Online Learning and Equilibrium Computation with Ranking Feedback

Mingyang Liu [1] **Yongshan Chen** [2] **Zhiyuan Fan** [1] **Gabriele Farina** [1] **Asuman E. Ozdaglar** [1] **Kaiqing Zhang** [2]

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 with privacy concerns. In this paper, we study an online learning setting 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 with the instantaneous utility ranking feedback in general. Moreover, we show that when the ranking model is relatively deterministic (*i.e.,* with a small temperature in the Plackett-Luce model), 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 effectiveness of our algorithms via numerical experiments.

[1]EECS, Massachusetts Institute of Technology [2]ECE, University of Maryland, College Park. Correspondence to: Mingyang Liu <liumy19@mit.edu>, Yongshan Chen <eternalfir@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. PMLR 267, 2025.

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 that the (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 (NFG), the time-average strategy will approximate the coarse correlated equilibrium (CCE) of the game (Cesa-Bianchi and 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 an environment with 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 exemplified 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 (cf. Figure 1 (a)) that recommends commodities to a stream of customers in an online fashion, where the customers at different timesteps may have different preferences for 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 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. 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.

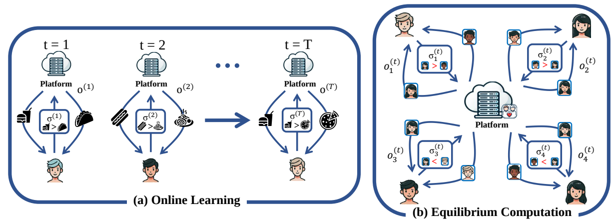

Figure 1: Two real-world examples of Online Learning and Equilibrium Computation with Ranking Feedback. In Figure (a), the online platform recommends choices of food to the customer at each timestep and receives a ranking feedback to improve the recommendation quality. Figure (b) illustrates an online dating app where the app recommends matches, the users rank other candidates, and the app uses the rankings to find matching equilibria among the users over time.

Ranking feedback may become even relevant in game-theoretic settings, when multiple humans continuously interact with each other, and the objective is to compute a certain equilibrium of the game. For example, consider an online dating platform recommending candidates for matching (cf. Figure 1 (b)). Each user may only have a ranking of the recommended candidates in each round, and the platform aims to find an equilibrium (a matching between users) so that all the users are satisfied. 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). Our focused setting to address these scenarios may appear related, but fundamentally different from the classical stable matching one (Gale and Shapley, 1962), see Appendix A for a detailed comparison.

In this paper, we seek to systematically study online learning and equilibrium computation with ranking feedback, where the loss vectors may be non-stochastically and even adversarially generated. This setting can be viewed as a generalization of the stochastic bandit with ranking feedback studied recently in Maran et al. (2024) (see a detailed comparison and related work in Appendix A). We aim to understand when regret minimization in our setting is possible, and also develop new algorithms with regret and equilibrium approximation guarantees. 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 show that: i) sublinear regret cannot be achieved with **InstUtil Rank** feedback, nor (up to logarithmic terms) with **AvgUtil Rank** feedback when the ranking model is too *deterministic* (*i.e.,* the temperature $\tau > 0$ of the ranking model in (PL) is small); ii) We propose new algorithms to achieve sublinear regret, under an additional assumption on the sublinear variation of the utility vectors; iii) Such an assumption can be avoided with full-information **AvgUtil Rank** feedback and a constant τ; iv) When all the players follow our no-regret learning algorithms in repeatedly playing a normal-form game, an approximate CCE can be computed. Our results are summarized in Table 1. Experiments that validate the effectiveness of our algorithms can be found in Appendix C.

2. Preliminaries

2.1. 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 $\mathcal{A} := \left\{a^1, a^2, \dots, a^{|\mathcal{A}|}\right\}$ with $|\mathcal{A}| > 1$. At each timestep t, the agent will commit to a strategy $\pi^{(t)} \in \Delta^{\mathcal{A}}$ and receive either a utility vector $\boldsymbol{u}^{(t)} \in [-1, 1]^{\mathcal{A}}$ or a realized utility value $u^{(t)}(a^{(t)})$ for the action $a^{(t)}$ taken from $a^{(t)} \sim \pi^{(t)}$, in the *full-information* and *bandit* setting, respectively. 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

$$R^{(T),\text{external}} := \max_{\widehat{\pi} \in \Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \pi^{(t)} \right\rangle. \tag{2.1}$$

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 $t \in [T]$.

2.2. 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 deterministic *black-box* oracle in our algorithms to be designed later. As a preliminary, we formally introduce such deterministic oracles here: we use $\text{Alg}\colon \bigcup_{t=0}^{\infty} \left(\mathbb{R}^{\mathcal{A}}\right)^t \to \Delta^{\mathcal{A}}$ to denote such an online learning algorithm, which is a mapping from a sequence of utility

Lower Bound	Full-Information	Bandit
InstUtil Rank	$\Omega(T)$ for $\tau \leq \mathcal{O}(1)$	
AvgUtil Rank	$\widetilde{\Omega}(T)$ for $\tau \leq \mathcal{O}\left(\frac{1}{T \log T}\right)$	$\Omega(T)$ for $\tau \leq \mathcal{O}\left(\frac{1}{\log T}\right)$

Upper Bound ($\tau = \mathcal{O}(1)$, Sublinear Regret)	Full-Information	Bandit
InstUtil Rank	Assumption 4.2	
AvgUtil Rank	✓	Assumption 4.2 ($q < \frac{1}{3}$)

Table 1: Summary of our contributions, including the *negative results* (top) and the *positive results* (bottom). The bottom table shows the necessary assumptions to achieve sublinear regret in that setting (✓means no assumption is required). $\tau > 0$ denotes the temperature of the ranking model in (PL).

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 $\pi^{(t+1)} = \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}\right)$ as the strategy at timestep $t+1$. Finally, we can denote the (external-)regret under Alg as

$$R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) := \max_{\widehat{\pi} \in \Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} - \text{Alg}\left(\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t-1}\right)\right\rangle, \tag{2.2}$$

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

3. Online Learning with Ranking Feedback

In online learning with ranking feedback, at each 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 $o^{(t)}$, and receive a permutation $\sigma^{(t)} \in \Sigma\left(o^{(t)}\right)$ from the environment, representing a *ranking* of those actions in $o^{(t)}$. In the full-information setting, $o^{(t)} = \mathcal{A}$, *i.e.,* the whole action set is proposed. In the bandit setting, $\left|o^{(t)}\right| = K < |\mathcal{A}|$. Suppose the agent's strategy at timestep t is $\pi^{(t)} \in \Delta^{\mathcal{A}}$, then we assume that in this bandit setting, the actions in $o^{(t)}$ are proposed in an *unbiased* way, such that $\mathbb{E}\left[\frac{\sum_{a\in o^{(t)}} u^{(t)}(a)}{K}\right] = \langle \boldsymbol{u}^{(t)}, \pi^{(t)}\rangle$, 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 k^{th} 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 according to

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

where $\boldsymbol{r}^{(t)} \in \mathbb{R}^{\mathcal{A}}$ is some utility vector based on which the ranking is determined, $\tau > 0$ is the *temperature* parameter 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 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**). The two feedback types may be motivated by different applications (cf. Section 1), and have been studied for dueling-bandits (Yue et al., 2012; Saha and Gaillard, 2022) and multi-armed bandits with ranking feedback (Maran et al., 2024), respectively. Both feedback types 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. *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

$\langle \boldsymbol{u}^{(t)}, \pi^{(t)} \rangle$. Note that this does not mean the agent can access the full vector $\boldsymbol{u}^{(t)}$, since this defeats the purpose of our ranking-feedback setting. Hence, the standard (external-)regret $R^{(T),\text{external}}$ defined in (2.1) will serve as the metric to evaluate the agent's performance in this online learning process.

Bandit 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 $\frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a)$, leading to the following performance metric of regret:

$$R^{(T)} := \max_{\widehat{\pi}\in\Delta^{\mathcal{A}}} \sum_{t=1}^{T} \left(\left\langle \boldsymbol{u}^{(t)}, \widehat{\pi} \right\rangle - \frac{1}{K} \sum_{a\in o^{(t)}} u^{(t)}(a) \right). \tag{3.1}$$

Note that such a definition is an external regret, which differs from the regret studied in (multi-)dueling-bandits (Yue et al., 2012; Du et al., 2020; Saha et al., 2021; Saha and Gaillard, 2022), and can be viewed as a generalization of the one considered in (Maran et al., 2024) when $K = 1$.

AvgUtil Rank: Ranking with Time-average Utility.

The second type of ranking-feedback is based on the *time-average* utility, which differs 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, when $\tau \to 0^+$ and the environment is stationary and stochastic, such a model aligns with the one studied in the recent work of Maran et al. (2024).

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), and the same (external-)regret $R^{(T),\text{external}}$ from (2.1) will be used as the metric.

