Settling the Minimax Regret in Online Learning to Rank with Top-k Feedback

by

Mingyuan Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science
(Honors Statistics)
in the University of Michigan
2018

Advisor:

Associate Professor Ambuj Tewari


THE UNIVERSITY OF MICHIGAN
ARTES · SCIENTIA · VERITAS
1817

2018

A C K N O W L E D G M E N T S

During my five years' undergraduate study at University of Michigan, I felt very lucky to meet and work with many great people, who helped me and influenced my life. I want to express my deepest gratitude to some of them.

First, I would like to thank Professor Ambuj Tewari, who introduced me with this project and advised my thesis, for his constant guidance and support. He is knowledgeable and is always willing to answer my questions. More importantly, he provided me with many sincere advices in both my study and my life. I could not thank him enough for his dedication and encouragement.

Second, I would like to thank all the professors who taught me. In particular, I want to recognize Professor Yaoyun Shi for the advice he provided for my graduate school applications, and Professor Gongjun Xu for the supervision of my independent study in Statistics. I also want to thank Stephen DeBacker and Gina Cornacchia, without whom my experience in Mathematics and Statistics would not have been so great.

Finally, I would like to thank my family and friends, for their constant support and encouragement, without whom I would not have been able to finish my degree smoothly.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

PL Pairwise Loss

DCG Discounted Cumulative Gain

P@n Precision@n Gain

AUC Area Under Curve

NDCG Normalized Discounted Cumulative Gain

AP Average Precision

ABSTRACT

Settling the Minimax Regret in Online Learning to Rank with Top-k Feedback

by

Mingyuan Zhang

Online learning to rank is a supervised machine learning problem where the rankers are trained on streaming data. Chaudhuri and Tewari (2017) developed a model for online learning to rank with highly limited feedback (top-k feedback), in both contextual setting and non-contextual setting.

In this thesis, we go deeper into non-contextual online learning to rank with top-k feedback, addressing some open problems posed by Chaudhuri and Tewari (2017). We provide a full characterization of minimax regret rates with the top k feedback model for all k for ranking measures Pairwise Loss, Discounted Cumulative Gain and Precision@n Gain. In addition, we give an efficient algorithm that achieves the minimax regret rate for Precision@n.

CHAPTER 1

Introduction

Learning to rank (Liu, 2011) is a supervised machine learning problem where we want to learn mappings that map a set of objects to a ranking. According, the output space in learning to rank problems consists of permutations of objects. Given true relevance scores of the objects, the accuracy of a ranked list is judged using ranking measures, such as Pairwise Loss (PL), Discounted Cumulative Gain (DCG), Precision@n Gain (P@n), and others. Most learning to rank algorithms are *offline*, i.e., they are designed to operate on the entire data in a single batch. However, interest in *online* algorithms, i.e., those that process the data incrementally, is rising due to a number of reasons. First, online algorithms often require less computation and storage. Second, many applications, especially on the web, produce ongoing streams of data making them excellent candidates for applying online algorithms. Third, basic online algorithms, such as the ones developed in this thesis, make excellent starting points for developing more sophisticated online algorithms that can deal with *non-stationarity*. Non-stationarity is a major problem on learning to rank settings since user preferences can easily change over time. Therefore, online learning to rank is a promising direction of research, in which rankers are updated on the fly as new data arrive.

The basic full feedback supervised learning setting assumes that the labeler, typically a human, provides the correct output for each example in the training dataset. Since the output in learning to rank problems is a permutation, it becomes practically impossible to get a fully specified permutation as a label from human labelers. Therefore, researchers

have looked into *weak supervision* or *partial feedback* settings where the correct permutation label is only partially revealed to the learning algorithm. For example, Chaudhuri and Tewari (2017) developed a model for online learning to rank with a particular case of partial feedback called *top-k feedback*. In this model, the online learning to rank problem is cast as an online partial-monitoring game (some other problems that can be cast as partial monitoring games are multi-armed bandits (Auer et al., 2003), and dynamic pricing (Kleinberg and Leighton, 2003)) between a learner and an oblivious adversary (who generates a sequence of outcomes before the game begins), played over T rounds. At each round, the learner outputs a ranking of objects whose quality with respect to the true relevance scores of the objects, is judged by some ranking measure. However, the learner receives highly limited feedback at the end of each round: only the relevance scores of the top k ranked objects are revealed to the learner. Here, $k \leq m$ and m is the number of objects.

The goal of the learner is to minimize its regret. The goal of regret analysis is to compute upper bounds on regret of explicit algorithms. If lower bounds on regret that match the upper bounds up to constants can be derived, then the *minimax regret* is identified, again up to constants. Previous work considered two settings: *non-contextual* (objects to be ranked are fixed) and *contextual* (objects to be ranked vary and get encoded as a context, typically in the form of a feature vector). In non-contextual setting, six ranking measures have been studied: PL, DCG, P@n, and their normalized versions Area Under Curve (AUC), Normalized Discounted Cumulative Gain (NDCG) and Average Precision (AP). Chaudhuri and Tewari (2017) showed that the minimax regret rates with the top k feedback model for PL, DCG and P@n are upper bounded by $O(T^{2/3})$ for all $1 \leq k \leq m$. In particular, for $k = 1$, the minimax regret rates for PL and DCG are $\Theta(T^{2/3})$. Moreover, for $k = 1$, the minimax regret rates for AUC, NDCG and AP are $\Theta(T)$. One of the open questions, as described in Chaudhuri and Tewari (2017), is to find the minimax regret rates for $k > 1$ for the six ranking measures.

