ADVERSARIAL ONLINE LEARNING WITH RANKING FEEDBACK AND AN APPLICATION TO EQUILIBRIUM COMPUTATION

Mingyang Liu[1], Yongshan Chen[2], Zhiyuan Fan[1], Gabriele Farina[1], Asuman E. Ozdaglar[1], Kaiqing Zhang[2]
[1] EECS, Massachusetts Institute of Technology
[2] ECE, University of Maryland, College Park

ABSTRACT

Online learning in arbitrary and possibly adversarial environments has been extensively studied in sequential decision-making, with a strong connection to equilibrium computation in game theory. Most existing online learning algorithms are based on *numeric* utility feedback from the environment, which may be unavailable in applications with humans in the loop and/or the presence of privacy concerns. In this paper, we study an online learning model where only a *ranking* of a set of proposed actions is provided to the learning agent at each timestep. We consider both ranking models based on either the *instantaneous* utility at each timestep, or the *time-average* utility until the current timestep, in both *full-information* and *bandit* feedback settings. Focusing on the standard (external-)regret metric, we show that sublinear regret cannot be achieved in general with the instantaneous utility ranking feedback. Moreover, we show that when the ranking model is relatively *deterministic* (*i.e.,* with a small temperature), sublinear regret cannot be achieved with the time-average utility ranking feedback, either. We then propose new algorithms to achieve sublinear regret, under the additional assumption that the utility vectors have a sublinear variation. Notably, we also show that when time-average utility ranking is used, such an additional assumption can be avoided in the full-information setting. As a consequence, we show that if all the players follow our algorithms, an approximate coarse correlated equilibrium of a normal-form game can be found through repeated play. Finally, we also validate the efficiency of our algorithms via numerical experiments.

1 INTRODUCTION

Online learning has been extensively studied as a model for sequential decision-making in arbitrary, and possibly adversarial environments (Shalev-Shwartz et al., 2012; Hazan et al., 2016). At each round of decision-making, the learning agent commits to a strategy and takes an action, then receives some *feedback* from the environment, oftentimes in a *numeric* form such as the utility vector (in the *full-information* setting) or the realized utility value (in the *bandit* setting). Numerous algorithms have been developed to achieve no-regret, *i.e.,* ensuring (external) regret grows sublinearly in time (Shalev-Shwartz et al., 2012; Hazan et al., 2016). Moreover, online learning is known to also have an inherent connection to equilibrium computation in Game Theory—when all the players are no-regret in repeatedly playing a normal-form game, the time-average strategy will approximate the coarse correlated equilibrium (CCE) of the game (Cesa-Bianchi & Lugosi, 2006).

However, such numeric feedback of utility values may not always be available in real-world applications. For example, when the feedback is provided by humans in the loop, it is much more convenient for them to *compare/rank* actions instead of numerically *scoring* them. This has been acknowledged and testified by the recent successes of reinforcement learning from human feedback (RLHF) in fine-tuning language models (Ouyang et al., 2022). Moreover, even if numeric utility values exist, sometimes they may not be accessible to the learning agent due to privacy or security concerns. For example, consider an online platform (c.f. Fig. 1 (a)) that recommends commodities to a stream of customers in an online fashion, where the customers at different timesteps may have

different preferences of the commodities. The platform aims to make good recommendations over time, while the customers may not be able/willing to reveal their actual valuation of the commodities. Depending on the types of the customers, *i.e.,* either being *one-shot* (arrive, rank, and leave forever), or being *long-lived* with memory, the utility used for ranking may either be the *instantaneous* one at each timestep, or the *time-average* one over the historical utility vectors so far. As a consequence, the platform needs to minimize the *regret* incurred by the recommendations with such ranking feedback, and it remains elusive what fundamental limits and effective algorithms are in such a setting.

The ranking feedback becomes more relevant in game-theoretic settings, when multiple humans interact continuously with each other, and the objective is to compute a certain equilibrium of the game. For example, consider a dating application platform recommending candidates for matching (c.f. Fig. 1 (b)). Each customer may only have a ranking of the recommended candidates at each round, and the platform aims to find an equilibrium (a matching between customers) so that all the customers are satisfied (Etesami & Srikant, 2024). Similar scenarios also appear in other matching platforms, *e.g.,* ride-sharing platforms that match drivers and passengers based on their preferences, such as the drivers' preference for trip lengths and the users' preferences for the drivers' driving manners (being prompt or cautious).

In this paper, we seek to systematically study online learning with ranking feedback, where the loss vectors may be arbitrarily generated. This setting can be viewed as a generalization of the stochastic bandit with ranking feedback studied recently in Maran et al. (2024). We aim to understand when regret minimization in our online setting is possible, and also develop new algorithms with regret and equilibrium approximation guarantees. Furthermore, we also study equilibrium computation with ranking feedback in repeated game playing, based on our online learning results. We summarize our contributions as follows.

Contributions. We consider two types of ranking feedback, categorized by how the rankings are made: one based on the *instantaneous utility* at each timestep (**InstUtil Rank**), and one based on the *time-average utility* until the current timestep (**AvgUtil Rank**). We will thoroughly study both the *full-information* and *bandit* feedback settings under both feedback types. In particular, we show the following.

i. It is impossible to achieve sublinear regret under **InstUtil Rank** feedback in both full-information and bandit settings.

ii. It is impossible to achieve sublinear regret under **AvgUtil Rank** feedback in both full-information and bandit settings when the ranking model is too *deterministic* (*i.e.,* the temperature $\tau > 0$ of the ranking model in (PL) is very small).

iii. We propose a new algorithm to achieve sublinear regret under **InstUtil Rank** feedback for both full-information and bandit settings, with an additional assumption on the sublinear variation of the utility vectors (c.f. Assumption 4.1).

iv. We propose a new algorithm to achieve sublinear regret under **AvgUtil Rank** feedback, without Assumption 4.1 for the full-information setting, and with Assumption 4.1 for the bandit setting.

v. When all the players follow our no-regret learning algorithms in repeatedly playing a normal-form game, an approximate CCE can be computed from the time-average strategies.

Our results are summarized in Table 1.

2 Online learning with ranking feedback

In this section, we will introduce the model of online learning with ranking feedback. The notations and preliminaries of online learning and games are postponed to Appendix A.

In online learning with ranking feedback, at timestep t, the agent does not have direct access to $\boldsymbol{u}^{(t)}$, nor the realized utility (the utility of the realized action at timestep t). Instead, at timestep t, she can propose a multiset (which may include repeated elements) of actions $o^{(t)} \subseteq \mathcal{A}$ and receives

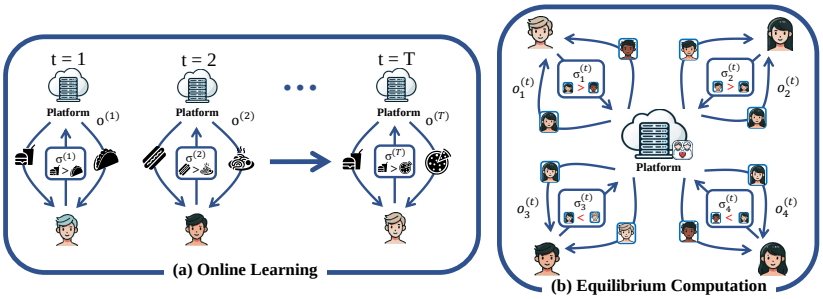

Figure 1: Two real-world examples of Online Learning and Equilibrium Computation with Ranking Feedback. Figure (a) is an example of a food recommendation application, where the online platform recommends choices of food to the customer at each timestep, and serves a stream of (possibly heterogeneous) customers in an online fashion. The customers will then rank the recommendations as feedback to the platform, so that the platform can improve the recommendation quality in the long run. Figure (b) is an example of a dating application, where the online platform recommends customers to match each other, so that they are all satisfied with the matching and have no incentives to deviate, *i.e.,* reaching an *equilibrium*. The customers will then rank the recommended candidates, and the platform aims to leverage such ranking feedback to find an equilibrium in multiple timesteps of such online interactions.

a permutation $\sigma^{(t)} \in \Sigma\left(o^{(t)}\right)$ from the environment, representing a *ranking* of those actions in $o^{(t)}$. In full-information setting, $o^{(t)} = \mathcal{A}$, *i.e.,* the whole action set is proposed. In bandit setting, $\left|o^{(t)}\right| = K$. Suppose the agent's strategy at timestep t is $\pi^{(t)} \in \Delta^{\mathcal{A}}$, then we assume that in the bandit setting, the actions in $o^{(t)}$ are proposed in an *unbiased* way, such that

$$\mathbb{E}\left[\frac{\sum_{a \in o^{(t)}} u^{(t)}(a)}{K}\right] = \left\langle \boldsymbol{u}^{(t)}, \pi^{(t)} \right\rangle, \tag{2.1}$$

which may be achieved if all the $a \in o^{(t)}$ are sampled i.i.d. from $\pi^{(t)}$ (with replacement). Let $\sigma^{(t)}(k) \in \mathcal{A}$ be the kth element of the permutation for any $k \in [K]$. Then, for any $k_1 < k_2 \in [K]$, action $\sigma^{(t)}(k_1)$ is preferred over action $\sigma^{(t)}(k_2)$. For notational simplicity, we define $a^i \overset{\sigma}{<} a^j$ if action a^i appears ahead of a^j in a permutation σ.

For the ranking model, we consider the standard Plackett-Luce (PL) model (Luce, 1959; Plackett, 1975), where at each timestep t, conditioned on the proposed action set $o^{(t)}$, the ranking $\sigma^{(t)}$ is sampled from

$$\mathbb{P}\left(\sigma^{(t)} \mid o^{(t)}\right) = \prod_{k_1=1}^{K} \frac{\exp\left(\frac{1}{\tau} r^{(t)}\left(\sigma^{(t)}\left(k_1\right)\right)\right)}{\sum_{k_2=k_1}^{K} \exp\left(\frac{1}{\tau} r^{(t)}\left(\sigma^{(t)}\left(k_2\right)\right)\right)}, \tag{PL}$$

where $\boldsymbol{r}^{(t)} \in \mathbb{R}^{\mathcal{A}}$ is some vector based on which the ranking is determined (as to be instantiated later), $\tau > 0$ is the *temperature* that determines how uncertain the ranking model is: when $\tau \to 0^+$, the model is absolutely certain, and the action with a larger utility in $\boldsymbol{r}^{(t)}$ will always be ranked in front of the actions with a smaller utility in the permutation. The utility vector $\boldsymbol{r}^{(t)}$ depends on the problem setting, which we will introduce next.

We will consider two types of ranking feedback throughout the paper, based on the choice of $\boldsymbol{r}^{(t)}$ in (PL): (i) ranking by the *instantaneous* utility (**InstUtil Rank**); (ii) ranking by the *time-average* utility (**AvgUtil Rank**). Both feedback forms can also be further separately defined for the *full-information* and *bandit* settings as below.

InstUtil Rank: Ranking with instantaneous utility.

The first type of ranking feedback we consider is based on the *instantaneous* utility function, *i.e.,* $\boldsymbol{r}^{(t)} = \boldsymbol{u}^{(t)}$ in (PL). Note that only the utilities at the *proposed* actions will be used

for ranking. This type is relevant when the feedback provider is oblivious or one-shot. For example, a stream of customers arrive in an online fashion, each of whom arrives, ranks, and then leaves, see *e.g.,* Mansour et al. (2015). When the environment is *stationary* and *stochastic*, the classical dueling-bandits model also used *instantaneous* utilities for comparison/ranking (Yue et al., 2012; Du et al., 2020).

Full-information setting. In this setting, *all* the actions can be evaluated and ranked at each timestep t, even for those she did not propose. Hence, her performance can be evaluated by $\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)} \right\rangle$. Note that this does not mean the agent can access the full vector $\boldsymbol{u}^{(t)}$, since this contradicts our ranking-feedback setting. Hence, the standard (external-)regret $R^{(T),\text{external}}$ defined in (A.1) will serve as the metric to evaluate the agent's performance in this online learning process.

Bandit setting. In this setting, only the *proposed actions* at each timestep t can be evaluated and ranked, with the associated elements in the vector $\boldsymbol{u}^{(t)}$. In particular, the proposed actions are evaluated by the average utility of $\frac{1}{K}\sum_{j=1}^{K} u^{(t)}\left(\sigma^{(t)}(j)\right)$, leading to the following regret metric for performance evaluation:

$$R^{(T)} := \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} \right\rangle - \frac{1}{K} \sum_{a\in o^{(t)}} u^{(t)}(a) \right). \tag{2.2}$$

Note that such a definition coincides with that in the (multi-)dueling-bandit problems (Yue et al., 2012; Du et al., 2020; Saha et al., 2021; Saha & Gaillard, 2022).

AvgUtil Rank: Ranking with time-average utility.

The second type of ranking-feedback is based on the *time-average* utility, which will differ for the full-information and bandit settings, as detailed below. This type is relevant when the feedback provider has memory and can use the history of utilities for ranking. For example, the customers are long-lived in the platform, see *e.g.,* Küçükgül et al. (2022) and Baldwin (2009). Notably, under bandit feedback, such a model aligns with the model in the recent work by Maran et al. (2024), when $\tau \to 0^+$ and the environment is stationary and stochastic.

Full-information setting. The time-average utility vector of $\boldsymbol{u}_{\text{avg}}^{(t)} := \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{u}^{(s)}$ will be used as the $\boldsymbol{r}^{(t)}$ in (PL) for ranking, even for those actions the agent did not propose at timestep t. Hence, $\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)} \right\rangle$ can be computed, and we will thus still use the (external-)regret $R^{(T),\text{external}}$ from (A.1) as the performance metric.

Bandit setting. In the bandit setting, only the proposed actions will be given to the environment to evaluate. For instance, the platform (learning agent) may recommend K restaurants among all possibilities to the user (environment) to try, so that the user will only know her evaluations of those K restaurants. As a result, the average utility is now defined as the *empirical mean* of the utility vectors over time. Formally, for each action $a \in \mathcal{A}$, we define

$$u_{\text{empirical}}^{(t)}(a) := \frac{\sum_{s=1}^{t} u^{(s)}(a) \sum_{a'\in o^{(t)}} \mathbb{1}\left(a = a'\right)}{\sum_{s=1}^{t} \sum_{a'\in o^{(t)}} \mathbb{1}\left(a = a'\right)}. \tag{2.3}$$

This $\boldsymbol{u}_{\text{empirical}}^{(t)}$ will then be used as the $\boldsymbol{r}^{(t)}$ in (PL) for ranking. Note that the discrepancy between the ranking models in the full-information and bandit settings in this case (contrast to that for **InstUtil Rank**), is due to that when looking at the history, the utility at actions *other than* those proposed at timestep t may be available and still be useful later. In contrast, for **InstUtil Rank**, only those proposed at timestep t are relevant for ranking, *i.e.,* only those elements of $\boldsymbol{u}^{(t)}(a)$ with $a \in o^{(t)}$ are used. Moreover, the ranking will be the *same* as returning the ranking of the proposed actions from the ranking of *all* actions (see the proof of Lemma E.1 for details). The regret metric will still be that in (2.2).

3 EQUILIBRIUM COMPUTATION WITH RANKING FEEDBACK

There is a mediator (platform) in the game that computes and recommends strategies to the players (*e.g.,* Uber recommends the matching between drivers and users), but with only access to the ranking feedback from the players, *e.g.,* humans. Specifically, when the strategy profile $\boldsymbol{\pi}$ is implemented by the players, player i's utility of taking action $a_i \in \mathcal{A}_i$ is $u_i^{\boldsymbol{\pi}}(a_i) := \sum_{\boldsymbol{a}' \in \times_{j=1}^N \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}') \mathbb{1}(a_i' = a_i) \prod_{j \neq i} \pi_j(a_j')$. However, instead of observing the utility directly, the mediator can only observe the ranking based on it.

Therefore, at each timestep t, the mediator will choose a strategy profile $\boldsymbol{\pi}$ and propose each player $i \in [N]$ a multiset $o_i^{(t)} = \left\{a_i^{(t),k}\right\}_{k=1}^K$ consisting of K actions.

- Full-information. Each player $i \in [N]$ will receive utility $\boldsymbol{u}_i^{\boldsymbol{\pi}^{(t)}}$, where $\boldsymbol{\pi}^{(t)} = \left(\pi_1^{(t)}, \ldots, \pi_N^{(t)}\right)$ is the strategy profile at timestep t.
- Bandit. The K actions are proposed sequentially. Therefore, when all players receive the kth action, player $i \in [N]$ receive the utility $\mathcal{U}_i\left(\left(a_j^{(t),k}\right)_{j=1}^N\right)$.

The player will evaluate her actions' utilities and report a ranking of the proposed actions according to either **InstUtil Rank** or **AvgUtil Rank**. The process will be repeated until the mediator finds an (approximate) equilibrium.

4 ONLINE LEARNING WITH **InstUtil Rank** FEEDBACK

The key challenge in achieving no-regret in the example is that the utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^T$ change arbitrarily fast over time. Hence, to obtain positive results, we may need to restrict how fast they change over time, as quantitatively characterized by the following assumption.

Assumption 4.1 (Sublinear variation of utility vectors)**.** *The utility vectors* $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^T$ *have a sublinear variation over time,* i.e.,

$$P^{(T)} := \sum_{t=2}^{T} \left\| \boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)} \right\| \leq \mathcal{O}(T^q), \tag{4.1}$$

for some $q < 1$.

Our result stated in Section 4 next will show that with Assumption 4.1, we can achieve sublinear regret, and thus close the gap. Moreover, in a game where the opponents all run no-regret learning algorithms such as projected gradient descent, Assumption 4.1 will be satisfied.

We will now present our algorithm for adversarial online learning under **InstUtil Rank**. The utility estimator in our algorithms is postponed to Appendix E.

4.1 FULL-INFORMATION SETTING

Under full-information feedback, the learning agent proposes the full action set $\mathcal{A}$ at each timestep. In this case, we can obtain $\widetilde{\boldsymbol{u}}^{(t)}$, an estimate of $\boldsymbol{u}^{(t)}$, by Algorithm 1, and obtain guarantees using Theorem E.2 under the setting of $p = 1$ (since all actions are proposed at each timestep). We now show that for an arbitrary adversarial online learning algorithm that can achieve sublinear external regret, we can construct an online learning algorithm with **InstUtil Rank** feedback based on it in a black-box way. A diagram of the algorithm is in Figure 2 and the details are specified in Algorithm 2. Specifically, we have the following theorem.

Theorem 4.2 (Informal)**.** *Consider* **InstUtil Rank** *with constant* $\tau > 0$*, full-information feedback, and Algorithm 2. For any* $\delta \in (0, 1)$*,* $T > 0$*, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, by choosing* m *properly, we have that with probability at least*

$(1-\delta)$, $R^{(T),\text{external}}$ *satisfies*

$$R^{(T),\text{external}} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) + \mathcal{O}\left(\left(P^{(T)}\right)^{\frac{1}{3}} T^{\frac{2}{3}} \left(\log\left(\frac{T}{\delta}\right)\right)^{\frac{1}{3}}\right).$$

Theorem 4.2 implies that when $P^{(T)}$, the variation of utility vectors, is sublinear, the regret of Algorithm 2 will be sublinear. The proof and formal version of the theorem are provided in Appendix H. The key technical step in the proof amounts to showing that $R^{(T),\text{external}}$ and $R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)$ are close to each other, since the associated utility vectors $\widetilde{\boldsymbol{u}}^{(t)}$ and $\boldsymbol{u}^{(t)}$ are close by Theorem E.2.

4.2 Bandit Setting

In the bandit-feedback setting, the utility of the online learning agent at timestep t is $\frac{1}{K}\sum_{k=1}^{K} u^{(t)}\left(\sigma^{(t)}(k)\right)$, *i.e.,* the average utility of the proposed actions. To achieve sublinear $R^{(T)}$, each proposed action will be sampled from $\pi^{(t)}$ independently *with replacement*. In other words, an action may be proposed multiple times at a single timestep. Therefore, to ensure each action will be proposed with a positive probability, we need to let $\pi^{(t)}(a) \geq \frac{\gamma}{|\mathcal{A}|}$ for some $\gamma > 0$ and every action $a \in \mathcal{A}$. To this end, we will let $\pi^{(t+1)} = (1-\gamma)\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right) + \gamma\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}$, *i.e.,* a convex combination of the strategy generated by the no-regret learning algorithm Alg and a uniform probability distribution over $\mathcal{A}$. The diagram can be found in Figure 2 and the details are in Algorithm 2.

Then, the regret bound for Algorithm 2 under bandit feedback is as follows.

Theorem 4.3 (Informal)**.** *Consider* **InstUtil Rank** *with constant $\tau > 0$, bandit feedback, and Algorithm 2. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, by choosing γ, m properly, with probability at least $(1-\delta)$, $R^{(T)}$ satisfies*

$$R^{(T)} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) + \mathcal{O}\left(\left(P^{(T)}\right)^{\frac{1}{5}} T^{\frac{4}{5}} \log\left(\frac{T}{\delta}\right)\right).$$

The proof and the formal version of Theorem 4.3 is provided in Appendix I. Similar to Theorem 4.2, when $P^{(T)}$ is sublinear in T, $R^{(T)}$ is also sublinear in T. The proof consists of three parts. To start, we use a concentration bound to argue that the difference $\left|R^{(T)} - R^{(T),\text{external}}\right|$ is small. Then, we focus on the quantity $\widetilde{R}^{(T)} := \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \widetilde{\boldsymbol{u}}^{(t)}, \widehat{\pi} - \pi^{(t)}\right\rangle$. This quantity is close to $R^{(T),\text{external}}$, since $\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|$ is bounded as shown in Theorem E.2. Finally, we show that $\left|\widetilde{R}^{(T)} - R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right|$ is bounded by $\mathcal{O}(\gamma T)$, by using the fact that $\pi^{(t+1)} = (1-\gamma)\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right) + \gamma\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}$.