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 out, 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

$$u_{\text{empirical}}^{(t)}(a) := \frac{\sum_{s=1}^{t} u^{(s)}(a) \sum_{a'\in o^{(t)}} \mathbb{1}(a = a')}{\sum_{s=1}^{t} \sum_{a'\in o^{(t)}} \mathbb{1}(a = a')}. \tag{3.2}$$

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 with **InstUtil Rank**), is due to the fact that when one is allowed to leverage 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. The regret metric will still be the one in (3.1). The background and formalism of equilibrium computation in the game-theoretic setting (with ranking feedback) can be found in Appendix B.

4. Hardness Results

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 4.1 in the following shows that for any temperature τ in (PL) not larger than a *constant*, there exists a sequence of utility vectors such that the expected regret is linear under **InstUtil Rank**, for both full-information and bandit settings.

Theorem 4.1. *Consider* **InstUtil Rank**. *For any $T > 0$, temperature $0 < \tau \leq 0.1$, and online learning algorithm, there exists a sequence of utilities $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$ such that $\min\left\{\mathbb{E}\left[R^{(T),\text{external}}\right], \mathbb{E}\left[R^{(T)}\right]\right\} \geq \Omega(T)$ in both full-information and bandit settings. The expectation is taken over the randomness of the algorithm and the ranking.*

To prove Theorem 4.1, we need to construct two sequences of utility vectors, which yield the same ranking under **InstUtil Rank** in expectation. However, being no-regret in one of them will result in linear regret in the other. The detailed proof can be found in Appendix D.

The key challenge in achieving no-regret in the hard instance above is that the utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$ change arbitrarily fast, *i.e.,* the accumulated variation grows *linearly* in 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 4.2 (Sublinear variation of utility vectors)**.** *The utility vectors $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$ have a sublinear variation over time,* i.e., *for some $q < 1$,*

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

Our result stated in Section 5 next will show that with Assumption 4.2, we can achieve sublinear regret, and thus close the gap. Moreover, note that in a game where the opponents all run common no-regret learning algorithms such as follow-the-regularized-leader (FTRL), Assumption 4.2 will be satisfied (cf. Lemma L.3).

Next, we show in Theorem 4.3 that when **AvgUtil Rank** is used, and τ is small enough, the minimal regret is still at least linear in T (up to logarithmic terms).

Theorem 4.3. *Consider* **AvgUtil Rank** *with full-information feedback. For any $T > 0$, temperature $0 < \tau \le \mathcal{O}\left(\frac{1}{T \log T}\right)$, and online learning algorithm, there exists $T' \ge T$ and a sequence of utilities $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T'}$ such that $\mathbb{E}\left[R^{(T'),\text{external}}\right] \ge \widetilde{\Omega}\left(T'\right)$. The expectation is taken over the randomness of the algorithm and the ranking.*

To prove Theorem 4.3, we need to construct $\log T$ sequences of utility vectors, with the same ranking feedback when τ is small. Then, we can show that at least one of them suffers an average regret $\widetilde{\Omega}(1)$.

Given Theorem 4.3, it is impossible to achieve $\widetilde{o}(T)$ with **AvgUtil Rank** when τ is very small. However, in Section 6, we will close the gap by showing that when τ is a *constant* (*i.e.*, $\mathcal{O}(1)$), we can achieve sublinear regret with **AvgUtil Rank**, even without Assumption 4.2.

Due to the different instantiations of $\boldsymbol{r}^{(t)}$ in the full-information and the bandit settings under **AvgUtil Rank**, we have a separate hardness result for the bandit setting, stronger than Theorem 4.3, as it allows a larger τ and avoids logarithmic terms. The result can be viewed as strengthening the hardness for the adversarial bandit setting in (Maran et al., 2024), which corresponds to the case with $\tau \to 0^+$.

Theorem 4.4. *Consider* **AvgUtil Rank** *with bandit feedback. For any $T > 0$, temperature $0 < \tau \le \mathcal{O}\left(\frac{1}{\log T}\right)$, and online learning algorithm, there exists a sequence of utilities $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{4T}$ such that $\mathbb{E}\left[R^{(4T)}\right] \ge \Omega\left(T\right)$. The expectation is taken over the randomness of the algorithm and the ranking.*

To prove Theorem 4.4, we need to construct two utility sequences such that sublinear regret in the first utility sequence will lead to insufficient exploration for the second sequence. As a result, when τ is small, those two sequences cannot be differentiated, and a linear regret must be incurred in one of them. The details are postponed to Appendix D due to space constraints.

5. Online Learning with InstUtil Rank Feedback

We start by introducing a novel utility estimation oracle to be used in our later algorithms.

5.1. Utility Estimation

A natural idea to learn under ranking feedback is to use the feedback to *estimate* the 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 $t \ge m$, we use the past m steps' permutations $\left\{\sigma^{(s)}\right\}_{s=t-m+1}^{m}$ to estimate the utility $\boldsymbol{u}^{(t)}$. Due to the non-convexity of the (PL), the key point to estimate utilities is to decompose the ranking of K actions into pair-wise rankings. Then, we can utilize the properties of the logistic function, such as monotonicity, to convert the estimation error on ranking probabilities back to utilities. The full algorithm can be found in Algorithm 1, with the following guarantee.

Theorem 5.1. *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 $t \ge m'$, when $m'p^4 \ge 2\log\left(\frac{2}{\delta}\right)$, with probability at least $1-\delta$, the estimate $\widetilde{\boldsymbol{u}}^{(t)}$ satisfies,*

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

When taking δ, p, τ as constants, the accumulated estimation error $\sum_{t=1}^{T}\left\|\widetilde{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\right\|_\infty$ will be bounded by $\mathcal{O}\left(\frac{T}{\sqrt{m'}} + m'P^{(T)}\right)$, which implies that sublinear accumulated estimation error is achieved when $P^{(T)}$ is sublinear (Assumption 4.2). Moreover, Theorem 5.1 implies that when $\tau \to 0^+$, the estimation error upper bound goes to $+\infty$. This makes sense intuitively: 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 upper bound also goes to $+\infty$, since the ranking is always sampled uniformly regardless of the utility vectors.

The full proof of Theorem 5.1 is deferred to Appendix E. Then, we show how to achieve sublinear regret in both full-information and bandit settings with **InstUtil Rank**, based on such an estimator.

5.2. Sublinear Regret with InstUtil Rank

This section shows that for any online learning algorithm that can achieve sublinear external regret with numeric utility feedback, we can construct an online learning algorithm with **InstUtil Rank** feedback based on it, in a black-box way.

With 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 5.1 with $p=1$ (since all actions are proposed at each timestep).

With bandit feedback, 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 $R^{(T)}$ (defined in (3.1)), 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 $\pi^{(t+1)} = (1-\gamma)\mathrm{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right) + \gamma\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}$, *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 of the algorithm can be found in Figure 5, and the details are in Algorithm 2.

Then, we have the following theorem.

Theorem 5.2. *Consider* **InstUtil Rank** *with constant $\tau > 0$. By running 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 the window size m and γ properly, we have that with probability at least $1-\delta$, $R^{(T),\mathrm{external}}$ satisfies*

$$
\begin{aligned}
R^{(T),\mathrm{external}} \leq & R^{(T),\mathrm{external}}\left(\mathrm{Alg}, \left(\widetilde{\boldsymbol{u}}^{(t)}\right)_{t=1}^{T}\right) \\
& + \mathcal{O}\left(\left(P^{(T)}\right)^{\frac{1}{3}} T^{\frac{2}{3}}\left(\log\left(\frac{T}{\delta}\right)\right)^{\frac{1}{3}}\right)
\end{aligned} \tag{Full-Info}
$$

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

The proof is deferred to Appendices F and G. Theorem 5.2 implies that when $P^{(T)}$, the variation of utility vectors, is sublinear, the regret of Algorithm 2 will be sublinear.

6. Online Learning with AvgUtil Rank Feedback

6.1. Utility Estimation

We also start by introducing a new utility estimation algorithm. Since $\sigma^{(t)}$ is generated based on $\boldsymbol{u}_{\mathrm{avg}}^{(t)}$, we will estimate $\boldsymbol{u}_{\mathrm{avg}}^{(t)}$ instead. We will still apply Algorithm 1, which will generate $\widetilde{\boldsymbol{u}}_{\mathrm{avg}}^{(t)}$, an estimate of $\boldsymbol{u}_{\mathrm{avg}}^{(t)}$, when the permutation is sampled under **AvgUtil Rank** feedback. Moreover, notice that

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

Therefore, $\sum_{s=t-m'+1}^{t-1}\left\|\boldsymbol{u}_{\mathrm{avg}}^{(s+1)} - \boldsymbol{u}_{\mathrm{avg}}^{(s)}\right\|_{\infty}$, the counterpart of $\sum_{s=t-m'+1}^{t-1}\left\|\boldsymbol{u}^{(s+1)} - \boldsymbol{u}^{(s)}\right\|_{\infty}$ in Theorem 5.1, can be bounded by $\sum_{s=t-m'+1}^{t-1}\frac{1}{s+1}$, which is irrelevant of $P^{(T)}$ in Assumption 4.2.