It is worth noting that the top k feedback model is neither full feedback (where the

outcome is uniquely determined by the feedback) nor bandit feedback (where the loss is determined by the feedback); the model falls under the framework of so-called *partial monitoring* (Cesa-Bianchi et al., 2006). Recent advances in classification of finite partial-monitoring games have shown that the minimax regret of any such game is 0, $\Theta(T^{1/2})$, $\Theta(T^{2/3})$, or $\Theta(T)$ up to a logarithmic factor, and is governed by two important properties: *global observability* and *local observability* (Bartók et al., 2014). In particular, Bartók et al. (2014) gave an almost complete classification of all finite partial-monitoring games by identifying four regimes: trivial, easy, hard and hopeless games, which correspond to the four minimax regret rates mentioned before, respectively. What was left from the classification is the set of games in oblivious adversarial setting with degenerate actions which are never optimal themselves, but can provide useful information. Lattimore and Szepesvari (2018) finished the characterization.

Our contributions: In this thesis, we focus on the non-contextual setting of the top k feedback model and we assume that the adversary is oblivious. We establish the minimax regret rates for all $1 \leq k \leq m$ for ranking measures PL, DCG, and P@n, by showing global observability and local observability properties. In addition, we provide an algorithm based on the NEIGHBORHOODWATCH2 algorithm in Lattimore and Szepesvari (2018). Our algorithm achieves the minimax rate for P@n and has per-round time complexity polynomial in m (for any fixed n).

CHAPTER 2

Preliminaries

We use the problem formulation by Chaudhuri and Tewari (2017), Bartók et al. (2014) and Lattimore and Szepesvari (2018).

2.1 Notation and Problem Setup

Let $\{e_i\}$ denote the standard basis. Let $\mathbf{1}$ denote the vector of all ones.

The fixed m objects to be ranked are $[m] := \{1, ..., m\}$. The permutation σ maps from ranks to objects, and its inverse σ^{-1} maps from objects to ranks. Specifically, $\sigma(i) = j$ means that object j is ranked i and $\sigma^{-1}(i) = j$ means that object i is ranked j. The relevance vector $R \in \{0, 1, ..., n\}^m$ represents relevance for each object. $R(i)$, i-th component of R, is the relevance for object i. For $n = 1$, R is binary-graded. For $n > 1$, R is multi-graded. In this thesis, we only study binary relevance, i.e., $n = 1$.

The learner can choose from $m!$ actions $\{\sigma | \sigma : [m] \to [m] \text{ is bijective}\}$ while the adversary can choose from $(n+1)^m$ outcomes $\{R | R \in \{0, 1, ..., n\}^m\}$. We use subscript t exclusively to denote time t, so σ_t is the action the learner chooses at round t and R_t is the outcome the adversary chooses at round t.

In a game G, the learner and the adversary play over T rounds. We will consider an *oblivious* adversary who chooses all the relevance vectors R_t ahead of the game (but they are not revealed to the learner at that point). In each round t, the learner predicts a permutation (ranking) σ_t according to a (possibly randomized) strategy π. The performance

of σ_t is judged against R_t by some ranking (loss) measure RL. At the end of round t, only the relevance scores of the top k ranked objects $(R_t(\sigma_t(1)), ..., R_t(\sigma_t(k)))$ are revealed to the learner. Therefore, the learner knows neither R_t (as in the full information game) nor $RL(\sigma_t, R_t)$ (as in the bandit game). The goal of the learner is to minimize the expected regret (where the expectation is over any randomness in learner's moves σ_t) defined as the difference in the realized loss and the loss of the best fixed action in hindsight:

$$\mathcal{R}_T(\pi, R_1, ...R_T) := \mathbb{E}_{\sigma_1,...,\sigma_T}\Big[\sum_{t=1}^{T} RL(\sigma_t, R_t)\Big] - \min_{\sigma} \sum_{t=1}^{T} RL(\sigma, R_t) \tag{2.1}$$

When the ranking measure is a gain, we can always negate the gain function so that it becomes a loss function. The worse-case regret of a learner's strategy is its maximum regret over all choices of $R_1, ..., R_T$. The minimax regret is the minimum worse-case regret over all strategies of the learner:

$$\mathcal{R}^*_T(G) = \inf_{\pi} \max_{R_1,...,R_T} \mathcal{R}_T(\pi, R_1, ...R_T) \tag{2.2}$$

where π is the learner's strategy to generate $\sigma_1, ..., \sigma_T$.