5 Online Learning with **AvgUtil Rank** Feedback

We will now focus on the setting of online learning with **AvgUtil Rank**.

5.1 Utility Estimation

Since $\sigma^{(t)}$ is generated based on $\boldsymbol{u}_{\text{avg}}^{(t)}$, we will estimate $\boldsymbol{u}_{\text{avg}}^{(t)}$ instead. We will still apply Algorithm 1, which will generate $\widetilde{\boldsymbol{u}}_{\text{avg}}^{(t)}$, an estimate of $\boldsymbol{u}_{\text{avg}}^{(t)}$, when the permutation is sampled under **AvgUtil Rank**. Moreover, notice that

$$\left\|\boldsymbol{u}_{\text{avg}}^{(t)} - \boldsymbol{u}_{\text{avg}}^{(t-1)}\right\|_{\infty} = \left\|\frac{\boldsymbol{u}^{(t)} + (t-1)\boldsymbol{u}_{\text{avg}}^{(t-1)}}{t} - \boldsymbol{u}_{\text{avg}}^{(t-1)}\right\|_{\infty} \leq \frac{1}{t}\left(\left\|\boldsymbol{u}^{(t)}\right\|_{\infty} + \left\|\boldsymbol{u}_{\text{avg}}^{(t-1)}\right\|_{\infty}\right) \leq \frac{2}{t}.$$

Therefore, the estimation error is bounded as follows.

Theorem 5.1. *Consider* **AvgUtil Rank** *and Algorithm 1. Suppose each action is proposed with probability at least $p > 0$ at each timestep $t \in [T]$ and let $\widetilde{\boldsymbol{u}}_{\text{avg}}^{(t)} = \text{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=t-m'+1}^{t}\right)$. Then, for any $\delta \in (0,1)$ and $t \geq m'$, when $m'p^4 \geq 2\log\left(\frac{2}{\delta}\right)$, with probability at least $1-\delta$, the estimate $\widetilde{\boldsymbol{u}}_{\text{avg}}^{(t)}$ satisfies,*

$$\left\|\widetilde{\boldsymbol{u}}_{\text{avg}}^{(t)} - \boldsymbol{u}_{\text{avg}}^{(t)}\right\|_\infty \leq \frac{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}{p}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m'}} + \sum_{s=t-m'+1}^{t-1}\frac{2}{s+1}.$$

When taken δ, p as constants, the accumulated estimation error $\sum_{t=1}^{T}\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|_\infty$ is bounded by $\mathcal{O}\left(\frac{T}{\sqrt{m'}} + m'\sum_{t=1}^{T}\frac{1}{t}\right) \leq \mathcal{O}\left(\frac{T}{\sqrt{m'}} + m'\log T\right)$. Therefore, to achieve sublinear accumulated estimation error, Assumption 4.1 is no longer required. The proof of Theorem 5.1 is simply substituting $\boldsymbol{u}^{(t)}$ in Theorem E.2 to $\boldsymbol{u}_{\text{avg}}^{(t)}$.

5.2 Full-Information Setting

Unlike the full-information setting under **InstUtil Rank**, where any (full-information) adversarial online learning algorithm can be leveraged, the algorithm for **AvgUtil Rank** needs to satisfy the following assumption.

Assumption 5.2. *The (full-information) adversarial online learning algorithm* Alg *needs to satisfy the following condition: for any $T > 0$, $t \in [T]$, sequences of utility vectors $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}, \left(\boldsymbol{u}'^{(s)}\right)_{s=1}^{t} \in \left(\mathbb{R}^{\mathcal{A}}\right)^t$, we have*

$$\left\|\text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right) - \text{Alg}\left(\left(\boldsymbol{u}'^{(s)}\right)_{s=1}^{t}\right)\right\| \leq L\left\|\sum_{s=1}^{t}\boldsymbol{u}^{(s)} - \sum_{s=1}^{t}\boldsymbol{u}'^{(s)}\right\|,$$

where $L = \Theta\left(T^{-c}\right)$ for some constant $c \in (0,1)$.

It can be verified that follow-the-regularized-leader (FTRL) with any strongly convex regularizer satisfies this assumption. The proof is postponed to Lemma M.3. The proposed action set is just $\mathcal{A}$ as in Section 4.1. The overall procedure is summarized in Algorithm 3.

Algorithm 3 has the following regret-bound guarantee.

Theorem 5.3 (Informal)**.** *Consider* **AvgUtil Rank** *with constant $\tau > 0$, full-information feedback, and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, that satisfies Assumption 5.2, by choosing m properly, we have that with probability at least $(1-\delta)$, $R^{(T),\text{external}}$ satisfies*

$$R^{(T),\text{external}} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + \mathcal{O}\left(LT^{\frac{5}{3}}\log\left(\frac{2T}{\delta}\right)\right).$$

Theorem 5.3 shows that with a small enough $L = \Theta(T^{-c})$ satisfying $c > 2/3$, $R^{(T),\text{external}}$ can be made sublinear. The proof and the formal version of Theorem 5.3 is provided in Appendix J. The proof sketch of Theorem 5.3 is as follows. Let $\overline{\pi}^{(t+1)} = \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right)$. Then, $\left|R^{(T),\text{external}} - R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right)\right|$ is bounded by $\mathcal{O}\left(\sum_{t=1}^{T}\left\|\pi^{(t+1)} - \overline{\pi}^{(t+1)}\right\|\right)$. Then, by Assumption 5.2, $\left\|\pi^{(t+1)} - \overline{\pi}^{(t+1)}\right\| \leq Lt\left\|\boldsymbol{u}_{\text{avg}}^{(t)} - \widetilde{\boldsymbol{u}}_{\text{avg}}^{(t)}\right\|$, and the result follows.

5.3 Bandit Setting

While applying Algorithm 1 to estimate the utility, we can only obtain an estimation of $\boldsymbol{u}_{\text{empirical}}^{(t)}$ instead of $\boldsymbol{u}_{\text{avg}}^{(t)}$. Therefore, to estimate $\boldsymbol{u}_{\text{avg}}^{(t)}$, we will divide the timesteps $\{1, 2, \ldots, t\}$ into $\lceil t/M \rceil$ blocks, with each block containing M timesteps except for the last one. Let $n^{(t)}(a) :=$

$\sum_{s=1}^{t} \#_{o^{(s)}}(a)$ for any $a \in \mathcal{A}$ as the number of times action a is proposed up to timestep t. Then, for each block $[s \cdot M + 1, (s+1)M]$ and $a \in \mathcal{A}$, we estimate $\frac{1}{M}\sum_{s'=s\cdot M+1}^{(s+1)M} u^{(s')}(a)$ by computing

$$\frac{\widetilde{u}_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - \widetilde{u}_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}.$$

The full algorithm is illustrated in Algorithm 3.

Theorem 5.4. *Consider* **AvgUtil Rank** *with constant $\tau > 0$, bandit feedback, and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, that satisfies Assumption 5.2, by choosing M, m, γ properly, we have that with probability at least $(1-\delta)$, $R^{(T)}$ satisfies*

$$R^{(T)} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + \widetilde{\mathcal{O}}\left(\left(\log\left(\frac{1}{\delta}\right)\right)^{2} L^{\frac{1}{3}} T^{\frac{23}{18}} \left(P^{(T)}\right)^{\frac{1}{6}}\right),$$

where $\widetilde{\mathcal{O}}$ hides logarithm of T.

When $P^{(T)} \leq \mathcal{O}\left(T^{q}\right)$ for some $q < \frac{1}{3}$, and $L = \Theta\left(T^{-c}\right)$ with $c \in \left(\frac{5}{6} + \frac{q}{2}, 1\right)$, Theorem 5.4 guarantees sublinear $R^{(T)}$. We need $c < 1$ because typically $R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) \leq \mathcal{O}\left(\frac{1}{L} + LT\right)$, *e.g.,* FTRL with any strongly convex regularizer. Furthermore, when all players are running Algorithm 3 in games, Theorem 5.4 guarantees that the individual regret of each player is sublinear, when $P^{(T)} \leq \mathcal{O}(LT)$ (satisfied by FTRL with any strongly convex regularizer). Because $q \leq 1 - c$ and $\left(\frac{5}{6} + \frac{1-c}{2}, 1\right)$ is non-empty when $c > \frac{2}{3}$. The details can be found in Section 6.

The proof of Theorem 5.4 is similar to that of Theorem 4.3, except for the technique to bound $\left\|\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} - \boldsymbol{u}_{\text{avg}}^{(t)}\right\|$. To bound the estimation error, we need to show that for each block $[s \cdot M + 1, (s+1)M]$ and $a \in \mathcal{A}$, the estimate

$$\frac{\widetilde{u}_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - \widetilde{u}_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}$$

is close to $\frac{1}{M}\sum_{s'=s\cdot M+1}^{(s+1)M} u^{(s')}(a)$. By using Theorem E.2, we can bound the difference between the estimate and

$$\frac{u_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - u_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)},$$

whose difference with $\frac{1}{M}\sum_{s'=s\cdot M+1}^{(s+1)M} u^{(s')}(a)$ can be bounded by the variation of the utility vectors in the block. The full proof and the formal version can be found in Appendix K.

6 Equilibrium Computation with Ranking Feedback

For a normal-form game $\left(N, \{\mathcal{A}_i\}_{i=1}^{N}, \{\mathcal{U}_i\}_{i=1}^{N}\right)$, the external regret of player $i \in [N]$ is

$$R_i^{(T),\text{external}} := \max_{\widehat{\pi}_i \in \Delta^{\mathcal{A}_i}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}_i^{(t)}, \widehat{\pi}_i - \pi_i^{(t)} \right\rangle, \tag{6.1}$$

where $\pi_i^{(t)} \in \Delta^{\mathcal{A}_i}$ is the strategy of player i at timestep t and $\boldsymbol{u}_i^{(t)}(a_i) = \sum_{\boldsymbol{a}' \in \times_{j=1}^{N} \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}') \mathbb{1}\left(a_i' = a_i\right) \prod_{j' \neq i} \pi_{j'}^{(t)}(a_{j'}')$ for any $a_i \in \mathcal{A}_i$. Then, it is known that the time-average joint strategy $\pi_{\text{avg}}^{(T)}$, where $\pi_{\text{avg}}^{(T)}(\boldsymbol{a}) := \frac{1}{T}\sum_{t=1}^{T}\prod_{i\in[N]} \pi_i^{(t)}(a_i)$ for any $\boldsymbol{a} \in \times_{i=1}^{N} \mathcal{A}_i$, is an ϵ-CCE, with

$$\epsilon := \max_{i \in [N]} \left\{\frac{1}{T} R_i^{(T),\text{external}}\right\}.$$

By applying the algorithm in Section 4 (for **InstUtil Rank** feedback) or Section 5 (for **AvgUtil Rank** feedback), we achieve sublinear $R_i^{(T),\text{external}}$ for each player $i \in [N]$. Note that $P^{(T)}$ in Assumption 4.1 can be bounded by the summation of all players' strategy variation. Specifically, we have the following result.

Lemma 6.1. *For any $T > 0$ and sequence of strategy profiles $\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(T)}\right)$, the variation of utility vectors of any player $i \in [N]$ satisfies that*

$$\sum_{t=2}^{T} \left\| \boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)} \right\| \leq \max_{j \in [N]} \sqrt{|\mathcal{A}_j|} \prod_{j'=1}^{N} |\mathcal{A}_{j'}| \sum_{t=2}^{T} \sum_{j=1}^{N} \left\| \pi_j^{(t)} - \pi_j^{(t-1)} \right\|. \tag{6.2}$$

The proof is postponed to Appendix L. Therefore, according to Lemma 6.1, to ensure $P^{(T)}$ to be sublinear in T, Alg need to additionally satisfy the following assumption.

Assumption 6.2 (Sublinear variation of strategies)**.** *The (full-information) adversarial online learning algorithm* Alg *needs to satisfy the following condition: for any $T > 0$, $t \in [T-1]$, and sequence of utility vectors $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t} \in \left([-1,1]^{\mathcal{A}}\right)^t$, we have*

$$\left\| \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right) - \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t+1}\right) \right\| \leq \eta,$$

where $\eta = \Theta\left(T^{-w}\right)$ for some constant $w \in (0,1)$.

Mirror descent (c.f. Wei et al. (2021, Lemma 1) and Liu et al. (2023, Lemma C.5)) and FTRL (see Lemma M.3 for the proof), along with any strongly convex regularizer, both satisfy this property. When Assumption 6.2 is satisfied, one can achieve sublinear regret with **InstUtil Rank**, under both full-information and bandit feedback. The formal statement is as follows.

Theorem 6.3. *Consider* **InstUtil Rank** *with constant $\tau > 0$ and Algorithm 2. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, that satisfies Assumption 6.2, by choosing M, m, γ according to Theorem 4.2 and Theorem 4.3 for different settings respectively, we have that with probability at least $(1-\delta)$, the algorithm finds an ϵ-CCE, with*

$$\epsilon \leq \max_{i \in [N]} \left\{ \frac{1}{T} R_i^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}_i^{(t)}\right)_{t=1}^{T}\right) \right\} + \mathcal{O}\left(\eta^{\frac{1}{3}} \left(\log\left(\frac{T}{\delta}\right)\right)^{\frac{1}{3}}\right) \tag{Full Information}$$

$$\epsilon \leq \max_{i \in [N]} \left\{ \frac{1}{T} R_i^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}_i^{(t)}\right)_{t=1}^{T}\right) \right\} + \mathcal{O}\left(\eta^{\frac{1}{5}} \log\left(\frac{T}{\delta}\right)\right). \tag{Bandit}$$

Under **AvgUtil Rank**, when all the players apply Algorithm 3 and both Assumption 5.2 and Assumption 6.2 are satisfied, the external regret of each player will be sublinear in T according to Theorem 5.3. Finally, we have the statement below.

Theorem 6.4. *Consider* **AvgUtil Rank** *with constant $\tau > 0$ and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any full-information no-regret learning algorithm with numeric utility feedback,* Alg*, that satisfies Assumption 5.2, by choosing M, m, γ according to Theorem 5.3, we have that with probability at least $(1-\delta)$, the algorithm finds ϵ-CCE under full-information, with*

$$\epsilon \leq \max_{i \in [N]} \left\{ \frac{1}{T} R_i^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}_i^{(t)}\right)_{t=1}^{T}\right) \right\} + \mathcal{O}\left(L T^{\frac{5}{3}} \log\left(\frac{2T}{\delta}\right)\right). \tag{Full Information}$$

When M, m, γ are chosen according toTheorem 5.4 and Assumption 6.2 is also satisfied, the following holds under bandit feedback,

$$\epsilon \leq \max_{i \in [N]} \left\{ \frac{1}{T} R_i^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}_i^{(t)}\right)_{t=1}^{T}\right) \right\} + \widetilde{\mathcal{O}}\left(\left(\log\left(\frac{1}{\delta}\right)\right)^2 \left(L^{\frac{1}{3}} \eta^{\frac{1}{6}} + L^{\frac{1}{2}}\right) T^{\frac{4}{9}}\right). \tag{Bandit}$$

Remark 6.5. *With the hardness in Theorem C.2, our no-regret results under* **AvgUtil Rank** *for both the online and game settings hold for a constant $\tau > 0$ (that cannot be arbitrarily small). However, we note that it may be possible when $\tau \to 0^+$ in the game setting: with such a* deterministic *ranking model, the* best-response *action against the history play of the opponents is now available, leading to the celebrated algorithm of* fictitious-play *(FP) (Robinson, 1951; Brown, 1951). FP is known to converge to an equilibrium in certain games (Robinson, 1951; Monderer & Shapley, 1996; Sela, 1999; Berger, 2005) (with (slow) convergence rates (Robinson, 1951; Daskalakis & Pan, 2014; Abernethy et al., 2021)), despite that it fails to be no-regret in the online setting (Fudenberg & Levine, 1995; 1998).*

REFERENCES

Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Fast convergence of fictitious play for diagonal payoff matrices. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1387–1404. SIAM, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Carliss Y Baldwin. The architecture of platforms: A unified view. *Platforms, Markets and Innovation/Edward Elgar Publishing Limited*, 2009.

Ulrich Berger. Fictitious play in 2×n games. *Journal of Economic Theory*, 120(2):139–154, 2005.

George W Brown. Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, 13(1): 374, 1951.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Constantinos Daskalakis and Qinxuan Pan. A counter-example to karlin's strong conjecture for fictitious play. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 2014.

Yihan Du, Siwei Wang, and Longbo Huang. Dueling bandits: From two-dueling to multi-dueling. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2020.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R Srikant. Exploration-driven policy optimization in RLHF: Theoretical insights on efficient data utilization. In *Forty-first International Conference on Machine Learning*, 2024.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient exploration for llms. *arXiv preprint arXiv:2402.00396*, 2024.

S Rasoul Etesami and R Srikant. Decentralized and uncoordinated learning of stable matchings: A game-theoretic approach. *arXiv preprint arXiv:2407.21294*, 2024.

Drew Fudenberg and David K Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, 1995.

Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Can Küçükgül, Özalp Özer, and Shouqiang Wang. Engineering social learning: Information design of time-locked sales campaigns for online platforms. *Management Science*, 68(7):4899–4918, 2022.

Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *International Conference on Learning Representations (ICLR)*, 2023.

R Duncan Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 565–582, 2015.

Davide Maran, Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Nicola Gatti, and Marcello Restelli. Bandits with ranking feedback. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Dov Monderer and Lloyd S Shapley. Fictitious play property for games with identical interests. *Journal of Economic Theory*, 68(1):258–265, 1996.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1029–1038. PMLR, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, 54(2):296–301, 1951.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both-world analyses for online learning from preferences. *arXiv preprint arXiv:2202.06694*, 2022.

Aadirupa Saha and Aditya Gopalan. Combinatorial bandits with relative feedback. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning (ICML)*, 2021.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 6263–6289. PMLR, 2023.

Aner Sela. Fictitious play in "one-against-all" multi-player games. *Economic Theory*, 14(3):635–651, 1999.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning (ICML)*, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Notations and Preliminaries

For any integer N, we define $[N] \coloneqq \{1, 2, ..., N-1, N\}$ to denote the set of positive integers no larger than N. We use bold notation $\boldsymbol{x}$ to denote a finite-dimensional vector, and x_i to denote the ith element of the vector. Moreover, we extend the space of the vector index from the integer space to any discrete set. Specifically, for any discrete set $\mathcal{S}$, let $\mathbb{R}^{\mathcal{S}}$ denote the $|\mathcal{S}|$ dimensional real space, where the $\mathcal{S} \ni s^{th}$ element of any $\boldsymbol{x} \in \mathbb{R}^{\mathcal{S}}$ can be written as x_s or $x(s)$. For any vector $\boldsymbol{x} \in \mathbb{R}^m$, let $\|\boldsymbol{x}\|_p$ be its L_p-norm and we use $\|\boldsymbol{x}\|$ to denote the L_2-norm by default. For any convex compact set $\mathcal{C} \subseteq \mathbb{R}^n$ and $\boldsymbol{x} \in \mathbb{R}^n$, let $\mathrm{Proj}_{\mathcal{C}}(\boldsymbol{x}_0) = \mathrm{argmin}_{\boldsymbol{x} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{x}_0\|$. For any event e, let $\mathbb{1}(e)$ be its indicator, which is equal to one when e holds and zero otherwise.

For any discrete set $\mathcal{S}$, let $|\mathcal{S}|$ denote its cardinality, $\Delta^{\mathcal{S}} \coloneqq \left\{\boldsymbol{x} \in \mathbb{R}^{\mathcal{S}} \colon \sum_{s \in \mathcal{S}} x_s = 1,\ x_s \geq 0 \text{ for all } s \in \mathcal{S}\right\}$ be the probability simplex over $\mathcal{S}$, and $\mathbf{1}(\mathcal{S})$ be an all-one vector with each index being elements in $\mathcal{S}$. Additionally, for any discrete set $\mathcal{S}$, let $\Sigma(\mathcal{S})$ be the set containing all the permutations of the elements in $\mathcal{S}$. We will use $\mathrm{sig}(x) \coloneqq \frac{\exp(x)}{1+\exp(x)} \colon \mathbb{R} \to \mathbb{R}$ to denote the logistic function.

A.1 Online Learning

We focus on online learning in a *non-stochastic* and potentially *adversarial* environment, where an agent interacts with the environment for multiple timesteps, by taking an action and then receiving some feedback at each timestep. The agent's action set is finite and denoted as $\mathcal{A} \coloneqq \left\{a^1, a^2, \ldots, a^{|\mathcal{A}|}\right\}$ with $|\mathcal{A}| > 1$. At each timestep t, the agent will commit to a strategy $\pi^{(t)} \in \Delta^{\mathcal{A}}$ and receive a utility vector $\boldsymbol{u}^{(t)} \in [-1, 1]^{\mathcal{A}}$ afterwards, in the full-information setting. The agent aims to minimize her (external) *regret*, which is the difference between her accumulated utility and the highest accumulated utility in hindsight by playing a fixed strategy across all timesteps. Formally, for any integer $T > 0$, the regret is defined as

$$R^{(T),\text{external}} \coloneqq \max_{\widehat{\pi} \in \Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \pi^{(t)} \right\rangle. \tag{A.1}$$

Since our goal is to minimize the regret, which is not affected if the vector $\boldsymbol{u}^{(t)}$ is offset by some constant at each timestep t. Hence, without loss of generality, we assume $u^{(t)}(a^{|\mathcal{A}|}) = 0$, *i.e.,* the last action always receives a zero utility for any $\boldsymbol{u}^{(t)}$ and $t \in [T]$.