6.2. Full-Information Setting

Unlike the case with **InstUtil Rank** feedback, where any (full-information) online learning algorithm can be leveraged, the algorithm with **AvgUtil Rank** feedback needs to be insensitive to the changes in accumulated utility, such as FTRL. Because we want the strategies generated by the online learner, given the estimated average utilities as input, to be close to those generated using the ground-truth utilities. Then, the regret is sublinear since our strategies are close to the strategies that yield a sublinear regret.

Assumption 6.1. *The (full-information) online learning algorithm* Alg *needs to satisfy the following condition: for any $T > 0$, $t \in [T]$, sequences of utilities $\left(\boldsymbol{u}^{(s)}\right)_{s=1}^{t}, \left(\boldsymbol{u}'^{(s)}\right)_{s=1}^{t} \in \left(\mathbb{R}^{\mathcal{A}}\right)^{t}$, we have*

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

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

It can be verified that follow-the-regularized-leader with any strongly convex regularizer satisfies this assumption (cf. Lemma L.3). Then, similar to Section 5.2, any online learning algorithm satisfying Assumption 6.1 can achieve a sublinear regret when equipped with the utility estimator in Algorithm 1. The overall procedure is summarized in Algorithm 3 and Figure 6, with the following guarantee.

Theorem 6.2. *Consider* **AvgUtil Rank** *with constant $\tau > 0$ and full-information feedback. By running 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 6.1, by choosing m properly, we have that with probability at least $1-\delta$, $R^{(T),\mathrm{external}}$ 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 6.2 shows that with a small enough $L = \Theta(T^{-c})$ satisfying $c > 2/3$, $R^{(T),\text{external}}$ can be made sublinear in T. The proof and the formal version of Theorem 6.2 are provided in Appendix H.

6.3. Bandit Setting

By appling Algorithm 1, we can only obtain an estimate of $\boldsymbol{u}^{(t)}_{\text{empirical}}$ instead of $\boldsymbol{u}^{(t)}_{\text{avg}}$. However, almost all no-regret learning algorithms made decisions according to the accumulated utility, such as mirror descent, FTRL, and regret matching (Zinkevich et al., 2007; Hazan et al., 2016). Let $n^{(t)}(a) := \sum_{s=1}^{t} \#_{o^{(s)}}(a)$ for any $a \in \mathcal{A}$ be the number of times action a has been proposed up to timestep t, where $\#_{o^{(s)}}(a)$ is the number of occurrences of a in $o^{(s)}$. A natural idea is to compute $n^{(t)}(a)u^{(t)}_{\text{empirical}}(a) - n^{(t-1)}(a)u^{(t-1)}_{\text{empirical}}(a)$ to get an estimate of $u^{(t)}(a)$. Nonetheless, the variance will be too large due to the multiplication of $n^{(t)}(a) \propto t$.

To address this issue, we divide the timesteps $\{1, 2, \dots, t\}$ into $\lceil t/M \rceil$ blocks, with each block containing M timesteps except for the last one. Then, for each block $\{s \cdot M + 1, s \cdot M + 2, \dots, (s+1)M\}$ (for $s \leq \lfloor \frac{t}{M} \rfloor - 1$) and $a \in \mathcal{A}$, we estimate $\frac{1}{M}\sum_{s'=s\cdot M+1}^{(s+1)M} u^{(s')}(a)$ by computing

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

In this way, the multiplier on $\widetilde{u}^{((s+1)\cdot M)}_{\text{empirical}}(a)$ is $\frac{n^{((s+1)\cdot M)}(a)}{n^{((s+1)\cdot M)}(a) - n^{(s\cdot M)}(a)} \propto \frac{t}{M}$. The trade-off of M is that the value above estimates the average utility of a *when a is chosen*, which may differ significantly from the true average utility of a in that block. Because when M is large, the accumulated utility variation is larger. The full algorithm is illustrated in Algorithm 3 and Figure 6.

Theorem 6.3. *Consider* **AvgUtil Rank** *with constant $\tau > 0$ and bandit feedback. By running 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 6.1, 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 logarithmic dependence on T.

By choosing $L = \Theta(T^{-c})$ with $c \in \left(\frac{5}{6} + \frac{q}{2}, 1\right)$, $\widetilde{\mathcal{O}}\left(\left(\log\left(\frac{1}{\delta}\right)\right)^2 L^{\frac{1}{3}} T^{\frac{23}{18}} \left(P^{(T)}\right)^{\frac{1}{6}}\right)$ is sublinear when $P^{(T)} \leq \mathcal{O}(T^q)$ for some $q < \frac{1}{3}$. Theorem 6.3 guarantees sublinear $R^{(T)}$. We need $c < 1$ because typically $R^{(T),\text{external}}\left(\text{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) \leq \mathcal{O}\left(\frac{1}{L} + LT\right)$, see *e.g.,* FTRL with any strongly convex regularizer (Hazan et al., 2016).

7. Equilibrium Computation with Ranking Feedback

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

$$R_i^{(T),\text{external}} := \max_{\widehat{\pi}_i \in \Delta^{\mathcal{A}_i}} \sum_{t=1}^{T} \left\langle \boldsymbol{u}_i^{(t)}, \widehat{\pi}_i - \pi_i^{(t)} \right\rangle, \tag{7.1}$$

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

Applying the algorithm in Section 5 (for **InstUtil Rank** feedback) or Section 6 (for **AvgUtil Rank** feedback), we achieve sublinear $R_i^{(T),\text{external}}$ for each player $i \in [N]$. Note that $P^{(T)}$ in Assumption 4.2 can be bounded by the summation of all players' strategy variation (cf. Lemma K.1). Thus, to ensure $P^{(T)}$ is sublinear in T, Alg needs to additionally satisfy the following assumption.

Assumption 7.1 (Sublinear variation of strategies)**.** *The (full-information) 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 $\eta = \Theta(T^{-w})$ for some constant $w \in (0,1)$.*

Mirror descent (cf. Wei et al. (2021, Lemma 1) and Liu et al. (2023, Lemma C.5)) and FTRL with any strongly convex regularizer (see Lemma L.3 for the proof), both satisfy this property. When Assumption 7.1 is satisfied, one can achieve sublinear regret with **InstUtil Rank**, under both full-information and bandit feedback. The formal statement is as follows.

Theorem 7.2. *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 7.1, by choosing* M, m, γ *according to Theorem 5.2, we have that with probability at least* $1-\delta$*, the algorithm finds an* ϵ*-CCE, with*

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

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

With **AvgUtil Rank** feedback, when all the players apply Algorithm 3 and both Assumption 6.1 and Assumption 7.1 are satisfied for the oracle Alg being used, the external regret of each player will be sublinear in T according to Theorem 6.2. Finally, we have the statement below.

Theorem 7.3. *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 6.1, by choosing* M, m, γ *according to Theorem 6.2, we have that with probability at least* $1-\delta$*, the algorithm finds an* ϵ*-CCE under full-information, with*

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

When M, m, γ *are chosen according to Theorem 6.3 and Assumption 7.1 is also satisfied, the following holds under bandit feedback,*

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

Lastly, we would like to remark that although the online learning setting can be hard with a small τ (cf. the hardness results in Theorem 4.3 and Theorem 4.4), computing an equilibrium is still possible even when $\tau \to 0^+$. A detailed discussion can be found in Remark K.3.

8. Conclusion and Limitations

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 4.2. Moreover, applying our algorithms to real-world datasets with ranking feedback, such as ride-sharing and match-dating, would also be of great interest.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

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)*, pages 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.

Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $log^2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pages 705–715. PMLR, 2021.

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

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

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*, pages 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23160–23167, 2025.

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.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American mathematical monthly*, 69(1):9–15, 1962.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

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

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

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.

Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021.

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*, pages 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*, pages 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*, pages 6263–6289. PMLR, 2023.

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

Vade Shah, Bryce L Ferguson, and Jason R Marden. Learning optimal stable matches in decentralized markets with unknown preferences. *arXiv preprint arXiv:2409.04669*, 2024a.

Vade Shah, Bryce L Ferguson, and Jason R Marden. Two-sided learning in decentralized matching markets. *arXiv preprint arXiv:2411.02377*, 2024b.

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.

Martin Zinkevich, Michael Johanson, Michael H. Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Neural Information Processing Systems (NeurIPS)*, 2007.

A. 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 and Gaillard, 2022; Saha and 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 *both* settings with instantaneous and time-average utilities for ranking. More importantly, the regret notions studied in these works were particularly designed for the dueling-bandit setting, and thus 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 and Gaillard, 2022; Saha and Gopalan, 2019; Du et al., 2020), while we focus on the *non-stochastic* setting where the environment is arbitrary and potentially *adversarial*, as in online learning (Shalev-Shwartz et al., 2012; Hazan et al., 2016). Due to the last two differences, the implication of these dueling-bandit algorithms in the game-theoretic setting is unclear, while our algorithms find an approximate CCE of the game, as a corollary of the no-(external-)regret guarantee.

Reinforcement Learning from Human Feedback 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 the classical Markov decision process model with online feedback (Novoseller et al., 2020; Saha et al., 2023; Xu et al., 2020). However, the utility/reward functions in these works are again stationary, and the regret notions extend those in the dueling-bandits literature, which are thus different from ours. Hence, these results do not apply to our adversarial online learning and game-theoretic settings.