2.2 Ranking Measures

We are interested in ranking measures that can be expressed in the form of $f(\sigma) \cdot R$ where $f : \mathbb{R}^m \to \mathbb{R}^m$, is composed of m copies of a univariate monotonically non-decreasing scalar-valued function $f^s : \mathbb{R} \to \mathbb{R}$. We say that f^s is monotonically non-decreasing iff $\sigma^{-1}(i) > \sigma^{-1}(j)$ implies $f^s(\sigma^{-1}(i)) \geq f^s(\sigma^{-1}(j))$. The monotonic non-increasing is defined analogously. Then, $f(\sigma)$ can be written as

$$f(\sigma) = [f^s(\sigma^{-1}(1)), ..., f^s(\sigma^{-1}(m))]$$

The definitions of ranking measures that we are going to study in this thesis are the following.

Pairwise Loss (PL) and Sum Loss (SL)

$$PL(\sigma, R) = \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbb{1}(\sigma^{-1}(i) < \sigma^{-1}(j)) \mathbb{1}(R(i) < R(j)) \tag{2.3}$$

$$SL(\sigma, R) = \sum_{i=1}^{m} \sigma^{-1}(i) R(i) \tag{2.4}$$

Ailon (2014) has shown that the regrets under PL and SL are equal. Therefore, we can study SL instead of PL. The minimax regret rate for PL is the same as that for SL.

$$\sum_{t=1}^{T} PL(\sigma_t, R_t) - \sum_{t=1}^{T} PL(\sigma, R_t) = \sum_{t=1}^{T} SL(\sigma_t, R_t) - \sum_{t=1}^{T} SL(\sigma, R_t) \tag{2.5}$$

Although we cannot express PL in the form of $f(\sigma) \cdot R$, we can do that for SL with $f(\sigma) = \sigma^{-1} = [\sigma^{-1}(1), ..., \sigma^{-1}(m)]$.

Discounted Cumulative Gain (DCG)

$$DCG(\sigma, R) = \sum_{i=1}^{m} \frac{R(i)}{\log_2(1 + \sigma^{-1}(i))} \tag{2.6}$$

DCG can be expressed in the form of $f(\sigma) \cdot R$ with

$$f(\sigma) = [\frac{1}{\log_2(1 + \sigma^{-1}(1))}, ..., \frac{1}{\log_2(1 + \sigma^{-1}(m))}]$$

Precision@n Gain (P@n)

$$P@n(\sigma, R) = \sum_{i=1}^{m} \mathbb{1}(\sigma^{-1}(i) \leq n) R(i) \tag{2.7}$$

P@n can also be expressed in the form of $f(\sigma) \cdot R$ with

$$f(\sigma) = [\mathbb{1}(\sigma^{-1}(1) \leq n), ..., \mathbb{1}(\sigma^{-1}(m) \leq n)]$$

Next, we will see that none of AUC (normalized PL), NDCG (normalized DCG), AP (normalized P@n) can be expressed in the form of $f(\sigma) \cdot R$.

Area Under Curve (AUC)

$$AUC(\sigma, R) = \frac{1}{N(R)} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbb{1}(\sigma^{-1}(i) < \sigma^{-1}(j)) \mathbb{1}(R(i) < R(j)) \tag{2.8}$$

where $N(R) = (\sum_{i=1}^{m} \mathbb{1}(R(i) = 1)) \cdot (\sum_{i=1}^{m} \mathbb{1}(R(i) = 0))$.

Normalized Discounted Cumulative Gain (NDCG)

$$NDCG(\sigma, R) = \frac{1}{Z(R)} \sum_{i=1}^{m} \frac{R(i)}{\log_2(1 + \sigma^{-1}(i))} \tag{2.9}$$

where $Z(R) = \max_{\sigma} \sum_{i=1}^{m} \frac{R(i)}{\log_2(1+\sigma^{-1}(i))}$.

Average Precision (AP)

$$AP(\sigma, R) = \frac{1}{\|R\|_1} \sum_{i=1}^{m} \frac{\sum_{j \leq i} \mathbb{1}(R(\sigma(j)) = 1)}{i} \mathbb{1}(R(\sigma(i)) = 1) \tag{2.10}$$

Remark 1. There are reasons why we are interested in this linear (in R) form of ranking measures. The algorithms that establish upper bound for the minimax regret rates require construction of unbiased estimator of the difference vector between two loss vectors that two different actions incur (Bartók et al., 2014; Lattimore and Szepesvari, 2018). The nonlinear (in R) form of ranking measures makes such a construction extremely hard.

2.3 A Quick Review of Partial Monitoring Games

Our results on minimax regret rates are developed based on the theory for general finite partial monitoring games developed in papers by Bartók et al. (2014) and Lattimore and Szepesvari (2018). Before presenting our results, it is necessary to reproduce the relevant definitions and notations as in Bartók et al. (2014), Chaudhuri and Tewari (2017) and Lattimore and Szepesvari (2018). For the sake of easy understanding, we adapt the definitions and notation to our setting.