A.1.1 Online Learning Algorithms with Numeric Feedback

Our results later will be *modular*, in the sense that *any* standard online learning algorithms with (full-information) numeric feedback, including projected gradient descent (PGD), multiplicative-weight update (MWU), and follow-the-regularized-leader (FTRL) in general (Hazan et al., 2016), can be used as a *black-box* oracle in our algorithms to be designed later. As a preliminary, we formally introduce such oracles here: we use $\mathrm{Alg} \colon \bigcup_{t=0}^{\infty} \left(\mathbb{R}^{\mathcal{A}}\right)^t \to \Delta^{\mathcal{A}}$ to denote such an online learning algorithm, which is a mapping from a sequence of utility vectors to the distribution over the action set $\mathcal{A}$. Therefore, given utility vectors $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}$ from timestep 1 to t, the algorithm will generate $\pi^{(t+1)} = \mathrm{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right)$ as the strategy at timestep $t+1$. Finally, we can denote the (external-)regret under Alg as

$$R^{(T),\text{external}}\left(\mathrm{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) \coloneqq \max_{\widehat{\pi} \in \Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \mathrm{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t-1}\right) \right\rangle, \tag{A.2}$$

which can be made sublinear in T for any utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$.

A.2 Normal-Form Games

An N-player normal-form game (NFG) can be characterized by a tuple $\left(N, \{\mathcal{A}_i\}_{i=1}^{N}, \{\mathcal{U}_i\}_{i=1}^{N}\right)$, where $\mathcal{A}_i := \left\{a_i^1, a_i^2, \ldots, a_i^{|\mathcal{A}_i|}\right\}$ is the (finite) action set for player $i \in [N]$; $\mathcal{U}_i \colon \bigtimes_{i=1}^{N} \mathcal{A}_i \to [-1, 1]$ ($\bigtimes$ is the Cartesian product of sets) is the utility function of player i, where $\mathcal{U}_i(a_1, a_2, ..., a_N)$ is the utility of player i when player $j \in [N]$ takes action a_j. We call $\boldsymbol{a} \colon = (a_1, a_2, ..., a_N)$ the *joint action* and let $\boldsymbol{a}_{-i} \colon = (a_1, ..., a_{i-1}, a_{i+1}, ..., a_N)$. Player $i \in [N]$ can choose a strategy $\pi_i \in \Delta^{\mathcal{A}_i}$, and we call $\bigtimes_{i=1}^{N} \Delta^{\mathcal{A}_i} \ni \boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ a *strategy profile*. When a strategy profile $\boldsymbol{\pi}$ is implemented, each player $i \in [N]$ has an expected utility of $\sum_{\boldsymbol{a} \in \bigtimes_{j=1}^{N} \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}) \prod_{j \in [N]} \pi_j(a_j)$. Lastly, we use the unbold notation $\pi \in \Delta^{\bigtimes_{i=1}^{N} \mathcal{A}_i}$ to denote the (possibly correlated) joint strategy of all players, where $\pi(\boldsymbol{a})$ is the probability of choosing the joint action $\boldsymbol{a} \in \bigtimes_{i=1}^{N} \mathcal{A}_i$.

In this paper, we focus on finding the ϵ-approximate *coarse correlated equilibrium* (ϵ-CCE) of the NFG, which is a probability distribution over the joint action set. It is formally defined as follows:

Definition A.1 (ϵ-CCE)**.** *For any joint strategy $\pi \in \Delta^{\bigtimes_{i=1}^{N} \mathcal{A}_i}$, it is an ϵ-CCE if*

$$\max_{i \in [N]} \max_{\widehat{\pi}_i \in \Delta^{\mathcal{A}_i}} \sum_{\boldsymbol{a} \in \bigtimes_{j=1}^{N} \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}) \left(\widehat{\pi}_i(a_i) \sum_{a_i' \in \mathcal{A}_i} \pi(a_i', \boldsymbol{a}_{-i}) - \pi(\boldsymbol{a}) \right) \leq \epsilon. \tag{ϵ-CCE}$$

When $\epsilon = 0$, we refer to it as a CCE. Moreover, when the joint strategy π can be written as the *product* of individual players' strategies, *i.e.,* $\pi(\boldsymbol{a}) = \prod_{i=1}^{N} \pi_i(a_i)$ for any $\boldsymbol{a} \in \bigtimes_{j=1}^{N} \mathcal{A}_j$, an ϵ-CCE reduces to an ϵ-Nash equilibrium (NE).

A.3 Related Work

Dueling Bandits. Using comparison and/or ranking feedback for sequential decision-making has mostly been studied under the framework of dueling bandits (Yue et al., 2012; Saha & Gaillard, 2022; Saha & Gopalan, 2019; Du et al., 2020; Saha et al., 2021; Dudík et al., 2015), where the agent takes two (or multiple) actions at each timestep, and receives a ranking of them as feedback. Different from our setting, the ranking feedback in these works was only based on the instantaneous utility at that timestep, while our results can address settings with both instantaneous and time-average utilities for ranking. More importantly, the regret notions studied in these works are particularly designed for the dueling-bandit setting, and are different from the classical external regret we focus on here. Finally, dueling bandits mostly focused on environments that are *stationary* and *stochastic* (Yue et al., 2012; Saha & Gaillard, 2022; Saha & Gopalan, 2019; Du et al., 2020), while we focus on the *non-stochastic* setting where the environment is arbitrary and potentially *adversarial*, as in the online learning setups focused on in Shalev-Shwartz et al. (2012); Hazan et al. (2016). Due to the last two differences, the implication of these algorithms in learning-in-games is unclear, while our algorithms converge to the CCE of the game, as a corollary of the no-(external-)regret guarantee.

Reinforcement Learning from Human Feedback (RLHF) and Preference-Based RL Inspired by the successes in aligning large-language-models (LLMs) (Ouyang et al., 2022), reinforcement learning from human feedback has received increasing attention. RLHF is usually instantiated as *preference-based* learning, where the humans rank the model outputs based on their preferences, and a reward model is then estimated from the feedback, which will be further used for model fine-tuning. This way, RLHF is oftentimes implemented in an *offline* fashion, where batch feedback data are used for reward model estimation (Ziegler et al., 2019; Bai et al., 2022; Ouyang et al., 2022; Zhu et al., 2023; Park et al., 2024). Recently, online versions of RLHF have also been developed (Dwaracherla et al., 2024; Du et al., 2024; Xie et al., 2024; Cen et al., 2024; Zhang et al., 2024), where the *exploration* issue was addressed with online feedback. In fact, beyond fine-tuning LLMs, preference-based RL has also been studied in classical Markov decision process models, with online feedback to provably trade-off exploration and exploitation (Novoseller et al., 2020; Saha et al.,

Linear Regret	Full Information	Bandit
InstUtil Rank	$\tau \leq \mathcal{O}(1)$	
AvgUtil Rank	$\tau \leq \mathcal{O}\left(\frac{1}{T \log T}\right)$	$\tau \leq \mathcal{O}\left(\frac{1}{\log T}\right)$

Sublinear regret	Full Information	Bandit
InstUtil Rank	Assumption 4.1	
AvgUtil Rank	✓	Assumption 4.1 ($q < \frac{1}{3}$)

Table 1: Summary of the contributions in this paper, including the negative results (top) and the positive results (bottom). The table on the top shows the conditions under which for any online learning algorithm, there exists an instance such that it will suffer a linear regret, with either *instantaneous* or *time-average* utility ranking feedback. The negative result for the bandit, time-average utility ranking feedback case, *i.e.,* the (**AvgUtil Rank**, Bandit) block in the top table, can be viewed as also strengthening the hardness result in the recent work Maran et al. (2024), which was developed for the case with $\tau \to 0^+$. The bottom table shows the conditions that are sufficient to achieve sublinear regret with *instantaneous* and *time-average* utility ranking feedback, respectively.

2023; Xu et al., 2020). However, the utility/reward functions in these works are stationary, and the regret notions that extend those in the dueling-bandits literature are different from ours. Hence, these algorithms do not apply to our adversarial online learning and game-theoretic settings.

Recent Work by Maran, Bacchiocchi, Stradi, Castiglioni, Gatti, and Restelli (2024). The work closest to ours is the very recent one Maran et al. (2024), which studied bandits with ranking feedback, also under the standard (external-)regret metric. Different from the ranking model in dueling bandits, the model in Maran et al. (2024) is based on the time-average utilities, a setting also considered in our paper. In contrast to our paper, Maran et al. (2024) focused on the *stochastic* bandits setting where the utility functions are stationary, while our focus is on the *adversarial* and game-theoretic settings, with *both* instantaneous and time-average utility based rankings. Moreover, the ranking model in Maran et al. (2024) corresponds to the case of $\tau \to 0^+$ in our framework. Notably, Maran et al. (2024) also provided a hardness result for the adversarial setting (for $\tau \to 0^+$), while our hardness results (with different instances) are stronger in that they allow a wider range of τ and also hold for the full-information setting (c.f. Table 1).

B Algorithms

Figure 2 is the illustration of learning with **InstUtil Rank** and Figure 3 is that of **AvgUtil Rank**. The utility estimator is Algorithm 1, the full algorithm for learning with **InstUtil Rank** is Algorithm 2, and that of **AvgUtil Rank** is Algorithm 3.

C Hardness for the Online Setting

In this section, we present hard instances to show that online learning in non-stochastic and potentially adversarial environments can be hard in general, under both **InstUtil Rank** and **AvgUtil Rank**, even when there are only two actions.

Theorem C.1 in the following shows that for any temperature not larger than a *constant*, there exists a set of utility vectors such that the expected regret is linear with **InstUtil Rank**, for both full-information and bandit settings.

Theorem C.1. *Consider* **InstUtil Rank**. *For any $T > 0$, temperature $0 < \tau \leq 0.1$, and online learning algorithm, there exists a sequence of utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$ such that*

$$\min\left\{\mathbb{E}\left[R^{(T),\text{external}}\right], \mathbb{E}\left[R^{(T)}\right]\right\} \geq \Omega\left(T\right)$$

in both full-information and bandit settings. The expectation is taken over the randomness of the algorithm and the ranking.

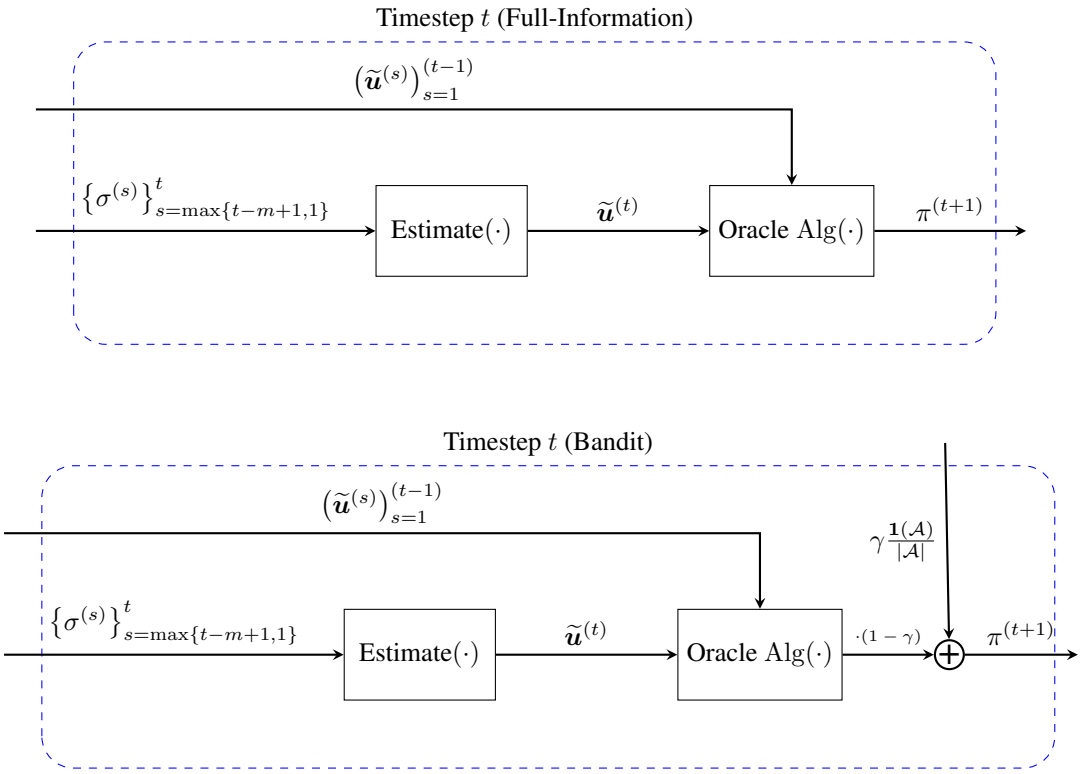

Figure 2: The diagram of Algorithm 2 with **InstUtil Rank** under full-information feedback (top) and bandit feedback (bottom). $\oplus$ represents the addition of $(1-\gamma)$ times the output the Alg and γ times a uniform distribution over $\mathcal{A}$.

Algorithm 1 Estimate$\left(\left\{\sigma^{(s)}\right\}_{s=1}^{m'}\right)$

Input: A set consisting of m' permutations : $\left\{\sigma^{(s)}\right\}_{s=1}^{m'}$ and temperature $\tau > 0$.
for $j = 1, 2, \ldots, |\mathcal{A}| - 1$ **do**
  for $s = 1, \ldots, m'$ **do**
    Calculate $n_{j,1}^{(s)}, n_{j,2}^{(s)}$ defined as

$$n_{j,1}^{(s)} \coloneqq \sum_{i,k\in[K]} \mathbb{1}\left(\sigma^{(s)}(i) = a^j, \sigma^{(s)}(k) = a^{|\mathcal{A}|} \text{and } i < k\right),$$
$$n_{j,2}^{(s)} \coloneqq \sum_{i,k\in[K]} \mathbb{1}\left(\sigma^{(s)}(i) = a^j, \sigma^{(s)}(k) = a^{|\mathcal{A}|} \text{and } i > k\right).$$

  end for
  Let $\mathcal{T}_j \coloneqq \left\{s \in [1, m'] \colon n_{j,1}^{(s)} + n_{j,2}^{(s)} > 0\right\}$
  Let $\mathrm{sig}^{-1}(x) \colon (0,1) \to \mathbb{R} \coloneqq \log\frac{x}{1-x}$ be the inverse function of sig. The utility of action a^j is then estimated as

$$\widetilde{u}(a^j) = \begin{cases} \mathrm{Proj}_{[-1,1]}\left(\tau \mathrm{sig}^{-1}\left(\frac{1}{|\mathcal{T}_j|}\cdot\sum_{s\in\mathcal{T}_j}\left(\frac{n_{j,1}^{(s)}}{n_{j,1}^{(s)}+n_{j,2}^{(s)}}\right)\right)\right) & |\mathcal{T}_j| > 0 \\ 0 & |\mathcal{T}_j| = 0. \end{cases}$$

end for
Return $\widetilde{\boldsymbol{u}} = \left(\widetilde{u}(a^1), \widetilde{u}(a^2), \ldots, \widetilde{u}\left(a^{|\mathcal{A}|-1}\right), 0\right)$

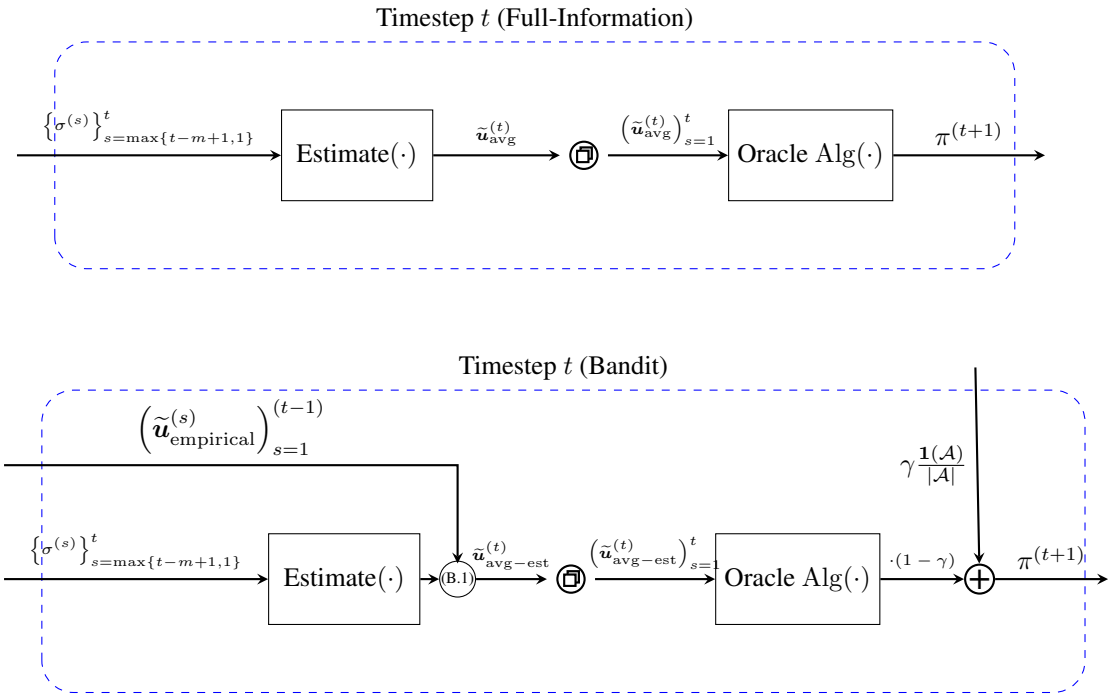

Figure 3: The diagram of Algorithm 3 with **AvgUtil Rank** under full-information feedback (top) and bandit feedback (bottom). represents copying the estimated utility vector for t times. represents the addition of $(1-\gamma)$ times the output the Alg and γ times a uniform distribution over $\mathcal{A}$.

Algorithm 2 Learning with **InstUtil Rank**

Input: Action space $\mathcal{A}$, any full-information no-regret algorithm Alg under numeric feedback, selected action number K, estimation window size m, and exploration rate γ.
Initialize $\pi^{(1)}$ as uniform distribution $\frac{\mathbf{1}}{|\mathcal{A}|}$ over $\mathcal{A}$
for timestep$t = 1, 2, \ldots, T$ **do**
  if Full-information setting **then**
    $K = |\mathcal{A}|$ in this case. Select all $|\mathcal{A}|$ actions.
  else if Bandit setting **then**
    Sample K actions independently with replacement from $\pi^{(t)}$.
  end if
  Receive a ranking feedback $\sigma^{(t)} = \left(\sigma^{(t)}(1), \sigma^{(t)}(2), \ldots, \sigma^{(t)}(K)\right)$ from the environment.
  $\widetilde{\boldsymbol{u}}^{(t)} = \text{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=\max\{t-m+1,1\}}^{t}\right)$ by calling Algorithm 1.
  if Full-information setting **then**
    $\pi^{(t+1)} \leftarrow \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right)$.
  else if Bandit setting **then**
    $\pi^{(t+1)} \leftarrow (1-\gamma)\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right) + \gamma\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}$.
  end if
end for

To prove Theorem C.1, we need to construct two sets of utility vectors, which are not distinguishable from **InstUtil Rank**, but being no-regret in one of them will result in linear regret in the other. The detailed proof can be found in Appendix F.

Next, we show in Theorem C.2 that when **AvgUtil Rank** is used, and τ is small enough, it is also impossible to achieve sublinear regret.

Algorithm 3 Learning with **AvgUtil Rank**

Input: Action space $\mathcal{A}$, any full-information no-regret algorithm Alg under numeric feedback, selected action number K, estimation window size m, exploration rate γ, and block size M.
Initialize $\pi^{(1)}$ as uniform distribution $\frac{\mathbf{1}}{|\mathcal{A}|}$ over $\mathcal{A}$
for timestep $t = 1, 2, \ldots, T$ **do**
  if Full-information setting **then**
    $K = |\mathcal{A}|$ in this case. Select all $|\mathcal{A}|$ actions.
  else if Bandit setting **then**
    Sample K actions independently with replacement from $\pi^{(t)}$.
  end if
  Receive a ranking feedback $\sigma^{(t)} = \left(\sigma^{(t)}(1), \sigma^{(t)}(2), \ldots, \sigma^{(t)}(|\mathcal{A}|)\right)$ from the environment.
  if Full-information setting **then**
    $\widetilde{\boldsymbol{u}}^{(t)}_{\mathrm{avg}} = \mathrm{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=\max\{t-m+1,1\}}^{t}\right)$ by calling Algorithm 1.
    $\pi^{(t+1)} \leftarrow \mathrm{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(t)}_{\mathrm{avg}}\right)_{s=1}^{t}\right)$, *i.e.,* the strategy generated by Alg when all utility vectors from timestep 1 to t are $\widetilde{\boldsymbol{u}}^{(t)}_{\mathrm{avg}}$.
  else if Bandit setting **then**
    $\widetilde{\boldsymbol{u}}^{(t)}_{\mathrm{empirical}} = \mathrm{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=\max\{t-m+1,1\}}^{t}\right)$ by calling Algorithm 1.
    Let $n^{(t)}(a) \coloneqq \sum_{s=1}^{t} \#_{o^{(s)}}(a)$ for any $a \in \mathcal{A}$ as the number of times action a is proposed up to timestep t. Then, the estimated average utility is

$$\widetilde{u}^{(t)}_{\mathrm{avg-est}}(a) \coloneqq \begin{cases} \frac{1}{\lfloor t/M \rfloor} \sum_{s=1}^{\lfloor t/M \rfloor} \frac{\widetilde{u}^{(s\cdot M)}_{\mathrm{empirical}}(a) n^{(s\cdot M)}(a) - \widetilde{u}^{((s-1)M)}_{\mathrm{empirical}}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)} & t \geq M \\ 0 & t < M \end{cases} \tag{B.1}$$

    $\pi^{(t+1)} \leftarrow (1-\gamma)\mathrm{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(t)}_{\mathrm{avg-est}}\right)_{s=1}^{t}\right) + \gamma \frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}$.
  end if
end for

Theorem C.2. *Consider* **AvgUtil Rank** *with full-information feedback. For any $T > 0$, temperature $0 < \tau \leq \mathcal{O}\left(\frac{1}{T \log T}\right)$, and online learning algorithm, there exists $T' \geq T$ and a sequence of utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T'}$ such that*

$$\min\left\{\mathbb{E}\left[R^{(T'),\mathrm{external}}\right], \mathbb{E}\left[R^{(T')}\right]\right\} \geq \Omega\left(\frac{T'}{\log T'}\right).$$

The expectation is taken over the randomness of the algorithm and the ranking.