Learning of Stable Matchings. Some of our motivating scenarios for the game-theoretic setting may also be modeled as the *stable matching* problem (Gale and Shapley, 1962), which has been extensively studied when the agents have full knowledge of their preferences. Recently, growing efforts have been devoted to *learning* in stable matching markets with unknown preferences, and through interactions between the agents (Liu et al., 2020; 2021; Basu et al., 2021; Jagadeesan et al., 2021; Etesami and Srikant, 2025; Shah et al., 2024b;a). Notably, (Etesami and Srikant, 2025; Shah et al., 2024b;a) also took a *game-theoretic* perspective, by developing learning dynamics for finding matchings in a decentralized, uncoordinated fashion. However, one key difference is that the learning agents (*e.g.,* the proposers or the platform) can still receive *numeric* feedback of the utilities each round, based on the matching result, while in our model, they can only receive the ranking feedback. Moreover, the learning dynamics in (Etesami and Srikant, 2025; Shah et al., 2024b;a) were specific to the matching model, while ours aim to address general normal-form games.

Recent Work by Maran, Bacchiocchi, Stradi, Castiglioni, Gatti, and Restelli (2024). The work closest to ours is the recent one by Maran et al. (2024), which studied multi-armed bandits with ranking feedback, also under the standard (external-)regret metric. Different from the ranking model in dueling bandits, the model of Maran et al. (2024) is based on time-average utilities, a setting also considered in our paper. More importantly, 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/online* and game-theoretic settings, with *both* instantaneous and time-average utility-based rankings. Furthermore, the ranking model in Maran et al. (2024) corresponds to the case of $\tau \to 0^+$ in our framework. Finally and notably, Maran et al. (2024) also provided a hardness result for the adversarial bandit setting (with $\tau \to 0^+$ in our framework), while our hardness results (with different hard instances) are stronger in the sense that they allow a wider range of τ for the bandit setting, and also cover the full-information setting (cf. Table 1).

B. Additional Notation and Preliminaries

Notation. For any integer $N > 0$, we define $[N] \coloneqq \{1, ..., N\}$ 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 i^{th} element of the vector. For any discrete set

$\mathcal{S}$, let $|\mathcal{S}|$ denote its cardinality, $\Delta^{\mathcal{S}} := \left\{\boldsymbol{x} \in \mathbb{R}^{\mathcal{S}} : \sum_{s\in\mathcal{S}} x_s = 1,\ x_s \geq 0 \text{ for all } s \in \mathcal{S}\right\}$ be the probability simplex over $\mathcal{S}$, and $\mathbf{1}(\mathcal{S})$ be an all-one vector with each index being elements in $\mathcal{S}$. For any ordered discrete set $\mathcal{S}$, we use $\mathbb{R}^{\mathcal{S}}$ to denote the $|\mathcal{S}|$ dimensional real space, where the $s^{th} \in \mathcal{S}$ element of any $\boldsymbol{x} \in \mathbb{R}^{\mathcal{S}}$ is denoted 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. 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 $\mathrm{sig}(x) := \frac{\exp(x)}{1+\exp(x)} : \mathbb{R} \to \mathbb{R}$ to denote the logistic function.

B.1. Normal-Form Games

An N-player normal-form game 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 $i \in [N]$; $\mathcal{U}_i\colon \bigtimes_{i=1}^N \mathcal{A}_i \to [-1, 1]$ ($\bigtimes$ denotes 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 the 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 an 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 B.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(n exact) CCE.

B.2. Equilibrium Computation with Ranking Feedback

There is a mediator (platform) in the game that computes strategies for the players, (*e.g.,* Uber recommends the candidate drivers and users to each other), 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 $u_i^{\boldsymbol{\pi}}(a_i) := \sum_{\boldsymbol{a}'\in\bigtimes_{j=1}^N \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}') \mathbb{1}(a_i' = a_i) \prod_{j\neq i} \pi_j(a_j')$. 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, and in different settings proceed differently as follows:

- **Full-information setting.** *All* the actions of each player $i \in [N]$ can be evaluated and ranked at each timestep t based on some utility vector, which is $\boldsymbol{u}_i^{\boldsymbol{\pi}^{(t)}}$ under **InstUtil Rank** and $\boldsymbol{u}_{\mathrm{avg}}^{(t)} := \frac{1}{t}\sum_{s=1}^t \boldsymbol{u}_i^{\boldsymbol{\pi}^{(s)}}$ under **AvgUtil Rank**, where $\boldsymbol{\pi}^{(t)} = \left(\pi_1^{(t)}, \ldots, \pi_N^{(t)}\right)$ is the strategy profile at timestep t.

- **Bandit setting.** For each player $i \in [N]$, only the K actions in $o_i^{(t)}$ that are proposed at timestep t will be evaluated and ranked, with the associated elements in some utility vector. Specifically, under **InstUtil Rank**, $\widehat{\boldsymbol{u}}_i^{(t)}$ defined below will be used: for each $a_i \in \mathcal{A}_i$

$$\widehat{u}_i^{(t)}(a_i) := \frac{1}{|o_{-i}^{(t)}|} \sum_{a_{-i}\in o_{-i}^{(t)}} \sum_{\boldsymbol{a}'\in\bigtimes_{j=1}^N \mathcal{A}_j} \mathcal{U}_i(\boldsymbol{a}') \mathbb{1}(a_i' = a_i) \mathbb{1}\left(a_{-i}' = a_{-i}\right);$$

under **AvgUtil Rank**, the corresponding empirical average utility is as computed in (3.2), with the $\boldsymbol{u}^{(s)}$ therein being replaced by the $\widehat{\boldsymbol{u}}_i^{(s)}$ above. As in the online setting, we assume that the actions are proposed in an *unbiased* way, *i.e.,* $\mathbb{E}\left[\frac{\sum_{a_{-i}\in o_{-i}^{(t)}} \mathcal{U}_i(a_i,a_{-i})}{|o_{-i}^{(t)}|}\right] = \left\langle \mathcal{U}_i(a_i,\cdot), \pi_{-i}^{(t)} \right\rangle$, for all $a_i \in \mathcal{A}_i$. In other words, $\widehat{\boldsymbol{u}}_i^{(t)}$ is an unbiased estimate of $\boldsymbol{u}_i^{\boldsymbol{\pi}^{(t)}}$.

The process will be repeated until the mediator finds an (approximate) equilibrium of the game, which is the average of the joint strategy over all timesteps.

C. Experiments

We evaluate our algorithms in two-player general-sum games with randomly generated utility, under all the combinations of full-information and bandit settings, as well as the **InstUtil Rank** and **AvgUtil Rank** feedback types. The CCE approximation ϵ for games with different parameters is provided in the following figures, with a 95% confidence interval. All the experiments are conducted on 37 cores of the Intel(R) Xeon(R) Platinum 8260 CPU @2.40 GH.

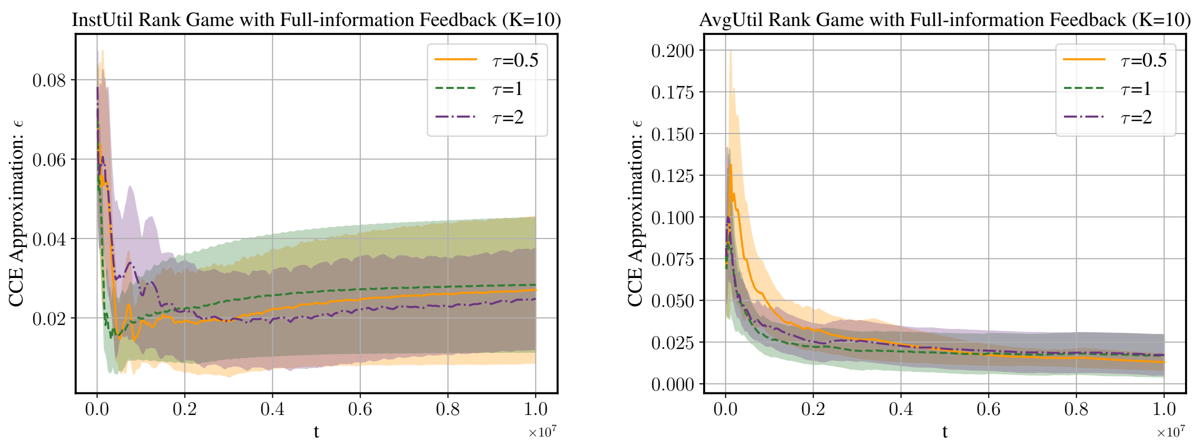

Figure 2: The exploitability for the full-information setting under both **InstUtil Rank** and **AvgUtil Rank** feedback. The performance is tested under different temperatures τ. Each parameter combination is tested 10 times with different random seeds.

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

To select 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 the 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.

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 5.2 and Theorem 6.3.