Recall that in the top k feedback model, there are $m!$ actions and 2^m outcomes (because we only consider binary relevance). Without loss of generality, we fix an ordering $(\sigma_i)_{1\leq i\leq m!}$ of all the actions and an ordering $(R_j)_{1\leq j\leq 2^m}$ of all the outcomes. Note that the subscripts in σ_i and R_j refer to the place in these fixed ordering and do not refer to time points in the game as in σ_t. It will be clear from the context whether we a referring to a place in the ordering or to a time point in the game. A game with ranking measure RL and top k $(1 \leq k \leq m)$ feedback can be defined by a pair of *loss matrix* and *feedback matrix*. The *loss matrix* is denoted by $L \in \mathbb{R}^{m!\times 2^m}$ with rows corresponding to actions and columns corresponding to outcomes. $L_{i,j}$ is the loss the learner suffers when the learner chooses action σ_i and the adversary chooses outcome R_j, i.e., $L_{i,j} = RL(\sigma_i, R_j)$. The *feedback matrix* is denoted by H of size $m! \times 2^m$ with rows corresponding to actions and columns corresponding to outcomes. $H_{i,j}$ is the feedback the learner gets when the learner chooses action σ_i and the adversary chooses outcome R_j, i.e., $H_{i,j} = (R_j(\sigma_i(1)), ..., R_j(\sigma_i(k)))$.

Loss matrix L and feedback matrix H together determine the difficulty of a game. In the following we will introduce some definitions to help understand the underlying structures of L and H.

Let l_i denote the column vector consisting of the i-th row of L. It is also called the loss vector for action i. Let Δ be the probability simplex in $\mathbb{R}^{2^m}$, that is, $\Delta = \{p \in \mathbb{R}^{2^m} : p \geq 0, \mathbf{1}^T p = 1\}$ where the inequality between vector is to be interpreted component-wise. Elements of Δ can be treated as *opponent strategies* as they are distributions over all outcomes. With loss vectors and Δ, we can then define what it means for a learner's action to be optimal.

Definition 1 (Optimal action)**.** Learner's action σ_i is said to be **optimal** under $p \in \Delta$ if $l_i \cdot p \leq l_j \cdot p$ for all $1 \leq j \leq m!$. That is, σ_i has expected loss not greater than that of any other learner's actions under p.

Identifying opponent strategies an action is optimal under gives the *cell decomposition* of Δ.

Definition 2 (Cell decomposition)**.** For learner's action σ_i, $1 \leq i \leq m!$, its **cell** is defined to be $C_i = \{p \in \Delta : l_i \cdot p \leq l_j \cdot p, \forall 1 \leq j \leq m!\}$. Then $\{C_1, ..., C_{m!}\}$ forms the **cell decomposition** of Δ.

It is easy to see that each cell is either empty or is a closed polytope. Based on properties of different cells, we can classify corresponding actions as following.

Definition 3 (Classification of actions)**.** Action σ_i is called **dominated** if $C_i = \emptyset$. Action σ_i is called **nondominated** if $C_i \neq \emptyset$. Action σ_i is called **degenerate** if it is nondominated and there exists action σ_j such that $C_i \subsetneq C_j$. Action σ_i is called **Pareto-optimal** if it is nondominated and not degenerate.

Dominated actions are never optimal. Cells of Pareto-optimal actions have $(2^m - 1)$ dimension, while those of degenerate actions have dimensions strictly less than $(2^m - 1)$.

Sometimes two actions might have the same loss vector, and we will call them duplicate actions. Formally, action σ_i is called **duplicate** if there exists action $\sigma_j \neq \sigma_i$ such that $l_i = l_j$. If actions σ_i and σ_j are duplicate to each other, one might think of removing one of them without loss of generality. Unfortunately, this will not work. Even though σ_i and σ_j have the same loss vector, they might have different feedbacks. Thus removing one of them might lead to a loss of information that the learner can receive.

Next we introduce the concept of *neighbors* defined in terms of Pareto-optimal actions.

Definition 4 (Neighbors)**.** Two Pareto-optimal actions σ_i and σ_j are **neighboring actions** if $C_i \cap C_j$ has dimension $(2^m - 2)$. The **neighborhood action set** of two neighboring actions σ_i and σ_j is defined as $N^+_{i,j} = \{k' : 1 \leq k' \leq m!, C_i \cap C_j \subseteq C_{k'}\}$.

All of the definitions above are with respect to the loss matrix L. The structure of L (i.e., the number of each type of actions) certainly plays an important role in determining the difficulty of a game. (For example, if a game has only one Pareto-optimal action, then simply playing the Pareto-optimal action in each round leads to zero regret.) However, that is only half of the story. In the other half, we will see the feedback matrix H determines how easily we can identify optimal actions.

In the following, we will turn our attention on the feedback matrix H. Recall that $H_{i,j}$ is the feedback the learner gets when the learner plays action σ_i and the adversary plays outcome R_j. Consider the i-th row of H, which is all possible feedbacks the learner could receive when playing action i. We want to infer what outcome the adversary chose from the feedback. Thus, the feedback itself does not matter; what matters is the number of distinct symbols in the i-th row of H. This will determine how easily we can differentiate among outcomes. Therefore, we will use *signal matrices* to standardize the feedback matrix H.

Definition 5 (Signal matrix)**.** Recall that in top k feedback model, the feedback matrix has 2^k distinct symbols $\{0,1\}^k$. Fix an enumeration $s_1, ..., s_{2^k}$ of these symbols. Then the **signal matrix** $S_i \in \{0,1\}^{2^k \times 2^m}$, corresponding to action σ_i, is defined as $(S_i)_{l,l'} = \mathbb{1}(H_{i,l'} = s_l)$.

