present a new lower bound for scale-free AMAB, which shows that the optimal worst case regret and fully adaptive regret are mutually exclusive, and the regret bounds achieved by `UMAB-G` are nearly tight.

Theorem 4.1. *Let $\mathcal{A}$ be any scale-free AMAB algorithm such that the regret of $\mathcal{A}$ is minimax optimal, i.e., for any loss sequence $\ell_1, \dots, \ell_T \in \mathbb{R}^n$, the regret of $\mathcal{A}$ is bounded by $\mathcal{R}^{\mathcal{A}}(\ell_{1:T}) \le \tilde{\mathcal{O}}(\ell_\infty\sqrt{nT})$. Then, there exists a loss sequence $\ell'_1, \dots, \ell'_T$ that satisfies $\sqrt{\sum_{t=1}^{T}\|\ell'_t\|_\infty^2} \ll \ell'_\infty\sqrt{T}$, but the regret of $\mathcal{A}$ satisfies $\mathcal{R}^{\mathcal{A}}(\ell'_{1:T}) \ge \Omega(\ell'_\infty\sqrt{nT})$.*

To the best of our knowledge, this is the first lower bound that demonstrate the trade-off between minimax regret and adaptive regret in MAB. Our results also demonstrate a fundamental gap between optimizing positive and negative losses in MAB, and can be of independent interest.

5 Experiments

We now complement our theoretical results and evaluate the performance of our algorithms `UMAB-G` (non-adaptive) and `UMAB-G-A` (adaptive). We compare to **all** existing scale-free/unbounded AMAB algorithms, including `SF-MAB` Putta and Agrawal [2022], `SF-MAB-A` Putta and Agrawal [2022], `AHB` Hadiji and Stoltz [2020], and `banker-OMD` Huang et al. [2023]. The figures show the average regret and `std` across 500 trails. Additional ablation studies comparing `UMAB-G` with `UMAB-G-A` are deferred to Appendix A.

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 in later rounds. This is because 1). Compared to Putta and 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 and Stoltz [2020], our exploration design leads to smaller regret than mixing with uniform distribution.

Amazon Sales: Next, we design an experiment using Amazon sales data. Similar to the above, we consider $n = 10$ Amazon stores and $T = 1258$ rounds. We assume that 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 losses we considered in this setting are all 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.

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 size 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.

6 Conclusion

We propose algorithms that achieve the optimal adaptive and minimax regrets in AMAB with unbounded losses, and prove that minimax optimality and full adaptivity are mutually exclusive. Real data experiments validate 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*, pages 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*, pages 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*, pages 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*, pages 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*, pages 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*, pages 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*, pages 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*, pages 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*, pages 894–904. PMLR, 2020.

Sudeep Raja Putta and Shipra Agrawal. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 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.

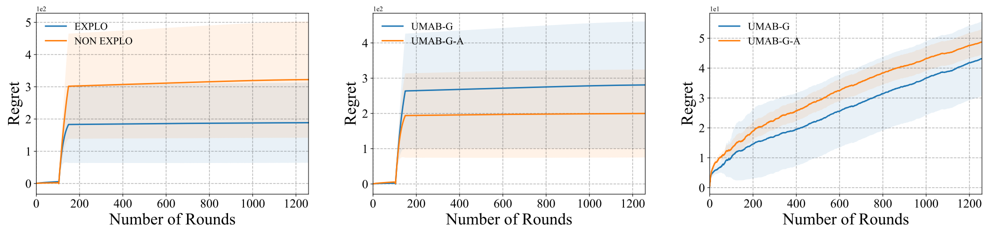
Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

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*, pages 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 Ablation Studies

In this section, we perform experimental evaluations on the design choices that we've made for our algorithms that improve their theoretical performances, and test whether these design choices also help in practice.

Impact of extra exploration 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:

$$\ell_t = \begin{cases} [0, -0.5]^\top, & \text{if } 1 \le t < 100 \\ [-10, 0]^\top, & \text{if } 100 \le t < 150 \\ [-0.05, 0]^\top, & \text{if } 150 \le t < 1258 \end{cases} \tag{9}$$

The intuition is to try deceive the 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. In particular, 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.