Given Theorem C.2, it is impossible to achieve sublinear regret with **AvgUtil Rank** when τ is very small. However, in Section 5, we will close the gap by showing that when τ is a *constant*, we can achieve sublinear regret with **AvgUtil Rank**, even without Assumption 4.1.

Due to the different instantiations of $\boldsymbol{r}^{(t)}$ in the full-information and the bandit settings under **AvgUtil Rank** in the ranking model (PL), we have a separate hardness result for the bandit setting, which is stronger than Theorem C.2 as it allows a larger τ.

Theorem C.3. *Consider* **AvgUtil Rank** *with bandit feedback. For any $T > 0$, temperature $0 < \tau \leq \mathcal{O}\left(\frac{1}{\log T}\right)$, and online learning algorithm, there exists a sequence of utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{4T}$ such that*

$$\min\left\{\mathbb{E}\left[R^{(4T),\mathrm{external}}\right], \mathbb{E}\left[R^{(4T)}\right]\right\} \geq \Omega\left(T\right).$$

The expectation is taken over the randomness of the algorithm and the ranking.

The proof is postponed to Appendix F.

D EXPERIMENTS

We evaluate our algorithms in two-player general-sum random-utility games under all combinations of full-information and bandit feedback, as well as **InstUtil Rank** and **AvgUtil Rank** models. The exploitability of different game parameters are provided in the following figures.

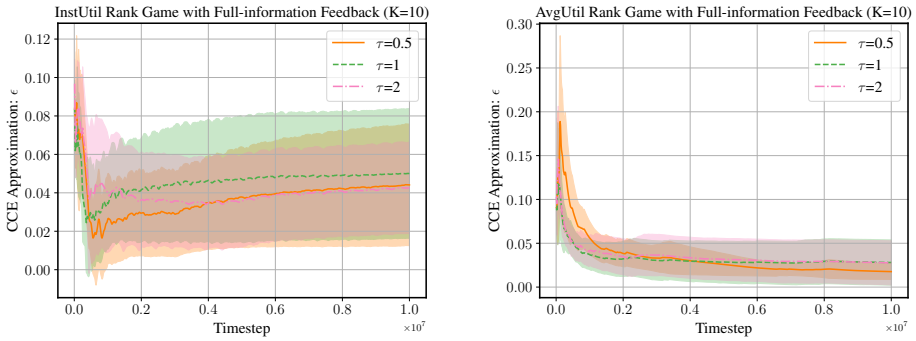

Figure 4: The exploitability for the full-information setting under both **InstUtil Rank** and **AvgUtil Rank** feedback. The performance is tested under different temperatures τ. Each parameter combination is tested 10 times with different random seeds. We pick the best m, and γ for each figure.

The utility estimation of each game utilizes Algorithm 1. The (full-information adversarial) no-regret learning oracle, Alg, for **InstUtil Rank** is PGD (Hazan et al., 2016) and for **AvgUtil Rank** is FTRL with L_2-regularization (Hazan et al., 2016).

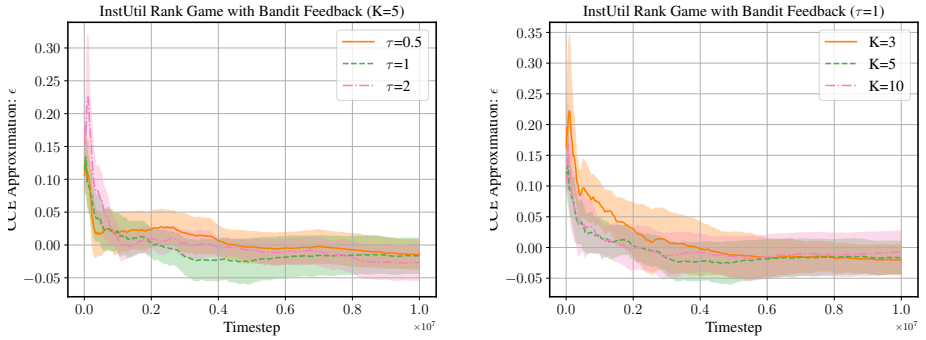

Figure 5: The exploitability for the bandit feedback setting under **InstUtil Rank** feedback. The performance is tested under different temperatures τ and sampled action size K. Each parameter combination is tested 10 times with different random seeds. We pick the best m, and γ for each figure.

To pick better hyper-parameters for different game settings, we performed a grid-search for **InstUtil Rank** on exploration rate γ and estimation window size m. For **AvgUtil Rank**, we perform the grid-search on exploration rate γ, estimation window size m, and the block size M. The parameters searched may differ depending on the full information or bandit feedback settings. All games are run for $T = 10^7$ iterations. Each player in the game has 10 actions. The learning rate was set to $\eta = \frac{1}{\sqrt{T}}$ in all experiments, except in the combination of **AvgUtil Rank** and bandit feedback, where it was set to $\eta = 10^{-6}$. Each parameter combination is tested 10 times with different random seeds. We pick the best m, M, and γ for each figure.

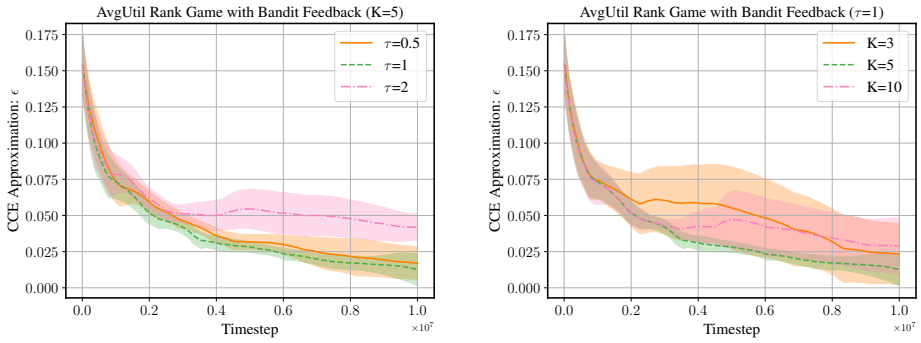

Figure 6: The exploitability for the bandit feedback setting under **AvgUtil Rank** feedback. The performance is tested under different temperatures τ and sampled action size K. Each parameter combination is tested 10 times with different random seeds. We pick the best m, M, and γ for each figure.

For all games, the exploitability decreases as t increases, which shows time-average joint strategy converges to CCE. The equilibrium of the bandit feedback setting for **AvgUtil Rank** is reached slower than **InstUtil Rank**, which fits the regret bound in Theorem 4.3 and Theorem 5.4.

The code for the experiments is provided at the anonymous github link Online-Learning-and-Equilibrium-Computation-with-Ranking-Feedback.[1]

E Utility Estimation

A natural idea to learn under ranking feedback is to use the feedback to *estimate* numeric utility vectors. At each timestep t, we propose using the ranking feedback from the last m steps to predict the utility vector $\boldsymbol{u}$. When $t \geq m$, we use the past m step's permutation $\left\{\sigma^{(s)}\right\}_{s=t-m+1}^{m}$ to estimate utility $\boldsymbol{u}^{(t)}$. Lemma E.1 below shows that when $K > 2$, for any two actions $a \neq b \in \mathcal{A}$ in the proposed action set $o^{(t)}$, in all action pairs a, b in $\sigma^{(t)}$, the proportion of pairs that action a appears before b, is equal to $\operatorname{sig}\left(\frac{u^{(t)}(a)-u^{(t)}(b)}{\tau}\right)$ in expectation. In other words, it is equal to the probability of the permutation (a, b) occurring when only proposing a, b.

Lemma E.1. *Let $\#_{\mathcal{S}}(a) := \sum_{a' \in \mathcal{S}} \mathbb{1}(a' = a)$ represent the number of elements in a multiset $\mathcal{S}$ that are equal to $a \in \mathcal{A}$. For any utility vector $\boldsymbol{u}$, temperature $\tau > 0$, a multiset of proposed actions $\mathcal{S}$ with cardinality $|\mathcal{S}| = K$, and any two actions $a \neq b \in \mathcal{S}$, we have*

$$\frac{1}{\#_{\mathcal{S}}(a) \cdot \#_{\mathcal{S}}(b)} \mathbb{E}\left[\sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \,\middle|\, \mathcal{S}\right] = \operatorname{sig}\left(\frac{u(a) - u(b)}{\tau}\right).$$

The expectation is taken over the randomness in the (PL) *model.*

The proof is postponed to Appendix G.1. With Lemma E.1, the general cases where $K > 2$ actions are proposed can be cast into the case with only two actions being proposed by counting out all existence of the action pairs. Therefore, to estimate $u^{(t)}(a^j)$, we will first construct an unbiased estimator of $\operatorname{sig}\left(\frac{u^{(s)}(a^j)-u^{(s)}(a^{|\mathcal{A}|})}{\tau}\right)$ using Lemma E.1, for all timesteps $s \in [t-m+1, t]$ when both $a^j, a^{|\mathcal{A}|} \in o^{(s)}$. Since we have assumed without loss of generality that $u^{(s)}(a^{|\mathcal{A}|}) = 0$, these values coincide with $\operatorname{sig}\left(\frac{u^{(s)}(a^j)}{\tau}\right)$. Then, by Hoeffding's inequality and monotonicity of the logistic function, with high probability, the mean of the logistic function estimators will be bounded between the

[1] https://anonymous.4open.science/r/Online-Learning-and-Equilibrium-Computation-with-Ranking-Feedback-FB0C

minimum and maximum of $\left\{\text{sig}\left(\frac{u^{(s)}(a^j)}{\tau}\right)\right\}_{s=t-m+1}^{t}$. By Assumption 4.1, since the utility vectors are changing slowly, that mean is guaranteed to be close to $\text{sig}\left(\frac{u^{(t)}(a^j)}{\tau}\right)$. With a good estimate of $\text{sig}\left(\frac{u^{(s)}(a^j)}{\tau}\right)$, we can then take an inverse of $\text{sig}(\cdot)$ to estimate $u^{(s)}(a^j)$. This estimation algorithm is summarized in Algorithm 1 and analyzed in Theorem E.2 below.

Theorem E.2. *Consider* **InstUtil Rank** *and Algorithm 1. Suppose each action is proposed with probability at least $p > 0$ at each timestep $t \in [T]$ and let $\widetilde{\boldsymbol{u}}^{(t)} = \text{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=t-m'+1}^{t}\right)$. Then, for any $\delta \in (0,1)$ and $t \geq m'$, when $m'p^4 \geq 2\log\left(\frac{2}{\delta}\right)$, with probability at least $1-\delta$, the estimate $\widetilde{\boldsymbol{u}}^{(t)}$ satisfies,*

$$\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|_\infty \leq \frac{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}{p}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m'}} + \sum_{s=t-m'+1}^{t-1}\left\|\boldsymbol{u}^{(s+1)} - \boldsymbol{u}^{(s)}\right\|_\infty .$$

When taken δ, p as constants, the accumulated estimation error $\sum_{t=1}^{T}\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|_\infty$ is bounded by $\mathcal{O}\left(\frac{T}{\sqrt{m'}} + m'P^{(T)}\right)$, which implies that sublinear accumulated estimation error is achieved only when $P^{(T)}$ is sublinear (Assumption 4.1). Moreover, Theorem E.2 implies that when $\tau \to 0^+$, the estimation error upperbound goes to $+\infty$. This makes intuitive sense: when $\tau \to 0^+$, only the action with the highest utility is chosen (deterministically), so it becomes impossible to estimate the gap between the utilities of any two actions. At the opposite end, when $\tau \to +\infty$, the estimation error upperbound also goes to $+\infty$, since the ranking is always sampled uniformly regardless of the utility vectors.

The proof of Theorem E.2 is provided in Appendix G. In the following, we will show how to achieve sublinear regret in both full-information and bandit settings with **InstUtil Rank**.

F Proof of Appendix C

$R^{(T)} \geq R^{(T),\text{external}}$ since we can choose $\pi^{(t)}(a) = \frac{\sum_{a' \in o^{(t)}} \mathbb{1}(a=a')}{K}$ for any action $a \in \mathcal{A}$ so that $\frac{1}{K}\sum_{a \in o^{(t)}} u^{(t)}(a) = \left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle$. Therefore, in the rest of this section, we will only show the lowerbound of $R^{(T),\text{external}}$.

F.1 Proof of Theorem C.1

Theorem C.1. *Consider* **InstUtil Rank**. *For any $T > 0$, temperature $0 < \tau \leq 0.1$, and online learning algorithm, there exists a sequence of utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$ such that*

$$\min\left\{\mathbb{E}\left[R^{(T),\text{external}}\right], \mathbb{E}\left[R^{(T)}\right]\right\} \geq \Omega\left(T\right)$$

in both full-information and bandit settings. The expectation is taken over the randomness of the algorithm and the ranking.

Proof. Consider an online learning problem with $\mathcal{A} = \{a, b\}$, so that the utility vector can be represented as $(u(a), u(b))$. There are two instances with $\tau = 0.1$.

In the first instance, there are two types of utility vectors $(-0.5, 0)$ and $(0.15, 0)$. At each timestep, the adversary will choose $(-0.5, 0)$ with probability $\frac{4}{13}$ and the other with probability $\frac{9}{13}$.

In the second instance, there are two types of utility vectors $(-0.02, 0)$ and $(0.1, 0)$. Let $\text{sig}(x)\colon \mathbb{R} \to \mathbb{R} := \frac{\exp(x)}{1+\exp(x)}$ be the logistic function. At each timestep, the adversary will choose $(-0.02, 0)$ with probability $\frac{4\text{sig}(-5)/13 + 9\text{sig}(1.5)/13 - \text{sig}(1)}{\text{sig}(-0.2) - \text{sig}(1)} \approx 0.58$ and the other with probability 0.42.

The expected utility of action b in both instances is 0. The expected utility of action a in the first instance is -0.05. The expected utility of action a in the second instance is 0.03.

Moreover, the probability of the online learning agent observing permutation (a, b) in the first instance is

$$\frac{4}{13}\text{sig}(-5) + \frac{9}{13}\text{sig}(1.5),$$

which is equal to the probability of observing it in the second instance

$$\frac{4\text{sig}(-5)/13 + 9\text{sig}(1.5)/13 - \text{sig}(1)}{\text{sig}(-0.2) - \text{sig}(1)}\text{sig}(-0.2) + \left(1 - \frac{4\text{sig}(-5)/13 + 9\text{sig}(1.5)/13 - \text{sig}(1)}{\text{sig}(-0.2) - \text{sig}(1)}\right)\text{sig}(1).$$

Therefore, for any algorithm that generates $\left(\pi^{(t)}\right)_{t=1}^{T}$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right] = \sum_{t=1}^{T}\mathbb{E}\left[\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right] \overset{(i)}{=} \sum_{t=1}^{T}\left\langle \mathbb{E}\left[\boldsymbol{u}^{(t)}\right], \mathbb{E}\left[\pi^{(t)}\right]\right\rangle.$$

(i) is because $\boldsymbol{u}^{(t)}$ is independent of $\pi^{(t)}$ given our generation processing both instances. Moreover,

$$\mathbb{E}\left[\pi^{(t)}\right] = \sum_{\sigma^{(1)},\ldots,\sigma^{(t-1)}\in\Sigma(\mathcal{A})} \mathbb{P}\left(\sigma^{(1)},\ldots,\sigma^{(t-1)}\right)\mathbb{E}\left[\pi \mid \sigma^{(1)},\ldots,\sigma^{(t-1)}\right].$$

The first term $\mathbb{P}\left(\sigma^{(1)},\ldots,\sigma^{(t-1)}\right)$ is equal in two instances according to the discussion above, and the second term $\mathbb{E}\left[\pi \mid \sigma^{(1)},\ldots,\sigma^{(t-1)}\right]$ is also equal since it only depends on the algorithm. Therefore, $\mathbb{E}\left[\pi^{(t)}\right]$ is the same in both instances.

However, $\mathbb{E}\left[\boldsymbol{u}^{(t)}\right] = (-0.05, 0)$ in the first instance but $(0.03, 0)$ in the second. Therefore, whenever achieving sublinear regret in the first instance, the algorithm will suffer a linear regret in the second instance, and vice versa. □

F.2 Proof of Theorem C.2

Theorem C.2. *Consider* **AvgUtil Rank** *with full-information feedback. For any $T > 0$, temperature $0 < \tau \leq \mathcal{O}\left(\frac{1}{T\log T}\right)$, and online learning algorithm, there exists $T' \geq T$ and a sequence of utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T'}$ such that*

$$\min\left\{\mathbb{E}\left[R^{(T'),\text{external}}\right], \mathbb{E}\left[R^{(T')}\right]\right\} \geq \Omega\left(\frac{T'}{\log T'}\right).$$

The expectation is taken over the randomness of the algorithm and the ranking.

Proof. We use $(u(a), u(b))$ to denote the utility vector when the action set is $\mathcal{A} = \{a, b\}$. In the following, we will show a hard instance for $\tau \to 0^+$, *i.e.,* we always observe the action with higher utility ranks first in the permutation. Then, we will show that $\tau \leq \mathcal{O}(\frac{1}{\log T})$ can be reduced to $\tau \to 0^+$.

The utility vector at timestep 1 is $\boldsymbol{u}^{(1)} = (0.5, 0)$. Then we will construct the rest of the utility vectors.

We call the following an action-a construction, since except for the last timestep, the observation is always action a is better. Let $K \in \mathbb{N}$ be the smallest integer such that $2^K \geq T$.

$$\begin{aligned}
\text{Sequence } 0 =& (0,1),(0,0)\\
\text{Sequence } 1 =& (1,0),(0,1),(0,1),(0,0)\\
\text{Sequence } 2 =& (1,0),(1,0),(1,0),(0,1),(0,1),(0,1),(0,1),(0,0)\\
&\ldots\\
\text{Sequence } K-1 =& \underbrace{(1,0),\ldots,(1,0)}_{2^{K-1}-1},\underbrace{(0,1),\ldots,(0,1)}_{2^{K-1}},(0,0)\\
\text{Sequence } K =& \underbrace{(1,0),\ldots,(1,0)}_{2^{K}-1},(0,0).
\end{aligned} \tag{F.1}$$

Lemma F.1 in the following shows that at least one of the sequences will incur a low average utility for the algorithm.

Lemma F.1. *Consider* (F.1)*. For any online learning algorithm, at least one of the $K+1$ sequences satisfies that the expected average utility per timestep is less than* $0.5 - \frac{1}{2(K+1)}$.

By Lemma F.1, there exists a sequence with length 2^k for some $k \leq K$ such that the average utility per timestep achieved by the algorithm is less than $0.5 - \frac{1}{2(K+1)}$. We will pick this sequence as the next 2^k utility vectors. If the current utility vector sequence is no less than T, then the hard instance is completed. Otherwise, we will construct the following action-b construction:

$$\begin{aligned}
\text{Sequence } 0 =& (1,0),(0,0)\\
\text{Sequence } 1 =& (0,1),(1,0),(1,0),(0,0)\\
\text{Sequence } 2 =& (0,1),(0,1),(0,1),(1,0),(1,0),(1,0),(1,0),(0,0)\\
&\dots\\
\text{Sequence } K-1 =& \underbrace{(0,1),...,(0,1)}_{2^{K-1}-1},\underbrace{(1,0),...,(1,0)}_{2^{K-1}},(0,0)\\
\text{Sequence } K =& \underbrace{(0,1),...,(0,1)}_{2^{K}-1},(0,0).
\end{aligned}$$

Similarly, except for the last observation, all the observations are action b being the best action. Similar to Lemma F.1, we can show that at least one of the sequences incurs average utility per timestep less than $0.5 - \frac{1}{2(K+1)}$. We will add that sequence to the end of our hard instance.

Let $T' \geq T$ be the length of the final instance. Therefore, the average regret will be $\frac{1}{2(K+1)} = \Omega(\frac{1}{\log T})$. Because from the construction, the best action should get at least $0.5 - \frac{1}{T}$ utility per timestep.