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

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

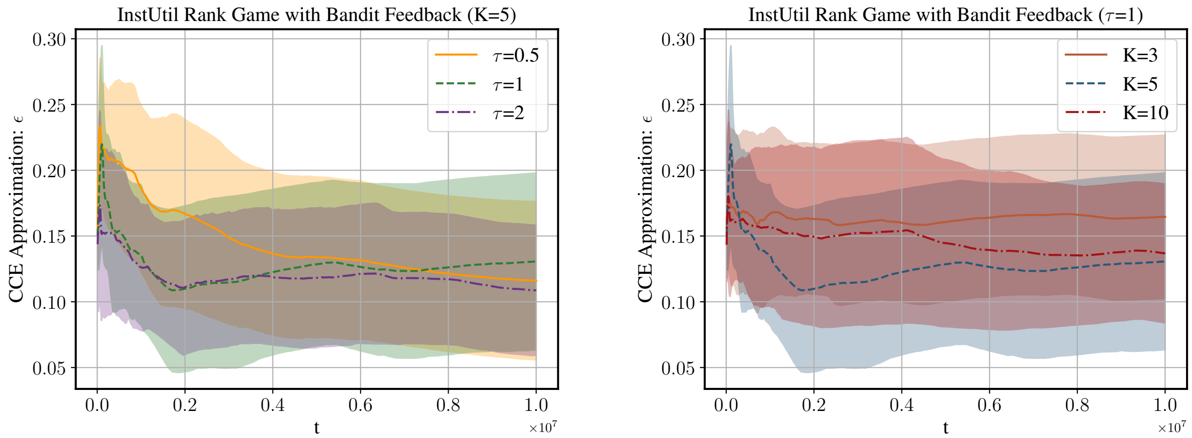

Figure 3: The exploitability for the bandit feedback setting under **InstUtil Rank** feedback. The performance is tested under different temperatures τ and sampled action size K. Each parameter combination is tested 10 times with different random seeds.

D. Proof of Section 4

Note that $\mathbb{E}\left[R^{(T)}\right] \geq \mathbb{E}\left[R^{(T),\text{external}}\right]$ by definition. Therefore, in the rest of this section, we will focus on showing the lower bound of $\mathbb{E}\left[R^{(T),\text{external}}\right]$.

D.1. Proof of Theorem 4.1

Theorem 4.1. *Consider* **InstUtil Rank**. *For any $T > 0$, temperature $0 < \tau \leq 0.1$, and online learning algorithm, there exists a sequence of utilities $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}$ such that $\min\left\{\mathbb{E}\left[R^{(T),\text{external}}\right], \mathbb{E}\left[R^{(T)}\right]\right\} \geq \Omega(T)$ 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 $(0.1, 0)$. Recall $\text{sig}(x)\colon \mathbb{R} \to \mathbb{R} := \frac{\exp(x)}{1+\exp(x)}$ is 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

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

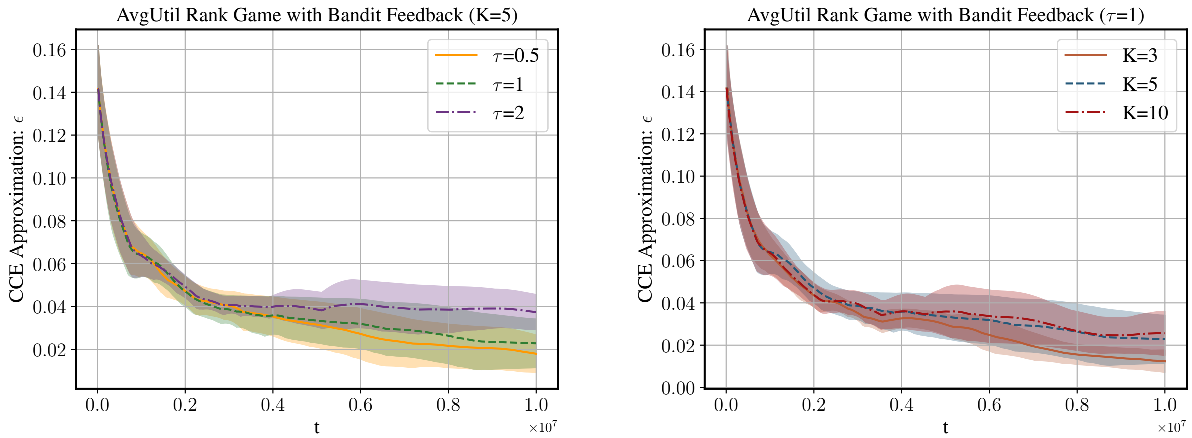

Figure 4: The exploitability for the bandit feedback setting under **AvgUtil Rank** feedback. The performance is tested under different temperatures τ and sampled action size K. Each parameter combination is tested 10 times with different random seeds.

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 process of generating both instances. Moreover,

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

The first term $\mathbb{P}\left(\sigma^{(1)},\ldots,\sigma^{(t-1)}\right)$ is equal in the two instances according to the discussion above, and the second term $\mathbb{E}\left[\pi^{(t)} \mid \sigma^{(1)},\ldots,\sigma^{(t-1)}\right]$ 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. $\square$

D.2. Proof of Theorem 4.3

Theorem 4.3. *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 utilities $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T'}$ such that $\mathbb{E}\left[R^{(T'),\text{external}}\right] \geq \widetilde{\Omega}\left(T'\right)$. 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 $\tau \leq \mathcal{O}(\frac{1}{T\log T})$ can be reduced to $\tau \to 0^+$.

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

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

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

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

Lemma D.1. *Consider* (D.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 D.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 establish 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, action b is the best action in all the observations. Similar to Lemma D.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 $\tau \leq \mathcal{O}\left(\frac{1}{T \log T}\right)$, from the construction above, the difference between the *cumulative* utility of the actions is always 0.5. By definition of **AvgUtil Rank**, $1 - \text{sig}\left(\mathcal{O}\left(\frac{0.5}{T\tau}\right)\right) \leq \mathcal{O}\left(\frac{1}{T}\right)$. 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 the case with $\tau \to 0^+$. $\square$

Lemma D.1. *Consider* (D.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 in this online learning setting, the strategy $\pi^{(t)}$ is determined by $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(t-1)}$. Therefore, in all the sequences in the action-a construction, since action a is the best in all the observations, for any two sequences $k_1 \leq k_2$, the expectation of the strategies is 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 the 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 $x^{(4)}, x^{(5)}, \ldots, x^{(7)}$ 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 $\frac{K}{2(K+1)} = \frac{1}{2} - \frac{1}{2K+2}$. □

D.3. Proof of Theorem 4.4

Theorem 4.4. *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 utilities $\left(\boldsymbol{u}^{(t)}\right)_{t=1}^{4T}$ such that $\mathbb{E}\left[R^{(4T)}\right] \geq \Omega\left(T\right)$. 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 to achieve sublinear expected regret, it must propose action a^1 for at least $0.9T$ timesteps during the first phase with probability at least $\frac{1}{2}$, since otherwise the expected external regret in the first phase is linear. During the second phase, it must propose action a^2 for at least $\frac{0.2T-0.1T}{0.2} = \frac{T}{2}$ timesteps due to the same reason with probability at least $\frac{1}{4}$. Then, at the end of the second phase, with probability at least $\frac{1}{4}$, $u_{\text{empirical}}^{(2T)}(a^2) - u_{\text{empirical}}^{(2T)}(a^1) \geq \frac{0.5T \cdot 0.2}{0.1T+0.5T} - 0.1 = \frac{1}{15}$.

Then, in the third phase of Instance 1, the algorithm needs to propose a^2 for at least $\frac{0.2T+2T-0.2T-0.1T}{1} = 1.9T$ timesteps 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, $u_{\text{empirical}}^{(4T)}(a^2) - u_{\text{empirical}}^{(4T)}(a^1) \geq \frac{0.5T \cdot 0.2}{0.1T+0.5T} - \frac{0.9T \cdot 0.1 + 0.1T \cdot 0.4}{0.9T+0.1T} \geq 0.03$. Therefore, when $\tau \to 0^+$, the observations 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 $t > 2T$ in Instance 2, we have $u_{\text{empirical}}^{(t)}(a^2) - u_{\text{empirical}}^{(t)}(a^1) \geq u_{\text{empirical}}^{(4T)}(a^2) - u_{\text{empirical}}^{(4T)}(a^1) \geq 0.03$. 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. □

E. Proof of Theorem 5.1

In this section, we proved the high probability bound for the utility estimation error, and with that, we gave the regret upper bound of our algorithm under **InstUtil Rank** feedback. Next, we will introduce the key lemma we used for utility estimation, Lemma E.1, which shows that the ranking of K actions can be decomposed into pair-wise rankings.

Algorithm 1 Utility Estimation with Action Permutations: Estimate$\left(\left\{\sigma^{(s)}\right\}_{s=1}^{m'}\right)$

Input: A set consisting of m' permutations of actions : $\left\{\sigma^{(s)}\right\}_{s=1}^{m'}$ with $|\sigma^{(s)}| = K$ for all $s \in [m']$, 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 $\mathrm{sig}(\cdot)$. 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 $\widetilde{\boldsymbol{u}} = \left(\widetilde{u}(a^1), \widetilde{u}(a^2), \ldots, \widetilde{u}\left(a^{|\mathcal{A}|-1}\right), 0\right)$

E.1. Pair-wise Utility Estimation

Lemma E.1 below shows that when the number of proposed actions $K > 2$, for any two actions $a \neq b \in o^{(t)}$, the proportion of pairs that action a appears before b in $o^{(t)}$, is equal to sig $\left(\frac{u^{(t)}(a) - u^{(t)}(b)}{\tau}\right)$ in expectation. In other words, the expected proportion is equal to the probability of the permutation (a, b) occurring when only proposing a, b.