At this point, one might attempt to construct unbiased estimators for loss vectors for all actions and then apply algorithms like Exp3 (Auer et al., 1998). Unfortunately this approach will not work in this setting. There are easy counterexamples (see Exhibit 1 in Appendix I of Lattimore and Szepesvari (2018)). Another approach is to construct unbiased estimators for differences between loss vectors. The idea is that we do not need to estimate the loss itself; instead it suffices to estimate how an action performs with respect to the optimal action in order to control the regret. It turns out this idea indeed works. The following two definitions capture the difficulty with which we can construct unbiased estimator for loss vector difference.

Definition 6 (Global observability)**.** A pair of actions σ_i and σ_j is called **globally observable** if $l_i - l_j \in \oplus_{1 \leq k' \leq m!} \text{Col}(S_{k'}^T)$, where Col refers to column space. The **global observability** condition holds if every pair of actions is globally observable.

Definition 7 (Local observability)**.** A pair of neighboring actions σ_i and σ_j is called **locally observable** if $l_i - l_j \in \oplus_{k' \in N_{i,j}^+} \text{Col}(S_{k'}^T)$. The **local observability** condition holds if every pair of neighboring actions is locally observable.

Global observability means that the loss vector difference can be estimated using feedbacks from all actions, while local observability means that it can be estimated using just feedbacks from the neighborhood action set. Clearly, local observability is a stronger condition, and local observability implies global observability.

We note that the above two definitions are given in Bartók et al. (2014). Later, when Lattimore and Szepesvari (2018) extended Bartók et al. (2014)'s work, they proposed different (but at least more general) definitions of *global observability* and *local observability*. We reproduce as follows.

Definition 8 (Alternative definitions of global observability and local observability)**.** Let $\Sigma = \{\sigma | \sigma : [m] \to [m] \text{ is bijective}\}$. Let $\mathcal{H}$ denote the set of symbols in H. A pair of actions σ_i and σ_j is called **globally observable** if there exists a function $f : \Sigma \times \mathcal{H} \to \mathbb{R}$

such that

$$\sum_{k'=1}^{m!} f(\sigma_{k'}, H_{k',l'}) = L_{i,l'} - L_{j,l'} \quad \text{for all } 1 \leq l' \leq 2^m$$

They are **locally observable** if in addition they are neighbors and $f(\sigma_{k'}, \cdot) = 0$ when $k' \notin N_{i,j}^{+}$. Again, the **global observability** condition holds if every pair of actions is globally observable, and the **local observability** condition holds if every pair of neighboring actions is locally observable.

Lemma 1. The alternative definitions of global observability and local observability generalize the original definitions.

To see why alternative definitions are more general, it is enough to note that $\sigma_{k'}$ and $H_{k',l'}$ contain the same information as $S_{k'}$ and $e_{l'}$ because observing $H_{k',l'}$ is equivalent to observing $S_{k'}e_{l'}$. The latter can be seen as a one-hot coding vector for the feedback. We will prove this lemma in Chapter 4 formally.

2.4 Classification Theorem for Finite Partial Monitoring

To make this thesis self-contained, in the section, we will state the important result that we use from the theory of finite partial monitoring games. The following theorem provides a full classification of all finite partial monitoring games into four categories.

Theorem 2. [Theorem 2 in Bartók et al. (2014) and Theorem 1 in Lattimore and Szepesvari (2018)] Let partial monitoring game $G = (L, H)$ have K nondominated actions. Then the minimax regret rate of G satisfies

$$R_T^*(G) = \begin{cases} 0, & \text{if } K = 1; \\ \widetilde{\Theta}(\sqrt{T}), & \text{if } K > 1, G \text{ is locally observable;} \\ \Theta(T^{2/3}), & \text{if } G \text{ is globally observable, but not locally observable;} \\ \Theta(T), & G \text{ is not globally observable.} \end{cases}$$

where there is a polylogarithmic factor inside $\widetilde{\Theta}(\cdot)$.

This theorem involves upper and lower bounds for each of the four categories. Several papers (Piccolboni and Schindelhauer, 2001; Antos et al., 2013; Cesa-Bianchi et al., 2006; Bartók et al., 2014; Lattimore and Szepesvari, 2018) contribute to this theorem. In particular, Bartók et al. (2014) summerizes and gives a nearly complete classification theorem. However, they failed to deal with degenerate games (i.e.the game that has degenerate or duplicate actions). Lattimore and Szepesvari (2018) addressed this gap in the literature.

With this classification theorem, it suffices for us to show the local or global observability conditions in order to establish minimax regret rates.

CHAPTER 3

Summary of Results

Under the non-contextual setting of top k feedback model, we have the following results.

3.1 Pairwise Loss (PL) and Sum Loss (SL)

Theorem 3. With respect to loss matrix L and feedback matrix H for SL, the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$.