When $\tau \leq \mathcal{O}\left(\frac{1}{\log T}\right)$, from the construction above, the difference between the cumulative utility of the actions is always 0.5. By definition of **AvgUtil Rank**, $1 - \text{sig}\left(\frac{0.5}{\tau}\right) = \mathcal{O}\left(\frac{1}{T}\right)$. Therefore, by union bound, with a non-negligible probability, all permutations will rank the action with a higher utility at first, so that the problem reduces to $\tau \to 0^+$. □

Lemma F.1. *Consider* (F.1)*. For any online learning algorithm, at least one of the $K+1$ sequences satisfies that the expected average utility per timestep is less than* $0.5 - \frac{1}{2(K+1)}$.

Proof. Note that this is online learning, so the strategy $\pi^{(t)}$ is determined by $\boldsymbol{u}^{(1)}, \dots, u^{(t-1)}$. Therefore, in all sequences in action-a construction, since all observations are action a being the best, for any two sequences $k_1 \leq k_2$, the expectation of the strategies are the same for the first 2^{k_1+1} utility vectors. For simplicity, we will use $x^{(t)}$ to denote the probability of choosing action-a at timestep t.

The average utility at sequence 0 is $\frac{1-x^{(1)}}{2}$. The average utility at sequence 1 is $\frac{x^{(1)}}{4} + \frac{1-x^{(2)}}{4} + \frac{1-x^{(3)}}{4}$. We can see that the utility contributed by $x^{(1)}$ to all sequences is

$$\frac{1-x^{(1)}}{2} + \frac{x^{(1)}}{4} + \frac{x^{(1)}}{8} + ... + \frac{x^{(1)}}{2^K} + \frac{x^{(1)}}{2^K} = \frac{1}{2}.$$

Similarly, the contribution of $x^{(2)}, x^{(3)}$ is $\frac{1}{4}$. The contribution of $x^{(4)}, x^{(5)}, \dots, x^{(7)}$ is $\frac{1}{8}$. Therefore, the total contribution of $x^{(1)}, ..., x^{(2^K-1)}$ is $\frac{K}{2}$. There are $K+1$ sequences in total, so that at least one of the sequences has average utility per timestep less than $\frac{K}{2(K+1)} = \frac{1}{2} - \frac{1}{2K+2}$. □

F.3 Proof of Theorem C.3

Theorem C.3. *Consider* **AvgUtil Rank** *with bandit feedback. For any $T > 0$, temperature $0 < \tau \leq \mathcal{O}\left(\frac{1}{\log T}\right)$, and online learning algorithm, there exists a sequence of utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{4T}$ such*

that

$$\min\left\{\mathbb{E}\left[R^{(4T),\text{external}}\right],\mathbb{E}\left[R^{(4T)}\right]\right\}\geq\Omega\left(T\right).$$

The expectation is taken over the randomness of the algorithm and the ranking.

Proof. Consider the following two instances. Both of them satisfy $\mathcal{A}=\left\{a^1,a^2\right\}$ and $K=1$.

$$\text{Instance 1}=\underbrace{(0.1,0),\ldots,(0.1,0)}_{T},\underbrace{(0,0.2),\ldots,(0,0.2)}_{T},\underbrace{(0,1),\ldots,(0,1)}_{2T}$$
$$\text{Instance 2}=\underbrace{(0.1,0),\ldots,(0.1,0)}_{T},\underbrace{(0,0.2),\ldots,(0,0.2)}_{T},\underbrace{(0.4,0.2),\ldots,(0.4,0.2)}_{2T}.$$

We call the first T timesteps as the first phase, the next T timesteps as the second phase, and the last $2T$ timesteps as the third phase.

For any online learning algorithm, it must propose action a^1 for at least $0.9T$ times during the first phase with probability at least $\frac{1}{2}$, since otherwise there is the expected external regret in the first phase is linear. During the second phase, it must propose action a^2 for at least $\frac{0.2T-0.1T}{0.2}=\frac{T}{2}$ times due to the same reason with probability at least $\frac{1}{4}$. Then, at the end of the second phase, with probability at least $\frac{1}{4}$, $u^{(2T)}_{\text{empirical}}(a^2)-u^{(2T)}_{\text{empirical}}(a^1)\geq\frac{0.5T\cdot0.2}{0.1T+0.5T}-0.1=\frac{1}{15}$.

Then, in the third phase of instance 1, the algorithm needs to propose a^2 for at least $\frac{0.2T+2T-0.2T-0.1T}{1}=1.9T$ times with probability at least $\frac{1}{8}$. In other words, a^1 is proposed by no more than $0.1T$ times. Then, in instance 2, at the end of the third phase, $u^{(4T)}_{\text{empirical}}(a^2)-u^{(4T)}_{\text{empirical}}(a^1)\geq\frac{0.5T\cdot0.2}{0.1T+0.5T}-\frac{0.9T\cdot0.1+0.1T\cdot0.4}{0.9T+0.1T}\geq0.03$. Therefore, when $\tau\to0^+$, the observation of instance 1 and instance 2 are the same with probability at least $\frac{1}{8}$. Then, with probability at least $\frac{1}{8}$, according to the discussion above, any learning algorithm will satisfy one of the following,

- Linear regret at timestep T.
- Linear regret at timestep $2T$.
- Linear regret at timestep $4T$ in either instance 1 or instance 2.

Moreover, for any $t>2T$ of instance 2, we have $u^{(t)}_{\text{empirical}}(a^2)-u^{(t)}_{\text{empirical}}(a^1)\geq u^{(4T)}_{\text{empirical}}(a^2)-u^{(4T)}_{\text{empirical}}(a^1)\geq0.03$. Therefore, when $\tau\leq\mathcal{O}\left(\frac{1}{\log T}\right)$, with high probability, the action with higher empirical average utility will always be ranked first. □

G Proof of Theorem E.2

In this section, we proved the high probability bound for the estimation error, and with that, we gave the regret upper bound of our algorithm for instantaneous utility. Under the assumption of sublinear variation of the utility function, our algorithm reached sublinear regret.

Theorem E.2 gave the estimation error bound of the utility vector for each timestep.

Theorem E.2. *Consider* **InstUtil Rank** *and Algorithm 1. Suppose each action is proposed with probability at least $p>0$ at each timestep $t\in[T]$ and let $\widetilde{\boldsymbol{u}}^{(t)}=\text{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=t-m'+1}^{t}\right)$. Then, for any $\delta\in(0,1)$ and $t\geq m'$, when $m'p^4\geq2\log\left(\frac{2}{\delta}\right)$, with probability at least $1-\delta$, the estimate $\widetilde{\boldsymbol{u}}^{(t)}$ satisfies,*

$$\left\|\widetilde{\boldsymbol{u}}^{(t)}-\boldsymbol{u}^{(t)}\right\|_\infty\leq\frac{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}{p}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m'}}+\sum_{s=t-m'+1}^{t-1}\left\|\boldsymbol{u}^{(s+1)}-\boldsymbol{u}^{(s)}\right\|_\infty.$$

Proof. For any $j \in [|\mathcal{A}| - 1]$, we assume that the probability for action a^j is chosen at each timestep is at least p. Let the number of action pair $\left(a^j, a^{|\mathcal{A}|}\right)$ chosen in the m' timestep be m_1, by Hoeffding's Inequality, we have that with probability at least $1 - \delta$:

$$m_1 \geq m'p^2 - \sqrt{\frac{m'}{2} \log\left(\frac{2}{\delta}\right)}.$$

For these chosen pairs, the probability that a ranks before $|\mathcal{A}_i|$ is $\frac{\exp\left(\frac{1}{\tau}u^{(t)}(a^j)\right)}{\exp\left(\frac{1}{\tau}u^{(t)}(a^j)\right)+\exp\left(\frac{1}{\tau}u^{(t)}(|\mathcal{A}_i|)\right)} = \frac{\exp\left(\frac{1}{\tau}u^{(t)}(a^j)\right)}{\exp\left(\frac{1}{\tau}u^{(t)}(a^j)\right)+1}$. We define $\mathrm{sig}(x)\colon \mathbb{R} \to \mathbb{R} := \frac{\exp(x)}{\exp(x)+1}$,let $S_{m_1} := \sum_{s=1}^{m'} \mathbb{1}\left(a^j \overset{\sigma^{(s)}}{<} a^{|\mathcal{A}|}\right)$, we then do a projection to $\frac{S_{m_1}}{m_1}$.

Due to the monotonicity of function sig,

$$\widetilde{u}^{(t)}(a^j) = \tau \mathrm{sig}^{-1}\left(\mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\frac{S_{m_1}}{m_1}\right)\right).$$

where $\widetilde{u}^{(t)}(a^j)$ is the estimation of $u^{(t)}(a^j)$. Then also by Hoeffding's Inequality, we have that with probability $1 - \delta$,

$$\left|\frac{S_{m_1}}{m_1} - \frac{1}{m_1}\sum_{s=1}^{m'} \mathbb{1}\left(a^j \overset{\sigma^{(s)}}{<} a^{|\mathcal{A}|}\right) \mathrm{sig}\left(\frac{1}{\tau}u^{(s)}(a^j)\right)\right| \leq \sqrt{\frac{1}{2m_1}\log\left(\frac{2}{\delta}\right)}.$$

Let $u^{(t),*}(a^j) \in \mathbb{R}$ be the scalar satisfying

$$\mathrm{sig}\left(\frac{1}{\tau}u^{(t),*}(a^j)\right) = \frac{1}{m_1}\sum_{s=1}^{m'} \mathbb{1}\left(a^j \overset{\sigma^{(s)}}{<} a^{|\mathcal{A}|}\right) \cdot \mathrm{sig}\left(\frac{1}{\tau}u^{(s)}(a^j)\right).$$

Since the logistic function is monotone and continuous, $u^{(t),*}(a^j)$ is unique and must exist. Then since $[-1, 1]$ is a convex set, with probability at least $1 - \delta$,

$$\begin{aligned}&\left|\mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\mathrm{sig}\left(\widetilde{u}^{(t)}(a^j)\right)\right) - \mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\mathrm{sig}\left(u^{(t),*}(a^j)\right)\right)\right| \\ \leq &\left|\mathrm{sig}\left(\widetilde{u}^{(t)}(a^j)\right) - \mathrm{sig}\left(u^{(t),*}(a^j)\right)\right| \leq \sqrt{\frac{1}{2m_1}\log\left(\frac{2}{\delta}\right)}.\end{aligned}$$

Lemma G.1 in the following shows that $\widetilde{u}^{(t)}(a^j)$ is bounded between the minimum and maximum of $\left\{u^{(s)}(a^j)\right\}_{s=1}^{m'}$. Then, by further utilizing the assumption that the variation of the utility vectors is small, we can bound the distance between our estimated utility vector and $\boldsymbol{u}^{(t)}$.

Lemma G.1. *Let* $x_1, \ldots, x_n \in [-1, 1]$, $\mathrm{sig}(x) := \frac{\exp(x)}{\exp(x)+1}$. *Let* $\mathrm{sig}_{\mathrm{avg}} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{sig}(x_i)$, *we have*

$$\min_{i\in[n]} x_i \leq \log\left(\frac{\mathrm{sig}_{\mathrm{avg}}}{1 - \mathrm{sig}_{\mathrm{avg}}}\right) \leq \max_{i\in[n]} x_i.$$

The proof is postponed to Appendix G.1.

By Lemma G.1, we have that $u^{(t),*} \in [-1, 1]$, so

$$\begin{aligned}\mathrm{sig}\left(u^{(t),*}(a^j)\right) &\in \left[\mathrm{sig}\left(-\frac{1}{\tau}\right), \mathrm{sig}\left(\frac{1}{\tau}\right)\right], \\ \mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\mathrm{sig}\left(u^{(t),*}(a^j)\right)\right) &= \mathrm{sig}\left(u^{(t),*}(a^j)\right).\end{aligned}$$

For any $u \in [-1, 1]$, we have

$$\begin{aligned}\frac{\mathrm{d}\mathrm{sig}\left(\frac{1}{\tau}u\right)}{\mathrm{d}u} =& \frac{1}{\tau}\mathrm{sig}\left(\frac{u}{\tau}\right)\left(1-\mathrm{sig}\left(-\frac{u}{\tau}\right)\right)\\ \geq& \frac{1}{\tau}\mathrm{sig}\left(-\frac{1}{\tau}\right)\left(1-\mathrm{sig}\left(\frac{1}{\tau}\right)\right)\\ =& \frac{1}{\tau}\left(\mathrm{sig}\left(-\frac{1}{\tau}\right)\right)^2 = \frac{1}{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}.\end{aligned}$$

Recall that with probability at least $1-\delta$,

$$\left|\mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\mathrm{sig}\left(\widetilde{u}^{(t)}(a^j)\right)\right) - \mathrm{sig}\left(u^{(t),*}(a^j)\right)\right| \leq \sqrt{\frac{1}{2m_1}\log\left(\frac{2}{\delta}\right)}.$$

Since function sig is monotonic, by Taylor expansion and the fact that $\widetilde{u}^{(t)}(a^j), u^{(t),*}(a^j) \in [-1, 1]$, we get that with probability at least $1-\delta$,

$$\left|\widetilde{u}^{(t)}(a^j) - u^{(t),*}(a^j)\right| \leq \tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{1}{2m_1}\log\left(\frac{2}{\delta}\right)}.$$

By Lemma G.1, we have

$$u^{(t),*}(a^j) \in \left[\min\left\{u^{(s)}(a^j)\right\}_{s=1}^{m'}, \max\left\{u^{(s)}(a^j)\right\}_{s=1}^{m'}\right],$$

which implies that

$$\begin{aligned}\left|\widetilde{u}^{(t)}(a^j) - u^{(t)}(a^j)\right| \leq& \tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{1}{2m_1}\log\left(\frac{2}{\delta}\right)} + \max_{s\in\{1,2,\ldots,m'-1\}}\left|u^{(s)}(a^j) - u^{(t)}(a^j)\right|\\ \leq& \tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{1}{2m_1}\log\left(\frac{2}{\delta}\right)} + \sum_{s=1}^{m'}\left|u^{(s+1)}(a^j) - u^{(s)}(a^j)\right|.\end{aligned}$$

When $m'p^4 \geq 2\log\left(\frac{2}{\delta}\right)$, with probability at least $1-\delta$,

$$m_1 \geq \frac{m'}{2}p^2.$$

So we have that with a probability at least $1-\delta$,

$$\left|\widetilde{u}^{(t)}(a^j) - u^{(t)}(a^j)\right| \leq \tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{1}{m'p^2}\log\left(\frac{2}{\delta}\right)} + \sum_{s=1}^{m'-1}\left|u^{s+1}(a^j) - u^{(s)}(a^j)\right|.$$

After estimating the utility of each action, we have

$$\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|_\infty \leq \frac{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}{p}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m'}} + \sum_{s=1}^{m'-1}\left\|\boldsymbol{u}^{(s+1)} - \boldsymbol{u}^{(s)}\right\|_\infty. \qquad \square$$

g the utility of eachTherefore, g the utility of eachTherefore,

Remark G.2. *Due to the monotonicity of function* sig*, the following two projections on* $\mathrm{sig}\left(\frac{x}{\tau}\right)$ *are equivalent:*

$$\begin{aligned}\mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\mathrm{sig}\left(\frac{x}{\tau}\right)\right) :=& \min\left(\max\left(\mathrm{sig}\left(\frac{x}{\tau}\right), \mathrm{sig}\left(-\frac{1}{\tau}\right)\right), \mathrm{sig}\left(\frac{1}{\tau}\right)\right),\\ \mathrm{sig}\left(\frac{\mathrm{Proj}_{[-1,1]}(x)}{\tau}\right) :=& \mathrm{sig}\left(\frac{\min\left(\max\left(x,-1\right),1\right)}{\tau}\right).\end{aligned}$$

G.1 OMITTED PROOFS

Lemma E.1. *Let $\#_{\mathcal{S}}(a) := \sum_{a' \in \mathcal{S}} \mathbb{1}(a' = a)$ represent the number of elements in a multiset $\mathcal{S}$ that are equal to $a \in \mathcal{A}$. For any utility vector $\boldsymbol{u}$, temperature $\tau > 0$, a multiset of proposed actions $\mathcal{S}$ with cardinality $|\mathcal{S}| = K$, and any two actions $a \neq b \in \mathcal{S}$, we have*

$$\frac{1}{\#_{\mathcal{S}}(a) \cdot \#_{\mathcal{S}}(b)} \mathbb{E}\left[\sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \,\middle|\, \mathcal{S}\right] = \operatorname{sig}\left(\frac{u(a) - u(b)}{\tau}\right).$$

The expectation is taken over the randomness in the (PL) *model.*

Proof. We will abuse the notion $\overset{\sigma}{<}$ from permutations to subsets of actions. When proposing a set of actions $\mathcal{S}$, let $a \overset{\mathcal{S}}{<} b$ denote the event that a is ahead of b in the permutation given by the environment. In pairwise BTL model, the probability that action a ranks before action b is that

$$\mathbb{P}_{\tau,\boldsymbol{u}}\left(a \overset{\{a,b\}}{<} b\right) = \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\exp\left(\frac{1}{\tau}u(a)\right) + \exp\left(\frac{1}{\tau}u(b)\right)}.$$

By definition, let the multiset of the K proposed actions be $\mathcal{S}$. Then, the probability of the K-wise permutation is

$$\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \mid \mathcal{S}) = \prod_{k_1=1}^{K} \frac{\exp\left(\frac{1}{\tau}u(\sigma(k_1))\right)}{\sum_{k_2=k_1}^{K} \exp\left(\frac{1}{\tau}u(\sigma(k_2))\right)}. \tag{G.1}$$

Let $\Sigma(\mathcal{S})$ be the set containing all permutations of $\mathcal{S}$.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \mid \mathcal{S}\right] \\
&= \sum_{\sigma \in \Sigma(\mathcal{S})} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \mid \mathcal{S}) \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \\
&= \sum_{\substack{\sigma \in \Sigma(\mathcal{S}): \\ \sigma(1)=a}} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \mid \mathcal{S}) \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) && \text{(G.2)} \\
&\quad + \sum_{\substack{\sigma \in \Sigma(\mathcal{S}): \\ \sigma(1)=b}} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \mid \mathcal{S}) \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) && \text{(G.3)} \\
&\quad + \sum_{\substack{\sigma \in \Sigma(\mathcal{S}): \\ \sigma(1) \notin \{a,b\}}} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \mid \mathcal{S}) \sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b). && \text{(G.4)}
\end{aligned}$$

Consider (G.2) first.

$$\sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)=a}} \mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma\,|\,\mathcal{S}\right)\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)$$

$$=\sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)=a}} \mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma\,|\,\mathcal{S}\right)\left(\#_{\mathcal{S}}(b)+\sum_{k_1=2}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)\right)$$

$$=\#_{\mathcal{S}}(b)\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=a\,|\,\mathcal{S}\right)+\sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)=a}} \mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma\,|\,\mathcal{S}\right)\sum_{k_1=2}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)$$

$$=\#_{\mathcal{S}}(b)\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=a\,|\,\mathcal{S}\right)$$

$$+\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=a\,|\,\mathcal{S}\right)\sum_{\sigma\in\Sigma(\mathcal{S}\setminus\{a\})}\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma\,|\,\mathcal{S}\setminus\{a\}\right)\sum_{k_1=2}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)$$

$$=\#_{\mathcal{S}}(b)\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=a\,|\,\mathcal{S}\right)+\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=a\,|\,\mathcal{S}\right)\mathbb{E}\left[\sum_{k_1=1}^{K-1}\sum_{k_2=k_1+1}^{K-1}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)\,|\,\mathcal{S}\setminus\{a\}\right].$$

Similarly, for (G.3), we have

$$\sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)=b}} \mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma\,|\,\mathcal{S}\right)\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)$$

$$=\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=b\,|\,\mathcal{S}\right)\mathbb{E}\left[\sum_{k_1=1}^{K-1}\sum_{k_2=k_1+1}^{K-1}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)\,|\,\mathcal{S}\setminus\{b\}\right].$$

Let $\mathrm{Unique}\left(\mathcal{S}\right)$ be the set of non-repeated elements in $\mathcal{S}$. Then, (G.4) can be written as,

$$\sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)\notin\{a,b\}}} \mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma\,|\,\mathcal{S}\right)\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)$$

$$=\sum_{\substack{c\in\mathrm{Unique}(\mathcal{S}):\\ c\notin\{a,b\}}} \mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma(1)=c\,|\,\mathcal{S}\right)\mathbb{E}\left[\sum_{k_1=1}^{K-1}\sum_{k_2=k_1+1}^{K-1}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)\,|\,\mathcal{S}\setminus\{c\}\right].$$

Let $\mathrm{sig}(x)\colon\mathbb{R}\to\mathbb{R}:=\frac{\exp(x)}{\exp(x)+1}$ be the logistic function. Next, we will use induction to show that for any actions $a\neq b\in\mathcal{S}$, the following holds.

$$\mathbb{E}\left[\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)\,|\,\mathcal{S}\right]=\#_{\mathcal{S}}(a)\#_{\mathcal{S}}(b)\mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right). \quad \text{(G.5)}$$

Base case. When $\#_{\mathcal{S}}(a)=0$ or $\#_{\mathcal{S}}(b)=0$, (G.5) trivially holds. When $|\mathcal{S}|=2$ and $\#_{\mathcal{S}}(a)=\#_{\mathcal{S}}(b)=1$,

$$\begin{aligned}\mathbb{E}\left[\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)\,|\,\mathcal{S}\right]&=\mathbb{P}_{\tau,\boldsymbol{u}}\left(\sigma=((a,b))\mid\{a,b\}\right)\\&=\mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right)\\&=\#_{\{a,b\}}(a)\#_{\{a,b\}}(b)\mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right).\end{aligned}$$

Induction step. When (G.5) holds for any $\mathcal{S}$ with $|\mathcal{S}| = K$. Then, we will show that it still holds for any $\mathcal{S}$ with $|\mathcal{S}| = K+1$. Firstly, we will introduce Lemma G.3 to quantify the marginal probability $\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S})$.