Lemma E.1. *Let $\#_{\mathcal{S}}(a) \coloneqq \sum_{a' \in \mathcal{S}} \mathbb{1}(a' = a)$ 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*

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

The expectation is taken over the distribution of σ under the ranking model (PL).

The proof can be found in Appendix E.2. 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 enumerating all possible action pairs. Therefore, to estimate $u^{(t)}(a^j)$ for some $a^j \in \mathcal{A}$, we will first construct an unbiased estimator of 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 sig $\left(\frac{u^{(s)}(a^j)}{\tau}\right)$. Then, by Hoeffding's inequality and the monotonicity of the logistic function $\mathrm{sig}(\cdot)$, with high probability, the mean of the logistic function estimators will be bounded between the minimum and maximum of $\left\{\mathrm{sig}\left(\frac{u^{(s)}(a^j)}{\tau}\right)\right\}_{s=t-m+1}^{t}$. By Assumption 4.2, since the utility vectors are changing slowly, that mean can be shown close to sig $\left(\frac{u^{(t)}(a^j)}{\tau}\right)$. With a good estimate of sig $\left(\frac{u^{(t)}(a^j)}{\tau}\right)$, we can then take an inverse of $\mathrm{sig}(\cdot)$ to estimate $u^{(t)}(a^j)$. This estimation algorithm is summarized in Algorithm 1 and analyzed in Theorem 5.1 below.

In the following, we will prove Theorem 5.1, which gives the estimation error bound of the utility vector for each timestep.

Theorem 5.1. *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 $\widetilde{\boldsymbol{u}}^{(t)} = \operatorname{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=t-m'+1}^{t}\right)$. 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,

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

Proof. Due to the symmetry of timesteps, we will only prove Theorem 5.1 for $t = m'$ for notational simplicity.

For any $j \in [|\mathcal{A}|-1]$, we assume that the probability for action a^j being chosen at each timestep is at least p. Let the number of action pairs $\left(a^j, a^{|\mathcal{A}|}\right)$ 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 $\frac{\exp\left(\frac{1}{\tau}u^{(m')}(a^j)\right)}{\exp\left(\frac{1}{\tau}u^{(m')}(a^j)\right)+\exp\left(\frac{1}{\tau}u^{(m')}(|\mathcal{A}_i|)\right)} = \frac{\exp\left(\frac{1}{\tau}u^{(m')}(a^j)\right)}{\exp\left(\frac{1}{\tau}u^{(m')}(a^j)\right)+1}$. We define

$$S_{m_1} \coloneqq \sum_{s=1}^{m'} \mathbb{1}\left(a^j, a^{|\mathcal{A}|} \in \sigma^{(s)}\right) \frac{\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\mathbb{1}\left(\sigma(k_1)=a\right)\cdot\mathbb{1}\left(\sigma(k_2)=b\right)}{\#_{\sigma^{(s)}}(a^j)\#_{\sigma^{(s)}}(a^{|\mathcal{A}|})}.$$

Due to the monotonicity of the function sig,

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

where $\widetilde{u}^{(m')}(a^j)$ is the estimation of $u^{(m')}(a^j)$. Then also by Hoeffding's Inequality, we have that with probability at least $1-\delta$,

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

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

$$\operatorname{sig}\left(\frac{1}{\tau}u^{(m'),*}(a^j)\right) = \frac{1}{m_1}\sum_{s=1}^{m'}\mathbb{1}\left(a^j, a^{|\mathcal{A}|} \in \sigma^{(s)}\right)\cdot\operatorname{sig}\left(\frac{1}{\tau}u^{(s)}(a^j)\right).$$

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

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

Lemma E.2 in the following shows that $\widetilde{u}^{(m')}(a^j)$ 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 $\boldsymbol{u}^{(m')}$.

Lemma E.2. *Let $x_1, \ldots, x_n \in [-1, 1]$ and $\mathrm{sig}_{\mathrm{avg}} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{sig}(x_i)$, we have*

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

The proof is postponed to Appendix E.2.

By Lemma E.2, we have that $u^{(m'),*} \in [-1, 1]$,

$$\mathrm{Proj}_{[\mathrm{sig}(-\frac{1}{\tau}),\mathrm{sig}(\frac{1}{\tau})]}\left(\mathrm{sig}\left(\frac{1}{\tau}u^{(m'),*}(a^j)\right)\right) = \mathrm{sig}\left(\frac{1}{\tau}u^{(m'),*}(a^j)\right).$$

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

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

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

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

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

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

By Lemma E.2, 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

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

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

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

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

$$\left|\widetilde{u}^{(m')}(a^j) - u^{(m')}(a^j)\right| \le \tau\left(e^{\frac{1}{\tau}}+1\right)^2\sqrt{\frac{1}{m'p^2}\log\left(\frac{2}{\delta}\right)} + \sum_{s=1}^{m'-1}\left|u^{(s+1)}(a^j) - u^{(s)}(a^j)\right|.$$

After estimating the utility of each action, we have

$$\left\|\widetilde{\boldsymbol{u}}^{(m')}-\boldsymbol{u}^{(m')}\right\|_\infty \leq \frac{\tau\left(e^{\frac{1}{\tau}}+1\right)^2}{p}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{m'}}+\sum_{s=1}^{m'-1}\left\|\boldsymbol{u}^{(s+1)}-\boldsymbol{u}^{(s)}\right\|_\infty . \qquad \square$$

Remark E.3. *Due to the monotonicity of the logistic function* sig*, the following two projections on* sig $\left(\frac{x}{\tau}\right)$ *are equivalent:*

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

E.2. Omitted Proofs

Lemma E.1. *Let $\#_{\mathcal{S}}\left(a\right) \coloneqq \sum_{a'\in\mathcal{S}}\mathbb{1}\left(a'=a\right)$ represent the number of elements in a multiset $\mathcal{S}$ that are equal to $a\in\mathcal{A}$. For any utility vector $\mathbf{u}$, 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*

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

The expectation is taken over the distribution of σ under the ranking model (PL).

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 PL model, the probability that action a ranks before action b is that

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

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

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

Recall that $\Sigma(\mathcal{S})$ denotes the set that contains all the permutations of the elements in $\mathcal{S}$. Hence, we have

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

We deal with (E.2) first:

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

Similarly, for (E.3), we have

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

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

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

Next, we will use induction to show that for any actions $a\neq b\in\mathcal{S}$, the following holds:

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

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

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

Lemma E.4. *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*

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

The proof is presented later in this section.

Induction step. When (E.5) holds for any $\mathcal{S}$ with $|\mathcal{S}| = K-1$. Then, we will show that it still holds for any $\mathcal{S}$ with $|\mathcal{S}| = K$. By Lemma E.4, (E.2) is equal to

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

Similarly, (E.3) is equal to

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

and (E.4) is equal to

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

Lastly, by summing them up, we have

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

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,

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

and we complete the induction. □

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

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

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

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

□

Lemma E.2. *Let $x_1,\ldots,x_n\in[-1,1]$ and $\mathrm{sig}_{\mathrm{avg}}\coloneqq\frac{1}{n}\sum_{i=1}^{n}\mathrm{sig}(x_i)$, we have*

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

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 $x_1\leq x_2\leq\cdots\leq x_n$. Thus, $\mathrm{sig}(x_1)\leq\mathrm{sig}_{\mathrm{avg}}\leq\mathrm{sig}(x_n)$.

Since $\mathrm{sig}(x)$ is monotonic and continuous, there exists 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

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

□

F. Proof of Theorem 5.2 (Full-Information)

In this section, we prove the regret upper bound under **InstUtil Rank** and full-information feedback.

Theorem F.1 (Formal version of Theorem 5.2 (Full-Information))**.** *Consider Algorithm 2 and full-information feedback. For any $\delta\in(0,1)$, $T>0$, and any no-regret learning algorithm with numeric utility feedback,* 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*

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

Proof. By Theorem 5.1, we have

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

When $t\geq m$, the estimation error between $\widetilde{\boldsymbol{u}}^{(t)}$ and $\boldsymbol{u}^{(t)}$ is given by Theorem 5.1. When $t<m$, it becomes trivial that the estimation error satisfies

$$\left\|\boldsymbol{u}^{(t)}-\widetilde{\boldsymbol{u}}^{(t)}\right\|_\infty\leq2.$$

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 $1-\delta$,

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

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. □

G. Proof of Theorem 5.2 (Bandit)

In this section, we prove the regret upper bound under **InstUtil Rank** and bandit feedback.