Comparison between `UMAB-G` and `UMAB-G-A` In the last part we investigate the difference between our algorithms with non-adaptive and adaptive exploration rates. Intuitively, the adaptive exploration rate is larger than the non-adaptive rate because it is of order $\mathcal{O}(1/\sqrt{t})$ instead of $\mathcal{O}(1/\sqrt{T})$ (assuming $\ell_\infty \ll T$). This makes adaptive exploration perform better in the adversary settings, e.g. as shown in Figure 2(b), where we use the same loss sequence in (9). However, if the loss sequence is benign, e.g. there exists one arm that is consistently better than the others, non-adaptive exploration will be better since it sacrifices 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 and Agrawal [2022]

In this subsection, we provide a detailed comparison between our work and Putta and 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 and Agrawal [2022] is to bound (3) by $\mathcal{O}(\ell_{t,k_t}^2/p_{t,k})$, resulting in an expectation regret $\mathcal{O}(\|\ell_t\|_2)$. In our work, we refine the analysis of (3), improving the bound to $\mathcal{O}(\ell_{t,k_t}^2)$, where the expectation is bounded by $\mathcal{O}(\|\ell_t\|_\infty^2)$. Considering the worst case scenario where $\|\ell_t\|_2^2 = n\|\ell_t\|_\infty^2$, our algorithm saves $\sqrt{n}$ in the regret.

Furthermore, Putta and Agrawal [2022] choose a uniform distribution for extra exploration. This approach ensures an exploration error (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 and Agrawal [2022] can be summarized as (under our notations definition)

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

which is suboptimal in n. In this study, we design to a new exploration strategy, as described in Algorithm 3. By Lemma D.4, 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 and 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., 3 in this paper) by

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

However, notice that the above only holds true if $\hat{\ell}_t \ge 0$. 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 $\mathbb{1}_{\neg clip(t)}$ denotes the probability of the clipping happening at round t. 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 $\hat{\ell}'_t$ to estimate $\mathbb{1}_{\neg clip(t)}\ell_t$, i.e., $\hat{\ell}'_{t,k} = \mathbb{1}_{\neg clip(t,k)}\ell_{t,k}/p_{t,k}$. However, we note that

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

which implies that $\hat{\ell}'_t$ is not an unbiased estimator of $\mathbb{1}_{\neg clip(t)}\ell_t$, 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. In particular, 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 3.1

C.1 Main proof and statement of technical lemmas

Recall (1), the expected regret can be denoted by

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

where $\mathbf{p}^\star$ denote the best fixed strategy. In particular, 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 (5), 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 smaller than $1/nT$ by definition.

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

Lemma C.1. *(Orabona and Pál [2018] Lemma 3) Let $a_1,\ldots,a_T \ge 0$. Then*

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

Using Lemma C.1 and $\eta_0,\ldots,\eta_T$ as in Algorithm 1, we have

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

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

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

completes the proof.

C.2 Proof of technical Lemmas

C.2.1 Proof of Lemma 3.2

For notations simplicity, we denote by

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

We first note

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

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 3.3

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 C.1

The proof refers to Lemma 3 in Orabona and 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 C.1 follows by summing over $t = 1, \ldots, T$ and choosing $\alpha = \sqrt{2}$. The proof is based on case analysis.

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

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

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 3.4

D.1 Main proof and statement of technical lemmas

By Lemma 3.3, we have

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

if $\ell_{t,k_t} \ge 0$. Alternatively, when $\ell_{t,k_t} < 0$, by Lemma 3.3 and 3.5 and inequality (6), we have

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

Moreover, we further note by Lemma 3.5,

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

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 3.1, we have

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

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

$$\mathbb{E}\Big[\sum_{t=1}^{T} \langle \hat{\ell}'_t, \mathbf{p}_t - \mathbf{p}^\dagger \rangle\Big] \leq \tilde{\mathcal{O}}\Big(\sqrt{n^2\ell_\infty^2 + n\sum_{t=1}^{T} \|\ell_t\|_\infty^2}\Big). \tag{10}$$

Now we summarize all the results.

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

Based on Lemma 3.6 and 3.7 and inequality (10), we have

1. (Non-Adaptive):

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

2. (Adaptive):

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

D.2 Proof of technical Lemmas

D.2.1 Proof of Lemma 3.5

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 $k_t^\star\in\arg\max_{k'\in[n]}p_{t,k'}$ as in Algorithm 3, it suffices to focus on the case $k_t=k_t^\star$, otherwise $p_{t,k_t}\leq p'_{t,k_t}$. When $k_t=k_t^\star$, 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 D.1. *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 D.2. *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 D.1 and D.2 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 D.1, $\tilde{p}_{t,k_t}/p_{t,k_t} \le 2$ if $-1/2\eta_t \le \ell'_{t,k_t} \le 0$. Moreover, by Lemma D.2, $p_{t+1,k_t}/p_{t,k_t} \le 3$ 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

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

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 3.6

Non-Adaptive exploration:

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

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

Adaptive exploration: We first introduce two auxiliary lemmas.