Lemma G.3. *For any utility vector $\boldsymbol{u}$, temperature $\tau > 0$, and a multiset of actions $\mathcal{S}$, the marginal probability of any action $a \in \mathcal{A}$ ranking at the first place of the permutation can be written as*

$$\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S}) = \#_{\mathcal{S}}(a) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)}.$$

The proof is postponed to the end of this section. By Lemma G.3, (G.2) is equal to

$$\begin{aligned}
&\#_{\mathcal{S}}(b)\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S}) \\
&+ \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S}) \sum_{\sigma \in \Sigma(\mathcal{S}\setminus\{a\})} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \,|\, \mathcal{S} \setminus \{a\}) \sum_{k_1=2}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \\
=&\#_{\mathcal{S}}(a)\#_{\mathcal{S}}(b) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} \\
&+ \#_{\mathcal{S}}(a) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} \left(\#_{\mathcal{S}\setminus\{a\}}(a) \cdot \#_{\mathcal{S}\setminus\{a\}}(b) \mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right)\right) \\
=&\#_{\mathcal{S}}(a)\#_{\mathcal{S}}(b) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} + \#_{\mathcal{S}}(a) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} (\#_{\mathcal{S}}(a) - 1) \cdot \#_{\mathcal{S}}(b) \mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right).
\end{aligned}$$

Similarly, (G.3) is equal to

$$\#_{\mathcal{S}}(b) \frac{\exp\left(\frac{1}{\tau}u(b)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} \#_{\mathcal{S}}(a) \cdot (\#_{\mathcal{S}}(b) - 1) \,\mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right),$$

and (G.4) is equal to

$$\left(1 - \frac{\#_{\mathcal{S}}(a)\exp\left(\frac{1}{\tau}u(a)\right) + \#_{\mathcal{S}}(b)\exp\left(\frac{1}{\tau}u(b)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)}\right) \#_{\mathcal{S}}(a) \cdot \#_{\mathcal{S}}(b) \mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right).$$

Lastly, by summing them up, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \,|\, \mathcal{S}\right] \\
=&\#_{\mathcal{S}}(a)\#_{\mathcal{S}}(b) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} - \#_{\mathcal{S}}(a) \frac{\exp\left(\frac{1}{\tau}u(a)\right) + \exp\left(\frac{1}{\tau}u(b)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)} \cdot \#_{\mathcal{S}}(b) \mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right) \\
&+ \#_{\mathcal{S}}(a) \cdot \#_{\mathcal{S}}(b) \mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right).
\end{aligned}$$

Note that $\mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right) = \frac{\exp\left(\frac{u(a)-u(b)}{\tau}\right)}{\exp\left(\frac{u(a)-u(b)}{\tau}\right)+1} = \frac{\exp\left(\frac{u(a)}{\tau}\right)}{\exp\left(\frac{u(a)}{\tau}\right)+\exp\left(\frac{u(b)}{\tau}\right)}$. Therefore,

$$\mathbb{E}\left[\sum_{k_1=1}^{K} \sum_{k_2=k_1+1}^{K} \mathbb{1}(\sigma(k_1) = a) \cdot \mathbb{1}(\sigma(k_2) = b) \,|\, \mathcal{S}\right] = \#_{\mathcal{S}}(a) \cdot \#_{\mathcal{S}}(b) \mathrm{sig}\left(\frac{u(a)-u(b)}{\tau}\right),$$

and we complete the induction. □

Lemma G.3. *For any utility vector $\boldsymbol{u}$, temperature $\tau > 0$, and a multiset of actions $\mathcal{S}$, the marginal probability of any action $a \in \mathcal{A}$ ranking at the first place of the permutation can be written as*

$$\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S}) = \#_{\mathcal{S}}(a) \frac{\exp\left(\frac{1}{\tau}u(a)\right)}{\sum_{a' \in \mathcal{S}} \exp\left(\frac{1}{\tau}u(a')\right)}.$$

Proof. Let $\Sigma(\mathcal{S})$ be the set containing all permutations of $\mathcal{S}$. By definition, for any action $a \in \mathcal{A}$,

$$\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S}) = \sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)=a}} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \,|\, \mathcal{S}) = \sum_{\substack{\sigma\in\Sigma(\mathcal{S}):\\ \sigma(1)=a}} \prod_{k_1=1}^{|\mathcal{S}|} \frac{\exp\left(\frac{1}{\tau}u\left(\sigma\left(k_1\right)\right)\right)}{\sum_{k_2=k_1}^{|\mathcal{S}|}\exp\left(\frac{1}{\tau}u\left(\sigma\left(k_2\right)\right)\right)}.$$

Since there are $\#_{\mathcal{S}}(a)$ action a in $\mathcal{S}$, by rearranging the terms, we have

$$\begin{aligned}
\mathbb{P}_{\tau,\boldsymbol{u}}(\sigma(1) = a \,|\, \mathcal{S}) =& \#_{\mathcal{S}}(a)\frac{\exp\left(\frac{1}{\tau}u\left(a\right)\right)}{\sum_{a'\in\mathcal{S}}\exp\left(\frac{1}{\tau}u\left(a'\right)\right)} \sum_{\sigma\in\Sigma(\mathcal{S}\setminus\{a\})} \prod_{k_1=1}^{|\mathcal{S}|-1} \frac{\exp\left(\frac{1}{\tau}u\left(\sigma\left(k_1\right)\right)\right)}{\sum_{k_2=k_1}^{|\mathcal{S}|-1}\exp\left(\frac{1}{\tau}u\left(\sigma\left(k_2\right)\right)\right)} \\
=& \#_{\mathcal{S}}(a)\frac{\exp\left(\frac{1}{\tau}u\left(a\right)\right)}{\sum_{a'\in\mathcal{S}}\exp\left(\frac{1}{\tau}u\left(a'\right)\right)} \sum_{\sigma\in\Sigma(\mathcal{S}\setminus\{a\})} \mathbb{P}_{\tau,\boldsymbol{u}}(\sigma \,|\, \mathcal{S}\setminus\{a\}) \\
=& \#_{\mathcal{S}}(a)\frac{\exp\left(\frac{1}{\tau}u\left(a\right)\right)}{\sum_{a'\in\mathcal{S}}\exp\left(\frac{1}{\tau}u\left(a'\right)\right)}.
\end{aligned}$$

□

Lemma G.1. *Let* $x_1, \ldots, x_n \in [-1, 1], \mathrm{sig}(x) := \frac{\exp(x)}{\exp(x)+1}$. *Let* $\mathrm{sig}_{\mathrm{avg}} := \frac{1}{n}\sum_{i=1}^{n}\mathrm{sig}(x_i)$, *we have*

$$\min_{i\in[n]} x_i \le \log\left(\frac{\mathrm{sig}_{\mathrm{avg}}}{1-\mathrm{sig}_{\mathrm{avg}}}\right) \le \max_{i\in[n]} x_i.$$

Proof. The logistic function $\mathrm{sig}(x)$ is increasing monotonically with respect to x, since $\frac{\mathrm{dsig}}{\mathrm{d}x} = \frac{\exp(x)}{(\exp(x)+1)^2} > 0$. Then, without loss of generality, let $x_1 \le x_2 \le \cdots \le x_n$. Then $\mathrm{sig}(x_1) \le \mathrm{sig}_{\mathrm{avg}} \le \mathrm{sig}(x_n)$.

Since $\mathrm{sig}(x)$ is monotonic and continuous, there exist only one $\zeta \in [x_1, x_n]$ such that $\mathrm{sig}(\zeta) = \mathrm{sig}_{\mathrm{avg}}$. As the inverse function of $\mathrm{sig}(x)$ is $\log\left(\frac{\mathrm{sig}(y)}{1-\mathrm{sig}(y)}\right)$, we have

$$\min_{i\in[n]} x_i \le \log\left(\frac{\mathrm{sig}_{\mathrm{avg}}}{1-\mathrm{sig}_{\mathrm{avg}}}\right) \le \max_{i\in[n]} x_i.$$

□

H Proof of Theorem 4.2

In this section, we prove the regret upper bound for our instantaneous reward algorithm under the full-information feedback.

Theorem H.1 (Formal version of Theorem 4.2)**.** *Consider Algorithm 2 and full-information feedback. For any* $\delta \in (0, 1)$, $T > 0$, *and any no-regret learning algorithm* Alg, *with probability at least* $(1-\delta)$, *by choosing* $m = \left(\frac{T}{P^{(T)}}\right)^{\frac{2}{3}}\left(\log\left(\frac{2T}{\delta}\right)\right)^{\frac{1}{3}}$, $R^{(T),\mathrm{external}}$ *satisfies*

$$\begin{aligned}
R^{(T),\mathrm{external}} \le& R^{(T),\mathrm{external}}\left(\mathrm{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) + \sqrt{2}\tau\left(e^{\frac{1}{\tau}}+1\right)^2\left(P^{(t)}\right)^{\frac{1}{3}}T^{\frac{2}{3}}\left(\log\left(\frac{2T}{\delta}\right)\right)^{\frac{1}{3}} \\
&+ 2\left(P^{(t)}\right)^{\frac{1}{3}}T^{\frac{2}{3}}\left(\log\left(\frac{2T}{\delta}\right)\right)^{\frac{1}{3}} + 2\left(P^{(t)}\right)^{-\frac{2}{3}}T^{\frac{2}{3}}\left(\log\left(\frac{2T}{\delta}\right)\right)^{\frac{1}{3}}. \qquad \text{(H.1)}
\end{aligned}$$

Proof. By Theorem E.2, we have

$$\begin{aligned}
\left|R^{(T),\mathrm{external}} - R^{(T),\mathrm{external}}\left(\mathrm{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right| =& \left|\max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \pi^{(t)}\right\rangle - \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \widetilde{\boldsymbol{u}}^{(t)}, \widehat{\pi} - \pi^{(t)}\right\rangle\right| \\
\le& \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\left|\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)} - \widetilde{\boldsymbol{u}}^{(t)}, \widehat{\pi} - \pi^{(t)}\right\rangle\right| \\
\le& \sum_{t=1}^{T}\left\|\boldsymbol{u}^{(t)} - \widetilde{\boldsymbol{u}}^{(t)}\right\|_{\infty} \cdot \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\left\|\widehat{\pi} - \pi^{(t)}\right\|_1.
\end{aligned}$$

When $t \geq m$, the estimation error between $\widetilde{\boldsymbol{u}}^{(t)}$ and $\boldsymbol{u}^{(t)}$ is given out by Theorem E.2. When $t < m$, it's trivial that the estimation error satisfies

$$\left\|\boldsymbol{u}^{(t)} - \widetilde{\boldsymbol{u}}^{(t)}\right\|_\infty \leq 2.$$

For any given δ, we require the utility estimation bound to hold with probability at least $1 - \frac{\delta}{T}$, then by union bound, with probability at least $1 - \delta$,

$$\left|R^{(T),\text{external}} - R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right| \leq 2\tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{2m_1}T} + 2m\left(P^{(T)}+1\right).$$

By choosing $m = \left(\frac{T}{P^{(T)}}\right)^{\frac{2}{3}}\left(\log\left(\frac{2T}{\delta}\right)\right)^{\frac{1}{3}}$, we conclude the proof. □

I Proof of Theorem 4.3

In this section, we prove the regret upper bound for our instantaneous reward algorithm under the bandit feedback.

Theorem I.1 (Formal version of Theorem 4.3)**.** *Consider Algorithm 2 and bandit feedback. For any $\delta \in (0,1)$, $T > 0$, and any no-regret learning algorithm* Alg*, with probability at least $(1-\delta)$, by choosing $\gamma = \left(\frac{P^{(T)}}{T}\right)^{\frac{1}{5}}$, $m = \frac{32|\mathcal{A}|^4}{K^4}\left(\frac{T}{P^{(T)}}\right)^{\frac{4}{5}}\log\left(\frac{2T}{\delta}\right)$, $R^{(T)}$ satisfies*

$$\begin{aligned} R^{(T)} \leq & R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) + \sqrt{2T\log\left(\frac{1}{\delta}\right)} + \left(\frac{\tau K\left(e^{\frac{1}{\tau}}+1\right)^2}{|\mathcal{A}|}+1\right)\left(P^{(T)}\right)^{\frac{1}{5}}T^{\frac{4}{5}} \\ & + \frac{64|\mathcal{A}|^4}{K^4}\left(P^{(T)}\right)^{\frac{1}{5}}T^{\frac{4}{5}}\log\left(\frac{2T}{\delta}\right) + \frac{64|\mathcal{A}|^4}{K^4}\left(\frac{T}{P^{(T)}}\right)^{\frac{4}{5}}\log\left(\frac{2T}{\delta}\right). \end{aligned} \tag{I.1}$$

Proof. Firstly, we define:

$$\begin{aligned} R^{(T)} &:= \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi}\right\rangle - \frac{1}{K}\sum_{j=1}^{K}\boldsymbol{u}^{(t)}\left(\sigma^{(t)}(j)\right)\right) \\ \widetilde{R}^{(T)} &:= \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \widetilde{\boldsymbol{u}}^{(t)}, \widehat{\pi} - \pi^{(t)}\right\rangle. \end{aligned}$$

Then,

$$\begin{aligned} R^{(T)} \leq & \underbrace{\left|R^{(T)} - R^{(T),\text{external}}\right|}_{\heartsuit} + \underbrace{\left|R^{(T),\text{external}} - \widetilde{R}^{(T)}\right|}_{\spadesuit} + \underbrace{\left|\widetilde{R}^{(T)} - R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right|}_{\diamondsuit} \\ & + \underbrace{R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)}_{\clubsuit}. \end{aligned}$$

Note that ♠ can be bounded by bounding $\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|_\infty$ as in Appendix H. ♣ is sublinear by definition of Alg. Next, we will introduce lemmas that bound ♥, ♦ individually. The proofs are postponed to Appendices I.1 and I.2.

Lemma I.2 (♥)**.** *For any $T > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\left|R^{(T)} - R^{(T),\text{external}}\right| \leq \sqrt{2T\log\left(\frac{1}{\delta}\right)}.$$

Lemma I.3 (♦). *The difference between* $\widetilde{R}^{(T)}$ *and* $R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)$ *satisfies:*

$$\left|\widetilde{R}^{(T)} - R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right| \leq 2\gamma T.$$

With Lemma I.2 and Lemma I.3, under the condition of Theorem 4.3, by letting Theorem E.2 to hold with probability $1 - \frac{\delta}{T}$ at each timestep, with probability at least $(1-\delta)$, the regret satisfies

$$R^{(T)} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) + \sqrt{2T\log\left(\frac{1}{\delta}\right)} + 2\frac{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}{p}\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{m}}T + 2m\left(P^{(T)}+1\right) + \gamma T.$$

In this case, each action $a \in \mathcal{A}$ is chosen with probability at least p that satisfies

$$\begin{aligned} p \geq & 1 - \left(1 - \frac{\gamma}{|\mathcal{A}|}\right)^K \geq 1 - \exp\left(-K\frac{\gamma}{|\mathcal{A}|}\right) \\ \geq & 1 - \left(1 - K\frac{\gamma}{|\mathcal{A}|} + \frac{1}{2}\left(K\frac{\gamma}{|\mathcal{A}|}\right)^2\right) = K\frac{\gamma}{|\mathcal{A}|} - \frac{1}{2}\left(K\frac{\gamma}{|\mathcal{A}|}\right)^2. \end{aligned}$$

Since $K\frac{\gamma}{|\mathcal{A}|} \leq 1$,

$$\frac{1}{2}K\frac{\gamma}{|\mathcal{A}|} \geq \frac{1}{2}\left(K\frac{\gamma}{|\mathcal{A}|}\right)^2 \quad \Rightarrow \quad p \geq \frac{K\gamma}{2\,|\mathcal{A}|}.$$

By letting $\gamma = \left(\frac{P^{(T)}}{T}\right)^{\frac{1}{5}}, m = \frac{32|\mathcal{A}|^4}{K^4}\left(\frac{T}{P^{(T)}}\right)^{\frac{4}{5}}\log\left(\frac{2T}{\delta}\right)$, we have

$$R^{(T)} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) + \mathcal{O}\left(T^{\frac{4}{5}}\left(P^{(T)}\right)^{\frac{1}{5}}\log\left(\frac{T}{\delta}\right)\right).$$

The condition $m \geq \frac{2\log\left(\frac{2T}{\delta}\right)}{p^4}$ is also satisfied since

$$mp^4 \geq m\frac{K^4\gamma^4}{16|\mathcal{A}|^4} = 2\log\left(\frac{2T}{\delta}\right). \qquad \square$$

I.1 Bounding ♥ : $\left|R^{(T)} - R^{(T),\text{external}}\right|$

We will show that $\left|R^{(T)} - R^{(T),\text{external}}\right|$ is sublinear by using a standard concentration bound.

Lemma I.2 (♥). *For any* $T > 0$ *and* $\delta \in (0,1)$*, with probability at least* $1-\delta$*:*

$$\left|R^{(T)} - R^{(T),\text{external}}\right| \leq \sqrt{2T\log\left(\frac{1}{\delta}\right)}.$$

Proof. Let

$$d^{(t)} := \frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a) - \left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle.$$

By definition, each element of $o^{(t)}$ is sampled *i.i.d.* from $\pi^{(t)}$ and the update-rule of $\pi^{(t)}$ is deterministic, $\mathbb{E}\left[d^{(t)} \mid \left\{\sigma^{(s)}\right\}_{s=1}^{t-1}\right] = 0$, so that $\left\{d^{(t)}\right\}$ is a martingale difference sequence.

We also have bound for $\left|\frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a)\right| \leq 1$, $\left|\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right| \leq 1$, so

$$\left|d^{(t)}\right| = \left|\frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a) - \left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right| \leq \left|\frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a)\right| + \left|\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right| \leq 2.$$

Furthermore, we have

$$\begin{aligned}\sum_{t=1}^{T} d^{(t)} =& \sum_{t=1}^{T}\left(\frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a)\right) - \sum_{t=1}^{T}\left(\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right)\\ =& \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi}\right\rangle - \left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right) - \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi}\right\rangle - \frac{1}{K}\sum_{j=1}^{K} u^{(t)}\left(\sigma^{(j)}\right)\right)\\ =& R^{(T),\text{external}} - R^{(T)}.\end{aligned}$$

Next, we will introduce Azuma-Hoeffding inequality to finish the concentration bound.

Theorem I.4 (Azuma-Hoeffding inequality)**.** *For any martingale difference sequence $Y_1, \ldots, Y_n$ with $\forall j \in [n], a_j \le Y_j \le b_j$, the following holds for any $w \ge 0$.*

$$\mathbb{P}\left(\sum_{j=1}^{n} Y_j \ge w\right) \le \exp\left(-\frac{2w^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right).$$

Then with Theorem I.4, we have that

$$\mathbb{P}\left(R^{(T),\text{external}} - R^{(T)} \ge \sqrt{2T\log\left(\frac{1}{\delta}\right)}\right) \le \delta.$$

Similarly, we can also prove

$$\mathbb{P}\left(R^{(T)} - R^{(T),\text{external}} \ge \sqrt{2T\log\left(\frac{1}{\delta}\right)}\right) \le \delta.$$

So we get that with probability at least $1-\delta$:

$$\left|R^{(T)} - R^{(T),\text{external}}\right| \le \sqrt{2T\log\left(\frac{1}{\delta}\right)}. \qquad \square$$

I.2 Bounding ♦ : $\left|\widetilde{R}^{(T)} - \widetilde{R}_{\gamma}^{(T)}\right|$

♦ can be bounded by $\mathcal{O}(\gamma T)$, since $\Delta^{\mathcal{A}}$ and $\Delta_{\gamma}^{\mathcal{A}}$ are close to each other.