Theorem G.1 (Formal version of Theorem 5.2 (Bandit))**.** *Consider Algorithm 2 and bandit feedback. For any $\delta \in (0,1)$, $T>0$, and any no-regret learning algorithm with numeric utility feedback,* 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*

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

Proof. Firstly, we define:

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

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

Then,

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

Note that $\spadesuit$ can be bounded by bounding $\left\|\widetilde{\boldsymbol{u}}^{(t)}-\boldsymbol{u}^{(t)}\right\|_\infty$ as in Appendix F. $\clubsuit$ is sublinear by the definition of Alg. Next, we will introduce lemmas that individually bound $\heartsuit$, $\diamondsuit$. The proofs are postponed to Appendices G.1 and G.2.

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

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

Lemma G.3 ($\diamondsuit$)**.** *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:*

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

With Lemma G.2 and Lemma G.3, under the conditions in Theorem 5.2, by letting Theorem 5.1 hold with probability $1-\frac{\delta}{T}$ at each timestep, with probability at least $(1-\delta)$, the regret satisfies

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

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

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

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

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

□

G.1. Bounding ♥

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

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

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

Proof. Let

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

By our algorithm design, 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.

Due to the bounds of $\left|\frac{1}{K}\sum_{a\in o^{(t)}} u^{(t)}(a)\right| \leq 1$, $\left|\left\langle \boldsymbol{u}^{(t)}, \pi^{(t)}\right\rangle\right| \leq 1$, we have

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

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 G.4 (Azuma-Hoeffding inequality)**.** *For any martingale difference sequence $Y_1, \ldots, Y_n$ such that $\forall j \in [n], a_j \leq Y_j \leq b_j$, the following holds for any $w \geq 0$.*

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

Then by Theorem G.4, with probability at least $1-\delta$

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

□

G.2. Bounding ♦

♦ can be bounded by $\mathcal{O}(\gamma T)$ by definition of $\pi^{(t)}$ in Algorithm 2.

Lemma G.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:*

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

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

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

□

H. Proof of Theorem 6.2

Theorem H.1 (Formal version of Theorem 6.2)**.** *Consider* **AvgUtil Rank** *with full-information feedback and Algorithm 3. For any $\delta \in (0,1)$, $T > 0$, and any no-regret learning algorithm with numeric utility feedback* Alg *that satisfies Assumption 6.1,* *with probability at least* $(1-\delta)$*, by choosing* $m = 2\frac{|\mathcal{A}|^4}{K^4}T^{\frac{2}{3}}\log\left(\frac{2T}{\delta}\right)$, $R^{(T),\mathrm{external}}$ *satisfies*

$$\begin{aligned} R^{(T),\mathrm{external}} \leq & R^{(T),\mathrm{external}}\left(\mathrm{Alg}, \left(\boldsymbol{u}^{(t)}\right)_{t=1}^{T}\right) + L\tau K\left(e^{\frac{1}{\tau}}+1\right)^2 L^{\frac{1}{3}}T + 4\frac{|\mathcal{A}|^4}{K^4}L^{-\frac{2}{3}}\log\left(\frac{2T}{\delta}\right) \\ & + 4\frac{|\mathcal{A}|^9}{K^8}L^{-\frac{1}{3}}\log\left(\frac{2T}{\delta}\right)(\log T+1) + 4\frac{|\mathcal{A}|^5}{K^4}\log\left(\frac{2T}{\delta}\right)L^{\frac{1}{3}}T. \end{aligned} \tag{H.1}$$

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

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

By Assumption 6.1 and Theorem 5.1, for any $t \geq m$, with probability at least $1-\delta$ we have

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

Then,

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

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

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

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

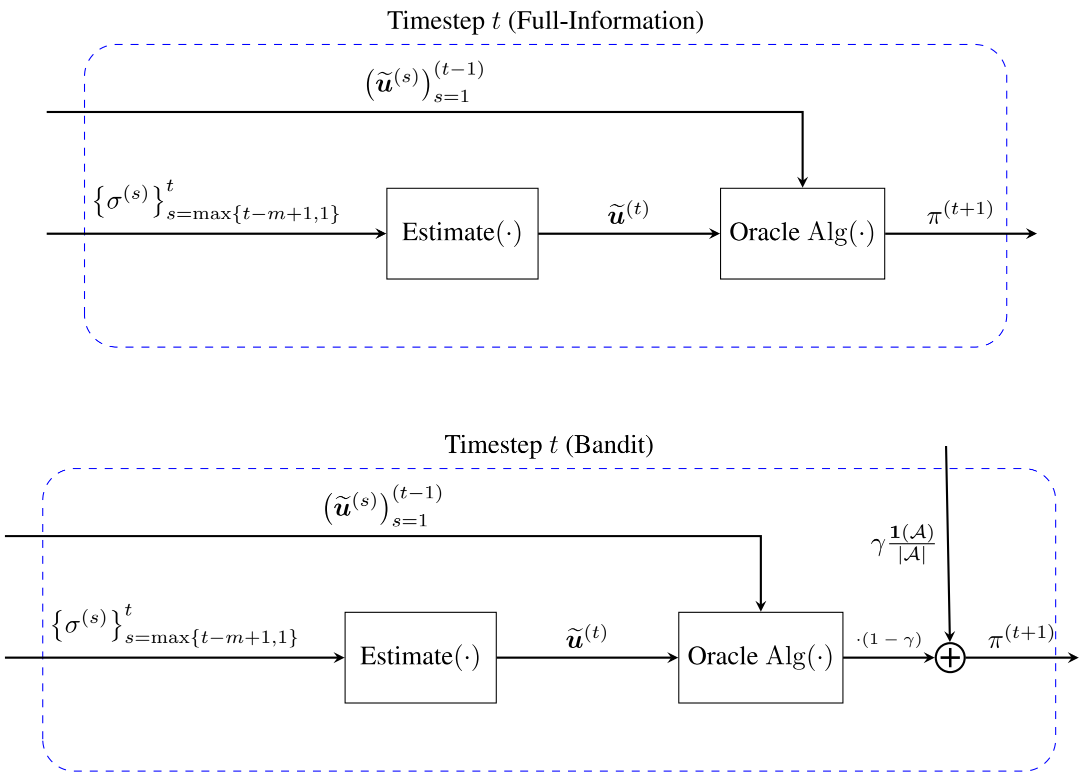

Figure 5: The diagram of Algorithm 2 with **InstUtil Rank** under full-information feedback (top) and bandit feedback (bottom). $\bigoplus$ represents the addition of $(1-\gamma)$ times the output the Alg and γ times a uniform distribution over $\mathcal{A}$.

I. Algorithms and Diagrams

In this section, we present the algorithms' diagrams and pseudo-code of learning with **InstUtil Rank** and **AvgUtil Rank** individually.

I.1. The Algorithm and Diagram for InstUtil Rank

We present the diagram and the algorithm pseudo-code of learning with **InstUtil Rank**: Figure 5 and Algorithm 2.

I.2. The Algorithm and Diagram for AvgUtil Rank

We present the diagram and the algorithm pseudo-code of learning with **AvgUtil Rank**: Figure 6 and Algorithm 3.

J. Proof of Theorem 6.3

Theorem J.1 (Formal version of Theorem 6.3)**.** *Consider* **AvgUtil Rank** *with bandit feedback and Algorithm 3. For any* $\delta \in (0,1)$, $T>0$, *and any no-regret learning algorithm with numeric utility feedback* Alg *that satisfies Assumption 6.1, 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*

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

Algorithm 2 Online Learning with **InstUtil Rank** Feedback

Input: Action space $\mathcal{A}$, any full-information no-regret learning algorithm Alg with numeric utility 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.
  $\widetilde{\boldsymbol{u}}^{(t)} = \mathrm{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=\max\{t-m+1,1\}}^{t}\right)$ by calling Algorithm 1.
  if Full-information setting **then**
    $\pi^{(t+1)} \leftarrow \mathrm{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right)$.
  else if Bandit setting **then**
    $\pi^{(t+1)} \leftarrow (1-\gamma)\mathrm{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(s)}\right)_{s=1}^{t}\right) + \gamma\frac{\mathbf{1}(\mathcal{A})}{|\mathcal{A}|}$.
  end if
end for

Algorithm 3 Online Learning with **AvgUtil Rank** Feedback

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**
    $\widetilde{\boldsymbol{u}}^{(t)}_{\text{avg}} = \mathrm{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=\max\{t-m+1,1\}}^{t}\right)$ by calling Algorithm 1.
    $\pi^{(t+1)} \leftarrow \mathrm{Alg}\left(\left(\widetilde{\boldsymbol{u}}^{(t)}_{\text{avg}}\right)_{s=1}^{t}\right)$, *i.e.,* the strategy generated by Alg by setting all utility vectors from timestep 1 to t as $\widetilde{\boldsymbol{u}}^{(t)}_{\text{avg}}$.
  else if Bandit setting **then**
    $\widetilde{\boldsymbol{u}}^{(t)}_{\text{empirical}} = \mathrm{Estimate}\left(\left\{\sigma^{(s)}\right\}_{s=\max\{t-m+1,1\}}^{t}\right)$ by calling Algorithm 1.
    Let $n^{(t)}(a) := \sum_{s=1}^{t} \#_{o^{(s)}}(a)$ for any $a \in \mathcal{A}$ as the number of times action a has been proposed up to timestep t. Then, the estimated average utility is