Theorem 1 in section 2.4 and the discussion in section 2.5 of Chaudhuri and Tewari (2017) have shown that for SL, the global observability holds for all $1 \leq k \leq m$. Combining our Theorem 3 and chaining with Theorem 2, we immediately have the minimax regret for SL:

$$\mathcal{R}_T^* = \begin{cases} \Theta(T^{2/3}), & k = 1, ..., m-2 \\ \widetilde{\Theta}(T^{1/2}), & k = m-1, m \end{cases}$$

By equation 2.5, PL has exactly the same minimax regret rates as SL.

Theorem 3 shows this game is hard for almost all values of k. In particular, since in reality small values of k are more interesting, it rules out the possibility of better regret for practically interesting cases for k. We also note that Chaudhuri and Tewari (2017) showed failure of local observability only for $k = 1$.

As for the time complexity, Chaudhuri and Tewari (2017) provided efficient (polynomial of m time) algorithm for PL and SL for values of k when global observability holds, so we have efficient algorithm for $k = 1, 2, ..., m-2$. For $k = m$, Suehiro et al. (2012) and Ailon (2014) have already shown efficient algorithms. The only case left out is $k = m-1$. Such large value of k is not interesting in practice, so we do not pursue this question.

3.2 Discounted Cumulative Gain (DCG)

Although DCG is a gain function, we can negate it to get a loss function. As in Chaudhuri and Tewari (2017), the results and the proofs for DCG are almost the same as those for SL. We have the following theorem.

Theorem 4. With respect to loss matrix L and feedback matrix H for DCG, the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$.

Corollary 10 of Chaudhuri and Tewari (2017) has shown that for DCG, the minimax regret rate is $O(T^{2/3})$ for $1 \leq k \leq m$. Combining with our Theorem 4 and chaining with Theorem 2, we immediately have the minimax regret for DCG:

$$\mathcal{R}_T^* = \begin{cases} \Theta(T^{2/3}), & k = 1, ..., m-2 \\ \widetilde{\Theta}(T^{1/2}), & k = m-1, m \end{cases}$$

Theorem 4 generalizes the results in Chaudhuri and Tewari (2017) that showed local observability fails only for $k = 1$, and rules out the possibility of better regret for values of k that are practically interesting. Also, there are efficient algorithms for $k = 1, 2, ..., m-2$ (Chaudhuri and Tewari, 2017) and for $k = m$ (Suehiro et al., 2012; Ailon, 2014). Again, we are not interested in designing an efficient algorithm for $k = m-1$.

3.3 Precision@n Gain (P@n)

Theorem 5. For fixed n such that $1 \leq n \leq m$, with respect to loss matrix L and feedback matrix H for P@n, the local observability holds for all $1 \leq k \leq m$.

It is not hard to see this game contains many duplicate actions (but no degenerate actions, as we will show) since P@n only cares about objects ranked in the top n position, irrespective of the order. The minimax regret does not directly follow from Theorem 2 of Bartók et al. (2014). However, a very recent paper Lattimore and Szepesvari (2018) has proved that locally observable games enjoy $\widetilde{\Theta}(T^{1/2})$ minimax regret, regardless of the existence of duplicate actions. This shows the minimax regret for P@n is

$$R_T^* = \widetilde{\Theta}(T^{1/2}) \quad \text{for } 1 \leq k \leq m$$

We note that Chaudhuri and Tewari (2017) only showed $T^{2/3}$ regret rates for P@n, so this result gives improvements over all values of k, including the practically relevant cases when k is small. In the next section, we will also give an efficient algorithm that realizes this regret rate.

3.4 Algorithm for Obtaining the Minimax Regret for P@n with Top-k Feedback

Lattimore and Szepesvari (2018) give an algorithm NEIGHBORHOODWATCH2 that achieves $\widetilde{\Theta}(T^{1/2})$ minimax regret for all finite partial monitoring games with local observability, including games with duplicate or degenerate actions. However, directly applying this algorithm to P@n would be intractable, since the algorithm has to spend $\Omega(\mathrm{poly}(K))$ time per round, where the number of actions K equals $m!$ in our setting with P@n.

We provide a modification before applying the algorithm NEIGHBORHOODWATCH2

so that it spends only $O(\text{poly}(m))$ time per round and obtains a minimax regret rate of $\widetilde{\Theta}(T^{1/2})$. Thus, it is more efficient.

We note that since top-k (for $k > 1$) feedback contains strictly more information than top-1 feedback does, it suffices to give an efficient algorithm for P@n with top-1 feedback which we will show in the following.

We first give a high-level idea why we can significantly reduce the time complexity from exponential in m to polynomial in m. It has to do with the structure of the game for P@n. We have the following observations for P@n.

Lemma 13 in Chapter 4: For P@n, each of learner's actions σ_i is Pareto-optimal.

Lemma 14 in Chapter 4: For action σ_i, define $A_i = \{a : \mathbb{1}(\sigma_i^{-1}(a) \leq n) = 1\}$ and $B_i = \{b : \mathbb{1}(\sigma_i^{-1}(b) \leq n) = 0\}$ be subsets of $\{1, 2, ..., m\}$. A pair of learner's actions $\{\sigma_i, \sigma_j\}$ is a neighboring action pair if there is exactly one pair of objects $\{a, b\}$ such that $a \in A_i$, $a \in B_j$, $b \in B_i$, and $b \in A_j$.