Lemma I.3 (♦)**.** *The difference between $\widetilde{R}^{(T)}$ and $R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)$ satisfies:*

$$\left|\widetilde{R}^{(T)} - R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right| \le 2\gamma T.$$

Proof. Let $\overline{\pi}^{(t+1)} = \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right)$. Then,

$$\begin{aligned}\left|\widetilde{R}^{(T)} - R^{(T),\text{external}}\left(\text{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right)\right| =& \left|\max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \widetilde{\boldsymbol{u}}^{(s)}, \widehat{\pi} - \pi^{(t)}\right\rangle - \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \widetilde{\boldsymbol{u}}^{(t)}, \widehat{\pi} - \overline{\pi}^{(t)}\right\rangle\right|\\ =& \left|\sum_{t=1}^{T}\left\langle \widetilde{\boldsymbol{u}}^{(t)}, \overline{\pi}^{(t)} - \pi^{(t)}\right\rangle\right|\\ =& \left|\sum_{t=1}^{T}\left\langle \widetilde{\boldsymbol{u}}^{(t)}, \overline{\pi}^{(t)} - \left((1-\gamma)\overline{\pi}^{(t)} + \gamma\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}\right)\right\rangle\right|\\ \le& \gamma\sum_{t=1}^{T}\left\|\widetilde{\boldsymbol{u}}^{(t)}\right\|_{\infty}\cdot\left(\left\|\overline{\pi}^{(t)}\right\|_{1} + \left\|\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}\right\|_{1}\right) \le 2\gamma T. \qquad \square\end{aligned}$$

J PROOF OF THEOREM 5.3

Theorem J.1 (Formal version of Theorem 5.3)**.** *Consider* **AvgUtil Rank** *with full-information feedback and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any no-regret learning algorithm* Alg *satisfying Assumption 5.2, with probability at least $(1-\delta)$, by choosing $m = 2\frac{|\mathcal{A}|^4}{K^4}T^{\frac{2}{3}}\log\left(\frac{2T}{\delta}\right)$, $R^{(T),\text{external}}$ satisfies*

$$R^{(T),\text{external}} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + L\tau K\left(e^{\frac{1}{\tau}}+1\right)^2 L^{\frac{1}{3}}T + 4\frac{|\mathcal{A}|^4}{K^4}L^{-\frac{2}{3}}\log\left(\frac{2T}{\delta}\right) \tag{J.1}$$

$$+ 4\frac{|\mathcal{A}|^9}{K^8}L^{-\frac{1}{3}}\log\left(\frac{2T}{\delta}\right)(\log T + 1) + 4\frac{|\mathcal{A}|^5}{K^4}\log\left(\frac{2T}{\delta}\right)L^{\frac{1}{3}}T.$$

Proof. Let $\overline{\pi}^{(t+1)} = \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right)$, *i.e.,* the strategy generated by Alg when the ground-truth utility vectors are given. Then,

$$\begin{aligned}
\left|R^{(T),\text{external}} - R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right)\right| &= \left|\max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \pi^{(t)}\right\rangle - \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}}\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \overline{\pi}^{(t)}\right\rangle\right| \\
&= \left|\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)}, \overline{\pi}^{(t)} - \pi^{(t)}\right\rangle\right| \\
&\leq \sum_{t=1}^{m-1}\left\|\boldsymbol{u}^{(t)}\right\|_{\infty}\cdot\left\|\overline{\pi}^{(t)} - \pi^{(t)}\right\|_1 + \sum_{t=m}^{T}\left\|\boldsymbol{u}^{(t)}\right\|\cdot\left\|\overline{\pi}^{(t)} - \pi^{(t)}\right\| \\
&\leq 2m + \sum_{t=m}^{T}\left\|\boldsymbol{u}^{(t)}\right\|\cdot\left\|\overline{\pi}^{(t)} - \pi^{(t)}\right\|.
\end{aligned}$$

By Assumption 5.2, when the conditions in Theorem 5.1 are met, for any $t \geq m$, with probability at least $1-\delta$ we have

$$\left\|\overline{\pi}^{(t)} - \pi^{(t)}\right\| \leq Lt\left\|\widetilde{\boldsymbol{u}}_{\text{avg}}^{(t)} - \boldsymbol{u}_{\text{avg}}^{(t)}\right\| \leq Lt\sqrt{|\mathcal{A}|}\left(\tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{2m}} + \sum_{s=t-m+1}^{t-1}\frac{2}{s+1}\right).$$

Then,

$$\begin{aligned}
&\left|R^{(T),\text{external}} - R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right)\right| \\
\leq& 2m + L|\mathcal{A}|\tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{2m}}T^2 + 2L|\mathcal{A}|\sum_{t=m}^{T}t\sum_{s=t-m+1}^{t}\frac{1}{s+1} \\
\leq& 2m + L|\mathcal{A}|\tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{2m}}T^2 + L|\mathcal{A}|\sum_{t=1}^{T}\frac{m(2t+m-1)}{t} \\
\leq& 2m + L|\mathcal{A}|\tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{2m}}T^2 + Lm^2|\mathcal{A}|\sum_{t=1}^{T}\frac{1}{t} + 2Lm|\mathcal{A}|T \\
\leq& L|\mathcal{A}|\tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{\log\left(\frac{2T}{\delta}\right)}{2m}}T^2 + 2m + Lm^2|\mathcal{A}|(\log T + 1) + 2|\mathcal{A}|mLT.
\end{aligned}$$

By choosing $m = 2\frac{|\mathcal{A}|^4}{K^4}T^{\frac{2}{3}}\log\left(\frac{2T}{\delta}\right)$, we have

$$R^{(T),\text{external}} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + \mathcal{O}\left(LT^{\frac{5}{3}}\log\left(\frac{2T}{\delta}\right)\right).$$

Moreover, now $m \geq \frac{2\log\left(\frac{2T}{\delta}\right)}{p^4}$, where $p = \frac{K}{|\mathcal{A}|}$. □

K PROOF OF THEOREM 5.4

Theorem K.1 (Formal version of Theorem 5.4)**.** *Consider* **AvgUtil Rank** *with bandit feedback and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any no-regret learning algorithm* Alg *satisfying Assumption 5.2, with probability at least $(1-\delta)$, by choosing $M = 4T^{\frac{5}{6}} \left(P^{(T)}\right)^{-\frac{1}{2}} |\mathcal{A}|^4 \log\left(\frac{6|\mathcal{A}|T}{\delta}\right)$, $m = 2T^{\frac{2}{3}}|\mathcal{A}|^4 \log\left(\frac{6}{\delta}\right)$, and $\gamma = L^{\frac{1}{3}} T^{\frac{5}{18}} \left(P^{(T)}\right)^{\frac{1}{6}}$, $R^{(T)}$ satisfies*

$$R^{(T)} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + L|\mathcal{A}|TW^{(T)} + 2\gamma\sqrt{|\mathcal{A}|}T + \sqrt{2T\log\left(\frac{3}{\delta}\right)}, \quad \text{(K.1)}$$

where

$$C_\delta := \frac{|\mathcal{A}| \log\left(\frac{3|\mathcal{A}|T}{\delta}\right)}{\gamma}$$

$$W^{(T)} := 4C_\delta \left(\log T + 1\right)\left(\frac{T^2}{M}\frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}} + 16KC_\delta\frac{m}{M}T\right)$$
$$+ M\left(\log T + 1\right)P^{(T)} + 2M\left(\log T + 2\right).$$

Proof. In the first part of the proof, we will bound $\left\|\widetilde{\boldsymbol{u}}^{(t)}_{\text{empirical}} - \boldsymbol{u}^{(t)}_{\text{empirical}}\right\|_\infty$. According to Theorem E.2 and union bound, since each action is proposed with probability at least $\frac{\gamma}{|\mathcal{A}|}$, with probability at least $1-\frac{\delta}{3}$, for any $t \geq m$, we have

$$\left\|\widetilde{\boldsymbol{u}}^{(t)}_{\text{empirical}} - \boldsymbol{u}^{(t)}_{\text{empirical}}\right\|_\infty \leq \frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}} + \sum_{s=t-m+1}^{t-1}\left\|\boldsymbol{u}^{(s+1)}_{\text{empirical}} - \boldsymbol{u}^{(s)}_{\text{empirical}}\right\|_\infty.$$

Let $\#_{o^{(t)}}(a)$ be the number of action $a \in \mathcal{A}$ proposed in $o^{(t)}$. Then, for any $t \in [T-1]$ and $a \in \mathcal{A}$, we have

$$\begin{aligned}
\left|u^{(t+1)}_{\text{empirical}}(a) - u^{(t)}_{\text{empirical}}(a)\right| &= \left|\frac{u^{(t)}_{\text{empirical}}(a)\sum_{s=1}^{t}\#_{o^{(s)}}(a) + u^{(t+1)}(a)\#_{o^{(t+1)}}(a)}{\sum_{s=1}^{t}\#_{o^{(s)}}(a) + \#_{o^{(t+1)}}(a)} - u^{(t)}_{\text{empirical}}(a)\right| \\
&\leq \left|\frac{u^{(t)}_{\text{empirical}}(a)\#_{o^{(t+1)}}(a)}{\sum_{s=1}^{t}\#_{o^{(s)}}(a) + \#_{o^{(t+1)}}(a)}\right| + \left|\frac{u^{(t+1)}(a)\#_{o^{(t+1)}}(a)}{\sum_{s=1}^{t}\#_{o^{(s)}}(a) + \#_{o^{(t+1)}}(a)}\right| \\
&\leq K\left|\frac{u^{(t)}_{\text{empirical}}(a)}{\sum_{s=1}^{t}\#_{o^{(s)}}(a) + \#_{o^{(t+1)}}(a)}\right| + K\left|\frac{u^{(t+1)}(a)}{\sum_{s=1}^{t}\#_{o^{(s)}}(a) + \#_{o^{(t+1)}}(a)}\right|.
\end{aligned}$$

Next, we will show that since each action will be proposed with probability at least $\frac{\gamma}{|\mathcal{A}|}$, with high probability, there is a lowerbound for $\#_{o^{(t+1)}}(a)$.

Lemma K.2. *Consider the case when action is proposed with probability at least $p > 0$ at each timestep. Then, for any $\delta > 0$, any action $a \in \mathcal{A}$, and $T > 0$, with probability at least $1-\delta$, the following holds for any $t \geq \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p}$,*

$$\exists t' \in [T] \quad , t - \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p} \leq t' \leq t \text{ and } a \in o^{(t')}. \quad \text{(K.2)}$$

For notational simplicity, let $C_\delta := \frac{|\mathcal{A}|\log\left(\frac{3|\mathcal{A}|T}{\delta}\right)}{\gamma}$. According to Lemma K.2, with probability at least $1-\frac{\delta}{3}$, for any timestep $t \geq C_\delta$, we have

$$\sum_{s=1}^{t}\#_{o^{(s)}}(a) \geq \left\lfloor\frac{t}{C_\delta}\right\rfloor \geq \frac{t}{2C_\delta}.$$

Therefore, for any $t \in [T-1]$ and $a \in \mathcal{A}$, we have

$$\left|u^{(t+1)}_{\text{empirical}}(a) - u^{(t)}_{\text{empirical}}(a)\right| \leq 4K\frac{C_\delta}{t+1}.$$

It holds for $t < C_\delta - 1$ because in this case $4K\frac{C_\delta}{t+1} \geq 4K \geq 2$ and all utilities are bounded in $[-1, 1]$. Finally, by Theorem E.2 and union bound, with probability at least $1 - \frac{2\delta}{3}$, we have

$$\left\|\widetilde{\boldsymbol{u}}^{(t)}_{\text{empirical}} - \boldsymbol{u}^{(t)}_{\text{empirical}}\right\|_\infty \leq \frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}} + 4KC_\delta \sum_{s=t-m+1}^{t-1}\frac{1}{s+1}.$$

Let $n^{(t)}(a) \coloneqq \sum_{s=1}^{t} \#_{o^{(s)}}(a)$ for any $a \in \mathcal{A}$ as the number of times action a is proposed up to timestep t. For any $a \in \mathcal{A}$ and $t \geq M$, let define

$$u^{(t)}_{\text{avg-est}}(a) \coloneqq \frac{1}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor}\frac{u^{(s\cdot M)}_{\text{empirical}}(a)n^{(s\cdot M)}(a) - u^{((s-1)M)}_{\text{empirical}}(a)n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}$$

$$\widetilde{u}^{(t)}_{\text{avg-est}}(a) \coloneqq \frac{1}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor}\frac{\widetilde{u}^{(s\cdot M)}_{\text{empirical}}(a)n^{(s\cdot M)}(a) - \widetilde{u}^{((s-1)M)}_{\text{empirical}}(a)n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}.$$

For $t < M$, we define $u^{(t)}_{\text{avg-est}}(a) = \widetilde{u}^{(t)}_{\text{avg-est}}(a) = 0$ for any action $a \in \mathcal{A}$. In the rest of the proof, we will bound $\left\|\widetilde{\boldsymbol{u}}^{(t)}_{\text{avg-est}} - \boldsymbol{u}^{(t)}_{\text{avg-est}}\right\|_\infty$ and $\left\|\boldsymbol{u}^{(t)}_{\text{avg-est}} - \boldsymbol{u}^{(t)}_{\text{avg}}\right\|_\infty$ individually.

K.1 $\left\|\widetilde{\mathbf{u}}^{(t)}_{\text{avg-est}} - \mathbf{u}^{(t)}_{\text{avg-est}}\right\|_\infty$ UPPERBOUND

For any $a \in \mathcal{A}$, we have

$$\begin{aligned}\left|\widetilde{u}^{(t)}_{\text{avg-est}}(a) - u^{(t)}_{\text{avg-est}}(a)\right| \leq & \frac{1}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor}\frac{n^{(s\cdot M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}\left|\widetilde{u}^{(s\cdot M)}_{\text{empirical}}(a) - u^{(s\cdot M)}_{\text{empirical}}(a)\right| \\ & + \frac{1}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor}\frac{n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}\left|\widetilde{u}^{((s-1)M)}_{\text{empirical}}(a) - u^{((s-1)M)}_{\text{empirical}}(a)\right|.\end{aligned}$$

According to Lemma K.2, $n^{(s\cdot M)}(a) - n^{((s-1)M)}(a) \geq \frac{M}{2C_\delta}$ when $M \geq C_\delta$. Therefore, when $M \geq C_\delta$, since $n^{(s\cdot M)}(a) \leq s \cdot M$, we have

$$\left|\widetilde{u}^{(t)}_{\text{avg-est}}(a) - u^{(t)}_{\text{avg-est}}(a)\right| \leq \frac{2C_\delta}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor} s\left(\left|\widetilde{u}^{(s\cdot M)}_{\text{empirical}}(a) - u^{(s\cdot M)}_{\text{empirical}}(a)\right| + \left|\widetilde{u}^{((s-1)M)}_{\text{empirical}}(a) - u^{((s-1)M)}_{\text{empirical}}(a)\right|\right).$$

K.2 $\left\|\mathbf{u}^{(t)}_{\text{avg-est}} - \mathbf{u}^{(t)}_{\text{avg}}\right\|_\infty$ UPPERBOUND

For any $a \in \mathcal{A}$,

$$\begin{aligned}&\left|\frac{1}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor}\frac{u^{(s\cdot M)}_{\text{empirical}}(a)n^{(s\cdot M)}(a) - u^{((s-1)M)}_{\text{empirical}}(a)n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)} - u^{(t)}_{\text{avg}}(a)\right| \\ \leq & \underbrace{\left|\frac{1}{\lfloor t/M \rfloor}\sum_{s=1}^{\lfloor t/M \rfloor}\frac{u^{(s\cdot M)}_{\text{empirical}}(a)n^{(s\cdot M)}(a) - u^{((s-1)M)}_{\text{empirical}}(a)n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)} - u^{(M\lfloor t/M \rfloor)}_{\text{avg}}(a)\right|}_{\spadesuit} \\ & + \underbrace{\left|u^{(t)}_{\text{avg}}(a) - u^{(M\lfloor t/M \rfloor)}_{\text{avg}}(a)\right|}_{\clubsuit}.\end{aligned}$$

Note that ♣ can be bounded by

$$\begin{aligned}\clubsuit =& \left| \frac{(M\lfloor t/M \rfloor) u_{\text{avg}}^{(M\lfloor t/M \rfloor)}(a) + \sum_{s=M\lfloor t/M \rfloor+1}^{t} u^{(s)}(a)}{t} - u_{\text{avg}}^{(M\lfloor t/M \rfloor)}(a) \right| \\ \leq& \frac{M}{t} \left| u_{\text{avg}}^{(M\lfloor t/M \rfloor)}(a) \right| + \frac{1}{t} \left| \sum_{s=M\lfloor t/M \rfloor+1}^{t} u^{(s)}(a) \right| \leq \frac{2M}{t}.\end{aligned}$$

For ♠, we have

$$\begin{aligned}\spadesuit =& \left| \frac{1}{\lfloor t/M \rfloor} \sum_{s=1}^{\lfloor t/M \rfloor} \left(\frac{u_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - u_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)} - \frac{1}{M} \sum_{s'=(s-1)M+1}^{s\cdot M} u^{(s')}(a) \right) \right| \\ \leq& \frac{1}{\lfloor t/M \rfloor} \sum_{s=1}^{\lfloor t/M \rfloor} \left| \frac{u_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - u_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)} - \frac{1}{M} \sum_{s'=(s-1)M+1}^{s\cdot M} u^{(s')}(a) \right|.\end{aligned}$$

When $n^{(s\cdot M)}(a) - n^{((s-1)M)}(a) > 0$, both $\frac{u_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - u_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)}$ and $\frac{1}{M} \sum_{s'=(s-1)M+1}^{s\cdot M} u^{(s')}(a)$ are in the convex hull of $\left\{ u^{(s')}(a) \right\}_{s'=(s-1)M+1}^{s\cdot M}$. Therefore,

$$\begin{aligned}& \left| \frac{u_{\text{empirical}}^{(s\cdot M)}(a) n^{(s\cdot M)}(a) - u_{\text{empirical}}^{((s-1)M)}(a) n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a) - n^{((s-1)M)}(a)} - \frac{1}{M} \sum_{s'=(s-1)M+1}^{s\cdot M} u^{(s')}(a) \right| \\ \leq& \max_{(s-1)M+1 \leq s', s'' \leq s\cdot M} \left| u^{(s')}(a) - u^{(s'')}(a) \right| \\ \leq& \sum_{s'=(s-1)M+1}^{s\cdot M-1} \left| u^{(s'+1)}(a) - u^{(s')}(a) \right|.\end{aligned}$$

Therefore, for any $t \geq M$ and $a \in \mathcal{A}$, we have

$$\left| u_{\text{avg-est}}^{(t)}(a) - u_{\text{avg}}^{(t)}(a) \right|_{\infty} \leq \frac{1}{\lfloor t/M \rfloor} \sum_{s=1}^{\lfloor t/M \rfloor} \sum_{s'=(s-1)M+1}^{s\cdot M-1} \left| u^{(s'+1)}(a) - u^{(s')}(a) \right| + \frac{2M}{t}.$$

By combining all pieces together, we have

$$\begin{aligned}& \left| \widetilde{u}_{\text{avg-est}}^{(t)}(a) - u_{\text{avg}}^{(t)}(a) \right| \\ \leq& \frac{2C_\delta}{\lfloor t/M \rfloor} \sum_{s=1}^{\lfloor t/M \rfloor} s \left(\left| \widetilde{u}_{\text{empirical}}^{(s\cdot M)}(a) - u_{\text{empirical}}^{(s\cdot M)}(a) \right| + \left| \widetilde{u}_{\text{empirical}}^{((s-1)M)}(a) - u_{\text{empirical}}^{((s-1)M)}(a) \right| \right) \\ &+ \frac{1}{\lfloor t/M \rfloor} \sum_{s=1}^{\lfloor t/M \rfloor} \sum_{s'=(s-1)M+1}^{s\cdot M-1} \left| u^{(s'+1)}(a) - u^{(s')}(a) \right| + \frac{2M}{t}.\end{aligned}$$

Then,

$$
\begin{aligned}
&\sum_{t=1}^{T}\left|\widetilde{u}_{\text{avg}-\text{est}}^{(t)}(a)-u_{\text{avg}}^{(t)}(a)\right| \\
=&\sum_{t=M}^{T}\left|\widetilde{u}_{\text{avg}-\text{est}}^{(t)}(a)-u_{\text{avg}}^{(t)}(a)\right|+\sum_{t=1}^{M-1}\left|\widetilde{u}_{\text{avg}-\text{est}}^{(t)}(a)-u_{\text{avg}}^{(t)}(a)\right| \\
\leq& 2C_\delta\sum_{s=1}^{\lfloor T/M\rfloor} s\left(\left|\widetilde{u}_{\text{empirical}}^{(s\cdot M)}(a)-u_{\text{empirical}}^{(s\cdot M)}(a)\right|+\left|\widetilde{u}_{\text{empirical}}^{((s-1)M)}(a)-u_{\text{empirical}}^{((s-1)M)}(a)\right|\right)\sum_{s'=1}^{\lfloor T/M\rfloor}\frac{M}{s'} \\
&+\sum_{s=1}^{\lfloor T/M\rfloor}\left(\sum_{s'=(s-1)M+1}^{s\cdot M-1}\left|u^{(s'+1)}(a)-u^{(s')}(a)\right|\right)\cdot\left(\sum_{s'=1}^{\lfloor T/M\rfloor}\frac{M}{s'}\right)+\sum_{t=1}^{T}\frac{2M}{t}+2M \\
\leq& 2C_\delta M\sum_{s=1}^{\lfloor T/M\rfloor} s\left(\left|\widetilde{u}_{\text{empirical}}^{(s\cdot M)}(a)-u_{\text{empirical}}^{(s\cdot M)}(a)\right|+\left|\widetilde{u}_{\text{empirical}}^{((s-1)M)}(a)-u_{\text{empirical}}^{((s-1)M)}(a)\right|\right)\left(\log\left(\lfloor T/M\rfloor\right)+1\right) \\
&+M\sum_{s=1}^{\lfloor T/M\rfloor}\left(\sum_{s'=(s-1)M+1}^{s\cdot M-1}\left|u^{(s'+1)}(a)-u^{(s')}(a)\right|\right)\cdot\left(\log\left(\lfloor T/M\rfloor\right)+1\right)+2M\left(\log T+1\right)+2M.
\end{aligned}
$$

When $s=1$, $s\left|\widetilde{u}_{\text{empirical}}^{((s-1)M)}(a)-u_{\text{empirical}}^{((s-1)M)}(a)\right|=0$ by definition. When $s>1$, since $M\geq 2m$, we have

$$
\frac{s}{(s-1)M-m+2}\leq\frac{s}{(s-1)M/2+2}\leq\frac{s}{(s-1)M/2}\leq\frac{4s}{s\cdot M}=\frac{4}{M}.
$$