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

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

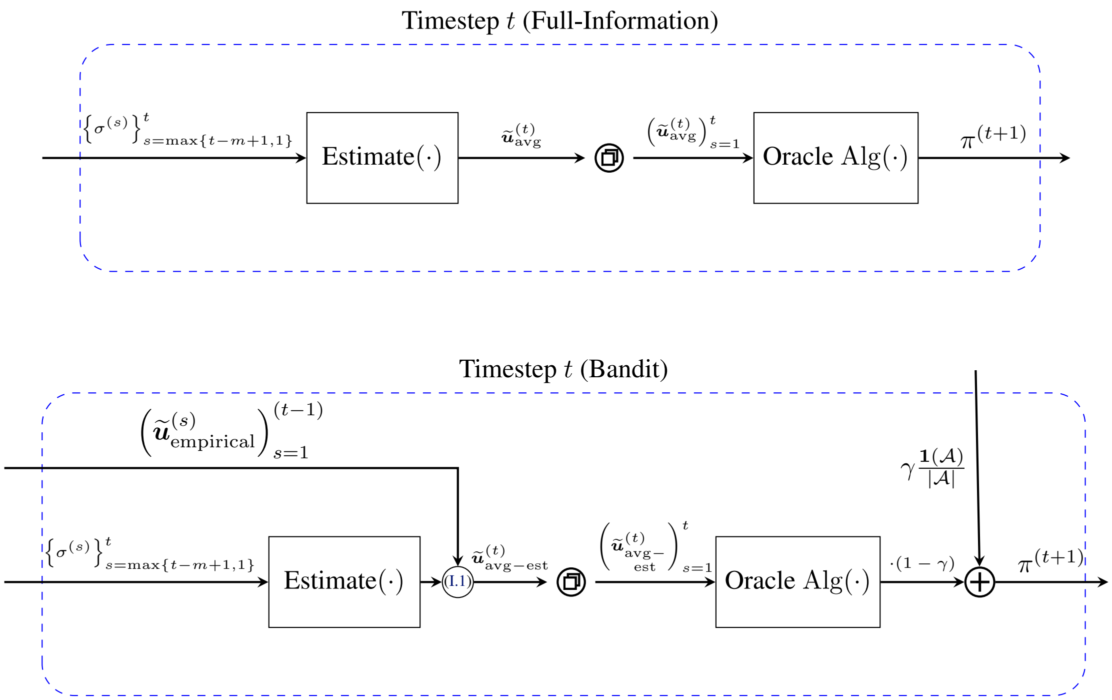

Figure 6: The diagram of Algorithm 3 with **AvgUtil Rank** under full-information feedback (top) and bandit feedback (bottom). ⧉ represents copying the estimated utility vector for t times. $\oplus$ represents the addition of $(1-\gamma)$ times the output the Alg and γ times a uniform distribution over $\mathcal{A}$.

where

$$
\begin{aligned}
C_\delta :=& \frac{|\mathcal{A}| \log\left(\frac{3|\mathcal{A}|T}{\delta}\right)}{\gamma} \\
W^{(T)} :=& 4C_\delta\left(\log T+1\right)\left(\frac{T^2}{M}\frac{\tau|\mathcal{A}|\left(e^{\frac{1}{\tau}}+1\right)^2}{\gamma}\sqrt{\frac{\log\left(\frac{6T}{\delta}\right)}{m}}+16KC_\delta\frac{m}{M}T\right) \\
&+M\left(\log T+1\right)P^{(T)}+2M\left(\log T+2\right).
\end{aligned}
$$

Proof. In the first part of the proof, we will bound $\left\|\widetilde{\boldsymbol{u}}_{\text{empirical}}^{(t)}-\boldsymbol{u}_{\text{empirical}}^{(t)}\right\|_\infty$. According to Theorem 5.1 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

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

Let $\#_{o^{(t)}}(a)$ be the number of action $a \in \mathcal{A}$ being 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 $\sum_{s=1}^{t}\#_{o^{(s)}}(a)$ for any timestep t.

Lemma J.2. *Consider the case when actions are 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 $t \geq \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p}$:*

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

Proof. For any $t \geq \frac{\log\left(\frac{|\mathcal{A}|T}{\delta}\right)}{p}$ and action $a \in \mathcal{A}$, the probability of (J.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 $1-\delta$, (J.2) holds for any $t \in [T]$ and any action $a \in \mathcal{A}$. $\square$

For notational simplicity, let $C_\delta := \frac{|\mathcal{A}|\log\left(\frac{3|\mathcal{A}|T}{\delta}\right)}{\gamma}$. According to Lemma J.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 $4K\frac{C_\delta}{t+1} \geq 4K \geq 2$ and all utilities are bounded in $[-1,1]$. Finally, by Theorem 5.1 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 $n^{(t)}(a) := \sum_{s=1}^{t}\#_{o^{(s)}}(a)$ 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 $t \geq M$, we define

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

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.

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

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

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

According to Lemma J.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

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

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 $\clubsuit$ can be bounded by

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

For $\spadesuit$, we have

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

When $n^{(s\cdot M)}(a)-n^{((s-1)M)}(a)>0$, both $\frac{u^{(s\cdot M)}_{\text{empirical}}(a)n^{(s\cdot M)}(a)-u^{((s-1)M)}_{\text{empirical}}(a)n^{((s-1)M)}(a)}{n^{(s\cdot M)}(a)-n^{((s-1)M)}(a)}$ 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

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

By combining all the pieces together, we have

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

Then,

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

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,

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

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

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

Further, by Assumption 6.1, we have

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

By combining all the pieces together, we have

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

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

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

It is 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 G.2, with probability at least $1-\frac{\delta}{3}$, we have

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

By a union bound argument, we complete the proof. □

K. Proof of Section 7

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

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

where $A = \max_j |\mathcal{A}_j|$.

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

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

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

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,

$$\left\|u_i^{(t)} - u_i^{(t-1)}\right\| \leq \sqrt{|\mathcal{A}_i|}\left\|\pi_{-i}^{(t)} - \pi_{-i}^{(t-1)}\right\|_1 \leq \sqrt{|\mathcal{A}_i|}\prod_{j=1}^{N}|\mathcal{A}_j|\sum_{j'\neq i}\left\|\pi_{j'}^{(t)} - \pi_{j'}^{(t-1)}\right\|_1.$$ □

K.1. Proof of Theorem 7.2 and Theorem 7.3

Before proving Theorem 7.2 and Theorem 7.3, we will show that when Assumption 7.1 is satisfied, the strategy variation is bounded.

Lemma K.2. *Suppose Assumption 7.1 is satisfied. For both full-information and bandit settings, Algorithm 2 satisfies the following,*

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

Suppose Assumption 6.1 is also satisfied, then the following holds for Algorithm 3 in the 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 the full-information setting, the proof simply follows from the fact that $\widetilde{\boldsymbol{u}}^{(t)} \in [-1,1]^{\mathcal{A}}$ and Assumption 7.1.

For Algorithm 2 and the bandit setting, we have

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

Thus, we can conclude the proof.

For Algorithm 3 and the bandit setting, for any $t \not\equiv 0 \pmod M$, we have $\widetilde{\boldsymbol{u}}^{(t)}_{\mathrm{avg-est}} = \widetilde{\boldsymbol{u}}^{(t-1)}_{\mathrm{avg-est}}$. Therefore, by Assumption 7.1, we have

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

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

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

Therefore,

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

where (i) uses Assumption 6.1. Then, the accumulated variation of $\pi^{(t)}$ over time 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 of $t \in [T]$ satisfying $t \equiv 0 \pmod M$. $\square$

With Lemma K.2, we can prove that $R_i^{(T),\text{external}}$ is sublinear for any player $i \in [N]$ by Theorem 5.2, Theorem 6.2, and Theorem 6.3. Then, by the folklore result that no-external-regret learning leads to approximate CCE (Hart and Mas-Colell, 2000; Blum and Mansour, 2007), Theorem 7.2 and Theorem 7.3 are proved. ☐

Remark K.3. *With the hardness in Theorem 4.3, under* **AvgUtil Rank** *feedback, both of our no-regret result for the online setting and the equilibrium computation result for the game setting hold for a constant $\tau > 0$ (that cannot be arbitrarily small). However, we note that the equilibrium computation result may still 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, precisely 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 and Shapley, 1996; Sela, 1999; Berger, 2005) (with (slow) convergence rates (Robinson, 1951; Daskalakis and Pan, 2014; Abernethy et al., 2021)), despite that it fails to be no-regret in the online setting (Fudenberg and Levine, 1995; 1998).*

L. Properties of Follow-the-Regularized-Leader (FTRL)

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

Definition L.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*

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

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

Next, we will introduce the well-known no-regret learning algorithm of follow-the-regularized-leader (Hazan et al., 2016; Shalev-Shwartz et al., 2012).

Definition L.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 L.3. *For any c_0-strongly convex and differentiable function $\psi\colon \Delta^{\mathcal{A}} \to \mathbb{R}$, (FTRL) satisfies Assumption 6.1 and Assumption 7.1 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,

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

where (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 6.1 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 7.1 for (FTRL). $\square$