Lemma 15 in Chapter 4: For neighboring action pair $\{\sigma_i, \sigma_j\}$, the neighborhood action set is $N_{i,j}^{+} = \{k : 1 \leq k \leq m!, l_k = l_i \text{ or } l_k = l_j\}$.

Lemma 14 says that P@n only cares about how action σ partitions $[m]$ into sets A and B; the order of objects within A (or B) does not matter. Furthermore, each ordering of objects in A and B corresponds to a unique action. Therefore, based on loss vectors, we can define equivalent classes over $m!$ actions such that all actions within a class share the same loss vector. In other words, each class collects actions duplicate to each other. A simple calculation shows all classes have the same number of actions, $n!(m-n)!$, and there are $\binom{m}{n}$ classes. Note that $\binom{m}{n}$ is $O(m^n)$ for fixed n, a polynomial of m. In real applications, n is usually very small, such as $1, 3, 5$.

In each of the equivalent classes, all the actions have the same partition of $[m]$ into sets A and B, where all objects in A are ranked before objects in B. For top-1 feedback setting, the algorithm only receives the relevance for the object ranked at the top. Therefore, in a class, the algorithm only needs to determine which object from A to be placed at the top

position. Clearly, there are just n choices as there are n objects in A, so we reduce the number of actions to consider in each class from $n!(m-n)!$ to n. Note that this reduction does not incur any loss of information. This is the key idea to simplify the time complexity. We only need to keep a distribution to sample from $n\binom{m}{n}$ (a polynomial of m for fixed n) actions, instead of sampling from $m!$ actions.

To make this section self-contained, we will include the algorithm NEIGHBORHOOD-WATCH2 in Lattimore and Szepesvari (2018) with some changes so that it is consistent with our notations.

Let $\mathcal{C}$ be the set of those $n\binom{m}{n}$ actions defined above. Let $\mathcal{A}$ be an arbitrary largest subset of Pareto-optimal actions from $\mathcal{C}$ such that $\mathcal{A}$ does not contain actions that are duplicates of each other. Note that $|\mathcal{A}| = \binom{m}{n}$ and $\mathcal{A}$ contains an action from each of the equivalent classes. Let $\mathcal{D} = \mathcal{C} \setminus \mathcal{A}$. For action a, let N_a be the set of actions consisting a and a's neighbors. By the alternative definition of local observability, there exists a function $v^{ab} : \Sigma \times \mathcal{H} \to \mathbb{R}$ for each pair of neighboring actions a, b such that the requirement in Definition 8 is satisfied. For notational convenience, let $v^{aa} = 0$ for all action a. Define $V = \max_{a,b} \|v^{ab}\|_\infty$. Since both Σ and $\mathcal{H}$ are finite sets, $\|v^{ab}\|_\infty$ is just $\max_{\sigma\in\Sigma, s\in\mathcal{H}} |v^{ab}(\sigma, s)|$. Lemma 17 shows $\|v^{ab}\|_\infty \leq 4$ for suitable choice of v^{ab}, so V can be upper bounded by 4.

Algorithm 1 NEIGHBORHOODWATCH2

Input L, H, η, γ

for $t = 1, \ldots, T$ **do**

For $a, k \in \mathcal{C}$ let $Q_{tka} = \mathbb{1}_{\mathcal{A}}(k)\dfrac{\mathbb{1}_{N_k \cap \mathcal{A}}(a)\exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{Z}_{ska}\right)}{\sum_{b \in N_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{Z}_{skb}\right)} + \mathbb{1}_{\mathcal{D}}(k)\dfrac{\mathbb{1}_{\mathcal{A}}(a)}{|\mathcal{A}|}$

Find distribution $\widetilde{P}_t$ such that $\widetilde{P}_t^\top = \widetilde{P}_t^\top Q_t$

Compute $P_t = (1-\gamma)\text{REDISTRIBUTE}(\widetilde{P}_t) + \frac{\gamma}{|\mathcal{C}|}\mathbf{1}$

Sample $A_t \sim P_t$ and receive feedback Φ_t

Compute loss-difference estimators for each $k \in \mathcal{A}$ and $a \in N_k \cap \mathcal{A}$:

$$\widehat{Z}_{tka} = \frac{\widetilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}} \quad \text{and} \quad \beta_{tka} = \eta V^2 \sum_{b \in N_{ak}^+} \frac{\widetilde{P}_{tk}^2}{P_{tb}} \quad \text{and} \quad \widetilde{Z}_{tka} = \widehat{Z}_{tka} - \beta_{tka}$$

end for

Note that REDISTRIBUTE function has no effect since there are no degenerate actions, so $P_t = (1-\gamma)\widetilde{P}_t + \frac{\gamma}{|\mathcal{C}|}\mathbf{1}$ and we omit the definition of REDISTRIBUTE function here.