Hence,

$$
\begin{aligned}
s\left|\widetilde{u}_{\text{empirical}}^{((s-1)M)}(a)-u_{\text{empirical}}^{((s-1)M)}(a)\right| \leq& s\frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}}+4KC_\delta\sum_{s'=(s-1)M-m+1}^{(s-1)M-1}\frac{s}{s'+1} \\
\leq& \frac{T}{M}\frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}}+16KC_\delta\frac{m}{M}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum_{t=1}^{T}\left|\widetilde{u}_{\text{avg}-\text{est}}^{(t)}(a)-u_{\text{avg}}^{(t)}(a)\right| \\
\leq& 4C_\delta M\cdot\lfloor T/M\rfloor\left(\log T+1\right)\left(\frac{T}{M}\frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}}+16KC_\delta\frac{m}{M}\right) \\
&+M\left(\log T+1\right)P^{(T)}+2M\left(\log T+1\right) \\
\leq& 4C_\delta\left(\log T+1\right)\left(\frac{T^2}{M}\frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}}+16KC_\delta\frac{m}{M}T\right) \\
&+M\left(\log T+1\right)P^{(T)}+2M\left(\log T+2\right).
\end{aligned}
$$

Lastly, similar to the proof in Appendix J, let $\overline{\pi}^{(t+1)} = \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right)$. Then, we have

$$\begin{aligned}
&\left|R^{(T),\text{external}} - R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right)\right| \\
=&\left|\sum_{t=1}^{T}\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)} - \overline{\pi}^{(t)}\right\rangle\right| \\
\leq&\sum_{t=1}^{T}\left\|\boldsymbol{u}^{(t)}\right\| \cdot \left\|\pi^{(t)} - \overline{\pi}^{(t)}\right\| \\
\leq&\sqrt{|\mathcal{A}|}\sum_{t=1}^{T}\left\|(1-\gamma)\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)}\right)_{s=1}^{t}\right) + \gamma\frac{\mathbf{1}\left(\mathcal{A}\right)}{|\mathcal{A}|} - \overline{\pi}^{(t)}\right\| \\
\leq&(1-\gamma)\sqrt{|\mathcal{A}|}\sum_{t=1}^{T}\left\|\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)}\right)_{s=1}^{t}\right) - \overline{\pi}^{(t)}\right\| + \gamma\sqrt{|\mathcal{A}|}\sum_{t=1}^{T}\left\|\frac{\mathbf{1}\left(\mathcal{A}\right)}{|\mathcal{A}|} - \overline{\pi}^{(t)}\right\| \\
\leq&\sqrt{|\mathcal{A}|}\sum_{t=1}^{T}\left\|\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)}\right)_{s=1}^{t}\right) - \overline{\pi}^{(t)}\right\| + 2\gamma\sqrt{|\mathcal{A}|}T.
\end{aligned}$$

Further, by Assumption 5.2, we have

$$\begin{aligned}
\sqrt{|\mathcal{A}|}\sum_{t=1}^{T}\left\|\text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)}\right)_{s=1}^{t}\right) - \overline{\pi}^{(t)}\right\| \leq& L\sqrt{|\mathcal{A}|}\sum_{t=1}^{T} t\left\|\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} - \boldsymbol{u}_{\text{avg}}^{(t)}\right\| \\
\leq& L|\mathcal{A}|T\sum_{t=1}^{T}\left\|\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} - \boldsymbol{u}_{\text{avg}}^{(t)}\right\|_{\infty}.
\end{aligned}$$

By combining all the pieces together, we have

$$\begin{aligned}
R^{(T),\text{external}} \leq& R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) \\
&+ \widetilde{\mathcal{O}}\left(\frac{\left(\log\left(\frac{1}{\delta}\right)\right)^{\frac{3}{2}}}{\gamma^2}\frac{LT^3}{M\sqrt{m}} + \frac{m\left(\log\left(\frac{1}{\delta}\right)\right)^2}{\gamma^2 M}LT^2 + LMP^{(T)}T + 2\gamma T\right),
\end{aligned}$$

where $\widetilde{\mathcal{O}}$ hides all the $\log T$ terms.

Let $M = 4T^{\frac{5}{6}}\left(P^{(T)}\right)^{-\frac{1}{2}}|\mathcal{A}|^4\log\left(\frac{6|\mathcal{A}|T}{\delta}\right)$, $m = 2T^{\frac{2}{3}}|\mathcal{A}|^4\log\left(\frac{6}{\delta}\right)$, and $\gamma = L^{\frac{1}{3}}T^{\frac{5}{18}}\left(P^{(T)}\right)^{\frac{1}{6}}$, we have

$$R^{(T),\text{external}} \leq R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + \mathcal{O}\left(\left(\log\left(\frac{1}{\delta}\right)\right)^2 L^{\frac{1}{3}}T^{\frac{23}{18}}\left(P^{(T)}\right)^{\frac{1}{6}}\right).$$

Easy to verify that $M \geq \max\{C_\delta, 2m\}$ and $m \geq \frac{2\log\left(\frac{6}{\delta}\right)}{\gamma^4}|\mathcal{A}|^4$.

Lastly, by Lemma I.2, with probability at least $1 - \frac{\delta}{3}$, we have

$$R^{(T)} \leq R^{(T),\text{external}} + \sqrt{2T\log\left(\frac{3}{\delta}\right)}.$$

By a union bound, we complete the proof. □

Lemma K.2. *Consider the case when action is proposed with probability at least $p > 0$ at each timestep. Then, for any $\delta > 0$, any action $a \in \mathcal{A}$, and $T > 0$, with probability at least $1 - \delta$, the following holds for any $t \geq \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p}$,*

$$\exists t' \in [T] \quad , t - \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p} \leq t' \leq t \text{ and } a \in o^{(t')}. \tag{K.2}$$

Proof. For any $t \geq \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p}$ and action $a \in \mathcal{A}$, the probability of (K.2) does not hold is at most

$$(1-p)^{\frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p}} \leq \exp\left(-\log\left(\frac{|\mathcal{A}|T}{\delta}\right)\right) = \frac{\delta}{|\mathcal{A}|T}.$$

Therefore, by union bound, with probability $1-\delta$, (K.2) holds for any $t \in [T]$ and any action $a \in \mathcal{A}$. □

L Proof of Section 6

Lemma 6.1. *For any $T > 0$ and sequence of strategy profiles $\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(T)}\right)$, the variation of utility vectors of any player $i \in [N]$ satisfies that*

$$\sum_{t=2}^{T} \left\| \boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)} \right\| \leq \max_{j \in [N]} \sqrt{|\mathcal{A}_j|} \prod_{j'=1}^{N} |\mathcal{A}_{j'}| \sum_{t=2}^{T} \sum_{j=1}^{N} \left\| \pi_j^{(t)} - \pi_j^{(t-1)} \right\|. \tag{6.2}$$

Proof. For any timestep t, player $i \in [N]$, and joint action $\boldsymbol{a}_{-i} \in \bigtimes_{j \neq i} \mathcal{A}_j$, let $\pi_{-i}^{(t)}(\boldsymbol{a}_{-i}) := \prod_{j \neq i} \pi_j^{(t)}(a_j)$.

Then, for any timestep t, player $i \in [N]$, and action $a_i \in \mathcal{A}_i$, we have

$$\begin{aligned}
\left| u_i^{(t)}(a_i) - \boldsymbol{u}_i^{(t-1)}(a_i) \right| &\leq \left| \sum_{\boldsymbol{a}' \in \bigtimes_{j=1}^{N} \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}') \, \mathbb{1}\left(a_i' = a_i\right) \left(\pi_{-i}^{(t)}(\boldsymbol{a}'_{-i}) - \pi_{-i}^{(t-1)}(\boldsymbol{a}'_{-i}) \right) \right| \\
&= \left| \left\langle \left(\mathcal{U}_i(a_i, \boldsymbol{a}'_{-i}) \right)_{\boldsymbol{a}'_{-i} \in \bigtimes_{j \neq i} \mathcal{A}_j}, \pi_{-i}^{(t)} - \pi_{-i}^{(t-1)} \right\rangle \right| \\
&\leq \left\| \left(\mathcal{U}_i(a_i, \boldsymbol{a}'_{-i}) \right)_{\boldsymbol{a}'_{-i} \in \bigtimes_{j \neq i} \mathcal{A}_j} \right\|_\infty \cdot \left\| \pi_{-i}^{(t)} - \pi_{-i}^{(t-1)} \right\|_1 \\
&\leq \left\| \pi_{-i}^{(t)} - \pi_{-i}^{(t-1)} \right\|_1 .
\end{aligned}$$

Further, for any $a, b, a', b' \in [0,1]$, we have $|ab - a'b'| = |ab - ab' + ab' - a'b'| \leq a\,|b - b'| + |a - a'|\,b' \leq |a - a'| + |b - b'|$. Therefore, by recursively using it, for any $\boldsymbol{a}_{-i} \in \bigtimes_{j \neq i} \mathcal{A}_j$, we have

$$\left| \pi_{-i}^{(t)}(\boldsymbol{a}_{-i}) - \pi_{-i}^{(t-1)}(\boldsymbol{a}_{-i}) \right| = \left| \prod_{j \neq i} \pi_j^{(t)}(a_j) - \prod_{j \neq i} \pi_j^{(t-1)}(a_j) \right| \leq \sum_{j \neq i} \left| \pi_j^{(t)}(a_j) - \pi_j^{(t-1)}(a_j) \right|.$$

Finally,

$$\left\| u_i^{(t)} - \boldsymbol{u}_i^{(t-1)} \right\| \leq \sqrt{|\mathcal{A}_i|} \left\| \pi_{-i}^{(t)} - \pi_{-i}^{(t-1)} \right\|_1 \leq \prod_{j=1}^{N} |\mathcal{A}_j| \sum_{j \neq i} \left\| \pi_j^{(t)} - \pi_j^{(t-1)} \right\|_1 .$$ □

L.1 Proof of Theorem 6.3 and Theorem 6.4

Before proving Theorem 6.3 and Theorem 6.4, we will show that when Assumption 6.2 is satisfied, then the strategy variation is bounded.

Lemma L.1. *Consider when Assumption 6.2 is satisfied. For both full-information and bandit, Algorithm 2 satisfies the following,*

$$\sum_{t=1}^{T-1} \left\| \pi^{(t)} - \pi^{(t+1)} \right\| \leq \mathcal{O}\left(\eta T\right).$$

When Assumption 5.2 is also satisfied, the following holds for Algorithm 3 in bandit setting,

$$\sum_{t=1}^{T-1} \left\| \pi^{(t)} - \pi^{(t+1)} \right\| \leq \mathcal{O}\left(\eta T + LT\right).$$

With Lemma L.1, we can prove that $R_i^{(T),\text{external}}$ is sublinear for any player $i \in [N]$ by Theorem 4.2, Theorem 6.3, Theorem 5.3, and Theorem 5.4. Then, according to the folklore theorem (Cesa-Bianchi & Lugosi, 2006), Theorem 6.3 and Theorem 6.4 are proved. □

Lemma L.1. *Consider when Assumption 6.2 is satisfied. For both full-information and bandit, Algorithm 2 satisfies the following,*

$$\sum_{t=1}^{T-1} \left\| \pi^{(t)} - \pi^{(t+1)} \right\| \leq \mathcal{O}\left(\eta T\right).$$

When Assumption 5.2 is also satisfied, the following holds for Algorithm 3 in bandit setting,

$$\sum_{t=1}^{T-1} \left\| \pi^{(t)} - \pi^{(t+1)} \right\| \leq \mathcal{O}\left(\eta T + LT\right).$$

Proof. For Algorithm 2 and full-information, the proof simply follows from $\widetilde{\boldsymbol{u}}^{(t)} \in [-1,1]^{\mathcal{A}}$ and Assumption 6.2.

For Algorithm 2 and bandit, we have

$$\left\| \pi^{(t+1)} - \pi^{(t)} \right\| = (1-\gamma) \left\| \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)} \right)_{s=1}^{t+1} \right) - \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)} \right)_{s=1}^{t} \right) \right\| \leq \eta.$$

Then, we conclude the proof.

For Algorithm 3 and bandit, for any $t \not\equiv 0 \pmod{M}$, we have $\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} = \widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)}$. Therefore, by Assumption 6.2, we have

$$\begin{aligned}
\left\| \pi^{(t-1)} - \pi^{(t)} \right\| &\leq \left\| \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t-1} \right) - \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} \right)_{s=1}^{t} \right) \right\| \\
&= \left\| \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t-1} \right) - \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t} \right) \right\| \leq \eta.
\end{aligned}$$

For any $t \equiv 0 \pmod{M}$, let $\boldsymbol{u} = \frac{\widetilde{u}_{\text{empirical}}^{(t)}(a) n^{(t)}(a) - \widetilde{u}_{\text{empirical}}^{(t-M)}(a) n^{(t-M)}(a)}{n^{(t)}(a) - n^{(t-M)}(a)}$, then we have

$$\begin{aligned}
\left\| \widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} - \widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right\| &= \left\| \left(\frac{(t/M-1)\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} + \boldsymbol{u}}{t/M} \right) - \widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right\| \\
&\leq \frac{M}{t} \left(\left\| \widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right\| + \|\boldsymbol{u}\| \right) \leq \frac{2M}{t} \sqrt{|\mathcal{A}|}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\| \pi^{(t-1)} - \pi^{(t)} \right\| \leq \left\| \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t-1} \right) - \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} \right)_{s=1}^{t} \right) \right\| \\
&\leq \left\| \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t-1} \right) - \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t} \right) \right\| + \left\| \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t-1)} \right)_{s=1}^{t} \right) - \text{Alg}\left(\left(\widetilde{\boldsymbol{u}}_{\text{avg-est}}^{(t)} \right)_{s=1}^{t} \right) \right\| \\
&\overset{(i)}{\leq} \eta + Lt \left(\frac{2M}{t} \sqrt{|\mathcal{A}|} \right) \\
&= \eta + 2LM\sqrt{|\mathcal{A}|}.
\end{aligned}$$

(i) uses Assumption 5.2. Then, the total variation is bounded by

$$\sum_{t=1}^{T-1} \left\| \pi^{(t+1)} - \pi^{(t)} \right\| \leq \mathcal{O}\left(\eta T + LM \frac{T}{M} \right) \leq \mathcal{O}\left(\eta T + LT\right),$$

since there are at most $\frac{T}{M}$ timesteps $t \in [T]$ satisfying $t \equiv 0 \pmod{M}$. □

M PROPERTIES OF FOLLOW THE REGULARIZED LEADER (FTRL)

Firstly, we will define the strongly convex function and its conjugate function.

Definition M.1. *For any integer n, a differentiable function $\psi(\boldsymbol{x})\colon \mathbb{R}^n \to \mathbb{R}$ is called c_0-strongly convex ($c_0 > 0$) when*

$$\psi(\boldsymbol{x}) \geq \psi(\boldsymbol{x}') + \langle \nabla\psi(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle + \frac{c_0}{2} \|\boldsymbol{x} - \boldsymbol{x}'\|^2 \tag{M.1}$$

holds for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$. Its conjugate function is defined as

$$\psi^*(\boldsymbol{y})\colon \mathbb{R}^n \to \mathbb{R} := \sup_{\boldsymbol{x}\in\mathbb{R}^n} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \psi(\boldsymbol{x}).$$

Specifically, if (M.1) holds for $c_0 = 0$, then we call ψ a convex function.

Next, we will introduce follow the regularized leader (FTRL).

Definition M.2 (Follow the Regularized Leader (FTRL))**.** *For any $T > 0$ and at any timestep $t \in \{0\} \cup [T-1]$, given the utility vectors $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}$, the strategy at timestep $t+1$, $\pi^{(t+1)}$, is defined as,*

$$\pi^{(t+1)} = \operatorname*{argmax}_{\pi\in\Delta^{\mathcal{A}}} \left(\lambda \sum_{s=1}^{t} \left\langle \boldsymbol{u}^{(s)}, \pi \right\rangle - \psi(\pi) \right), \tag{FTRL}$$

for some constant $\lambda > 0$. Typically, λ is taken to be $\Theta\left(T^{-r}\right)$ for some constant $r > 0$.

Now, we can introduce the smoothness of (FTRL).

Lemma M.3. *For any c_0-strongly convex and differentiable function $\psi\colon \Delta^{\mathcal{A}} \to \mathbb{R}$,* (FTRL) *satisfies Assumption 5.2 and Assumption 6.2 with $L = \frac{\lambda}{c_0}$ and $\eta = \frac{\lambda}{c_0}\sqrt{|\mathcal{A}|}$.*

Proof. By first-order optimality, at any timestep $t \in \{0\} \cup [T-1]$ for any two sequences of utility vectors $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}$ and $\left(\boldsymbol{u}'^{(s)}\right)_{s=1}^{t}$, let the corresponding strategy generated by (FTRL) be $\pi^{(t+1)}$ and $\pi'^{(t+1)}$ respectively, we have

$$\left\langle \lambda \sum_{s=1}^{t} \boldsymbol{u}^{(s)} - \nabla\psi\left(\pi^{(t+1)}\right), \pi'^{(t+1)} - \pi^{(t+1)} \right\rangle \leq 0$$
$$\left\langle \lambda \sum_{s=1}^{t} \boldsymbol{u}'^{(s)} - \nabla\psi\left(\pi'^{(t+1)}\right), \pi^{(t+1)} - \pi'^{(t+1)} \right\rangle \leq 0.$$

By summing them up and rearranging the terms, we have

$$\left\langle \lambda \sum_{s=1}^{t} \boldsymbol{u}'^{(s)} - \lambda \sum_{s=1}^{t} \boldsymbol{u}^{(s)}, \pi'^{(t+1)} - \pi^{(t+1)} \right\rangle \geq \left\langle \nabla\psi\left(\pi'^{(t+1)}\right) - \nabla\psi\left(\pi^{(t+1)}\right), \pi'^{(t+1)} - \pi^{(t+1)} \right\rangle.$$

Since ψ is c_0-strongly convex, we have

$$\psi\left(\pi^{(t+1)}\right) \geq \psi\left(\pi'^{(t+1)}\right) + \left\langle \nabla\psi\left(\pi'^{(t+1)}\right), \pi^{(t+1)} - \pi'^{(t+1)} \right\rangle + \frac{c_0}{2}\left\|\pi^{(t+1)} - \pi'^{(t+1)}\right\|^2$$
$$\psi\left(\pi'^{(t+1)}\right) \geq \psi\left(\pi^{(t+1)}\right) + \left\langle \nabla\psi\left(\pi^{(t+1)}\right), \pi'^{(t+1)} - \pi^{(t+1)} \right\rangle + \frac{c_0}{2}\left\|\pi^{(t+1)} - \pi'^{(t+1)}\right\|^2.$$

By summing them up and rearranging the terms, we have

$$\left\langle \nabla\psi\left(\pi'^{(t+1)}\right) - \nabla\psi\left(\pi^{(t+1)}\right), \pi'^{(t+1)} - \pi^{(t+1)} \right\rangle \geq c_0 \left\|\pi^{(t+1)} - \pi'^{(t+1)}\right\|^2.$$

Therefore,

$$\begin{aligned} c_0 \left\|\pi^{(t+1)} - \pi'^{(t+1)}\right\|^2 &\leq \left\langle \lambda \sum_{s=1}^{t} \boldsymbol{u}'^{(s)} - \lambda \sum_{s=1}^{t} \boldsymbol{u}^{(s)}, \pi'^{(t+1)} - \pi^{(t+1)} \right\rangle \\ &\overset{(i)}{\leq} \left\| \lambda \sum_{s=1}^{t} \boldsymbol{u}'^{(s)} - \lambda \sum_{s=1}^{t} \boldsymbol{u}^{(s)} \right\| \cdot \left\|\pi'^{(t+1)} - \pi^{(t+1)}\right\|. \end{aligned}$$

(i) is by Hölder's Inequality. Then,

$$\left\| \lambda \sum_{s=1}^{t} \boldsymbol{u}'^{(s)} - \lambda \sum_{s=1}^{t} \boldsymbol{u}^{(s)} \right\| \leq c_0 \left\| \pi^{(t+1)} - \pi'^{(t+1)} \right\|,$$

so that (FTRL) satisfies Assumption 5.2 with $L = \frac{\lambda}{c_0}$. Furthermore, note that the results above also hold for sequences of utility vectors of different lengths (not necessarily equal to length t simultaneously). As a result, we have

$$\left\| \pi^{(t+1)} - \pi^{(t)} \right\| \leq \frac{\lambda}{c_0} \left\| \boldsymbol{u}^{(t)} \right\| \leq \frac{\lambda}{c_0} \sqrt{|\mathcal{A}|},$$

for any $t \in \{0\} \cup [T-1]$, which implies that $\eta = \frac{\lambda}{c_0}\sqrt{|\mathcal{A}|}$ in Assumption 6.2 for (FTRL). □

N Concluding Remarks

In this paper, we studied online learning and equilibrium computation with ranking feedback, which is particularly relevant to application scenarios with humans in the loop. Focusing on the classical (external-)regret metric, we designed novel hardness instances to show that achieving sublinear regret can be hard in general, in a few different ranking models and feedback settings. We then developed new algorithms to achieve sublinear regret under an additional assumption on the sublinear variation of the utility, leading to an equilibrium computation result in the repeated game setting. We believe our work paves the way for promising avenues of future research. For example, it would be interesting to close the gap between the lower-bound and the positive result for **AvgUtil Rank** under bandit feedback, *i.e.,* either show the hardness when τ is a *constant* or achieve sublinear regret for constant τ without Assumption 4.1. Moreover, applying our algorithms to real-world datasets with ranking feedback, such as ride-sharing and match-dating, would also be of great interest.