Lemma D.3. *Let $a_1, \dots, a_T \geq 0$. Then*

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

Lemma D.4. *Given any action sequence $k_1, \dots, k_T$, with the adaptive exploration rate as in Algorithm 2, there is*

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

The detailed proof of Lemma D.3 and D.4 would be provided later. Now we can prove Lemma 3.6.

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

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 D.3 with $a_t = |\langle\hat{\ell}'_t, \mathbf{c}_t\rangle|$. The last inequality is by Lemma D.4. Taking expectation on the both sides, there is

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

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., $|\ell_t'| \le |\ell_t|$. The last inequality is due to $\langle|\ell_t|, |\mathbf{c}_t|\rangle \le \|\ell_t\|_\infty\|\mathbf{c}_t\|_1 \le 2n\|\ell_t\|_\infty$. The whole proof is completed.

D.2.3 Proof of Lemma 3.7

Recall

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

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

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

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

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

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 $C_{t+1} \ge \ell_{t,k_t}$ holds, which also means $|C_{t+1}| \le |\ell_{t,k_t}|$. Thus, we have

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

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 D.1

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

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

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

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

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 D.2

Similar to the proof of Lemma D.1, 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 D.3

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.

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

D.2.7 Proof of Lemma D.4

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

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 $\mathbf{p}_{t,k_t^\star} \geq 1/n$. 2). $c_{t,k_t^\star} \geq -n$ and $\rho_t \leq 1/2n^2$ for all $t \in [T]$. 3). $p_{t,k} + \rho_t c_{t,k} \geq \rho_t$ for all $k \in [n] \setminus \{k_t^\star\}$ by Algorithm 3. The last inequality is because ρ_t is nonincreasing.

E Proof of Theorem 4.1

It suffices to focus on the case $n = 2$. Consider an environment E_0 with loss sequence $\ell_1, \ldots, \ell_T$, in which $\ell_{t,1} = 0$ and $\ell_{t,2} = 1$ for all $t \in [T]$. By assumption, we should have $\mathcal{R}^{\mathcal{A}}(\ell_{1:T}) \leq \tilde{\mathcal{O}}(\sqrt{T}) = r$. In this case, we can find an interval $\mathcal{I} \in [T]$ with $|\mathcal{I}| = T/r$ such that $\mathbb{E}[\sum_{t \in \mathcal{I}} \mathbb{1}\{k_t = 2\}] \leq 1/2$. Indeed, otherwise, let $[T] = \mathcal{I}_1 \cup, \ldots, \cup \mathcal{I}_r$ be a partition of interval $[T]$ into disjoint and consecutive intervals of size T/r, we will have

$$
\mathcal{R}^{\mathcal{A}}(\ell_{1:T}) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{k_t = 2\}\right] = \sum_{j=1}^{2\sqrt{2T}} \mathbb{E}\left[\sum_{t \in \mathcal{I}_j} \mathbb{1}\{k_t = 2\}\right] > \sqrt{2T}.
$$

which contradicts to the assumption.

Given the above, with probability at least $1/2$, algorithm $\mathcal{A}$ will not pull the arm 2 within the interval $\mathcal{I}$ when it runs against E_0. We define such an event by e. Now, consider a new environment E_1 with loss sequence $\ell'_1, \ldots, \ell'_T$, in which $\ell'_{t,1} = \ell_{t,1}$ for $t \in [T]$, $\ell'_{t,2} = \ell_{t,2}$ for $t \in [T] \setminus \mathcal{I}$, and $\ell'_{t,2} = -L$ for $t \in \mathcal{I}$. By the argument above, with probability at least $1/2$, algorithm $\mathcal{A}$ operates the same on E_0 and E_1. In this case, we note that

$$
\mathbb{E}\left[\mathcal{R}^{\mathcal{A}}(\ell'_{1:T}) | e\right] \geq L|\mathcal{I}| - T
$$

and

$$
\mathbb{E}\left[\mathcal{R}^{\mathcal{A}}(\ell'_{1:T}) | \neg e\right] \geq -T.
$$

Thus we finally have

$$
\mathbb{E}\left[\mathcal{R}^{\mathcal{A}}(\ell'_{1:T})\right] \geq \frac{L|\mathcal{I}|}{2} - T = \frac{LT}{2r} - T
$$

Take $L = \Omega(\sqrt{T})$, we have $\mathbb{E}\left[\mathcal{R}^{\mathcal{A}}(\ell'_{1:T})\right] \geq \Omega(L\sqrt{T})$. We further notice that

$$
\sqrt{\sum_{t=1}^{T} \|\ell'_t\|_\infty^2} \leq \sqrt{L^2\sqrt{T} + T} \ll L\sqrt{T}.
$$

This completes the proof.