Theorem 6 (Derived from Theorem 2 in Lattimore and Szepesvari (2018))**.** Let $K = |\mathcal{C}| = n\binom{m}{n}$. For top-1 feedback model with P@n, suppose the algorithm above is run on $G = (L, H)$ with $\eta = \frac{1}{V}\sqrt{\log(K)/T}$ and $\gamma = \eta K V$. Then

$$\mathcal{R}_T^* \leq O\Big(\frac{KV}{\epsilon_G}\sqrt{T\log(K)}\Big)$$

where ϵ_G is a constant specific to the game G, not depending on T (Lemma 16 shows $\frac{1}{\epsilon_G}$ can be defined as $4m$ in this setting.). Moreover, the time complexity in each round is $O(poly(m))$, and $\frac{V}{\epsilon_G} \leq 16m$ which is $O(poly(m))$.

CHAPTER 4

Proofs

4.1 Proofs for Section 2

Proof for Lemma 1.

Proof. We prove the alternative definition of global observability (Definition 8) generalizes the original definition (Definition 6). It will follow that the alternative definition of local observability (Definition 8) generalizes the original definition (Definition 7).

Consider top k feedback model, so each signal matrix $S_{k'}$ is 2^k by 2^m. Definition 6 says a pair of actions σ_i and σ_j is globally observable if $l_i - l_j \in \oplus_{1 \le k' \le m!} \text{Col}(S_{k'}^T)$. So we can write $l_i - l_j$ as

$$l_i - l_j = \sum_{k'=1}^{m!} \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^T)_{l'}$$

where $c_{k',l'} \in \mathbb{R}$ is constant, and $(S_{k'}^T)_{l'}$ is the l'-th column of $S_{k'}^T$. Define

$$f(\sigma_{k'}, H_{k',k''}) = \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^T)_{k'',l'}, \quad \text{for } 1 \le k'' \le 2^m \tag{4.1}$$

where $(S_{k'}^T)_{k'',l'}$ is the element in row k'' and column l' of $S_{k'}^T$. Then $l_i - l_j$ can also be written as

$$l_i - l_j = \sum_{k'=1}^{m!} \begin{bmatrix} f(\sigma_{k'}, H_{k',1}) \\ \dots \\ f(\sigma_{k'}, H_{k',2^m}) \end{bmatrix}$$

satisfying Definition 8. □

4.2 Proofs for Theorem 3 in Section 3.1

In this section, the loss function is SL unless otherwise stated.

Lemma 7 (Chaudhuri and Tewari (2017), Lemma 2)**.** For SL, each of the leaner's actions σ_i is Pareto-optimal.

Proof. We reproduce their proof as concepts in the proof are useful for proving our results.

For $p \in \Delta$, $l_i \cdot p = \sum_{j=1}^{2^m} p_j(\sigma_i^{-1} \cdot R_j) = \sigma_i^{-1} \cdot \sum_{j=1}^{2^m} p_j R_j = \sigma_i^{-1} \cdot \mathbb{E}[R]$, where the expectation is taken with respect to p. Then $l_i \cdot p$ is minimized when $\mathbb{E}[R(\sigma_i(1))] \geq \mathbb{E}[R(\sigma_i(2))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))]$. Therefore, the cell of σ_i is $C_i = \{p \in \Delta : \mathbf{1}^T p = 1, \mathbb{E}[R(\sigma_i(1))] \geq \mathbb{E}[R(\sigma_i(2))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))]\}$. C_i has only one equality constraint, and hence has dimension $(2^m - 1)$. This shows σ_i is Pareto-optimal. □

Determining whether the game is locally observable requires knowing all neighboring action pairs. We now characterize neighboring actions pairs for SL.

Lemma 8. A pair of actions σ_i and σ_j is a neighboring action pair if and only if there is exactly one pair of objects $\{a, b\}$, whose positions differ in σ_i and σ_j, such that a is placed just before b in σ_i, and b is placed just before a in σ_j.

Our Lemma 8 strengthens Lemma 3 of Chaudhuri and Tewari (2017) by showing the condition (*there is exactly* ...) is also necessary.

Proof. The "if" part is Lemma 3 of Chaudhuri and Tewari (2017). We only need to prove the "only if" part.

Assume the condition (*there is exactly* ...) fails to hold. Note that σ_i and σ_j cannot be the same, so they differ in at least two positions. Then there are two cases. 1. σ_i and σ_j differ in exactly two positions, but the two positions are not consecutive. 2. σ_i and σ_j differ

in more than two positions.

Consider case 1. Without loss of generality, assume σ_i and σ_j differ in positions k and k' where $k' - k > 1$. Then there is a unique pair of objects $\{a, b\}$ such that $\sigma_i(k) = a$, $\sigma_i(k') = b$, $\sigma_j(k) = b$ and $\sigma_j(k') = a$, and $\sigma_i(l) = \sigma_j(l)$ for $l \neq k, k'$. From Lemma 7, σ_i has cell

$$C_i = \{p \in \Delta : \mathbb{E}[R(\sigma_i(1))] \geq ... \geq \mathbb{E}[R(\sigma_i(k))] \geq ... \geq \mathbb{E}[R(\sigma_i(k'))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))]\}$$

and σ_j has cell

$$C_j = \{p \in \Delta : \mathbb{E}[R(\sigma_j(1))] \geq ... \geq \mathbb{E}[R(\sigma_j(k))] \geq ... \geq \mathbb{E}[R(\sigma_j(k'))] \geq ... \geq \mathbb{E}[R(\sigma_j(m))]\}$$

Then

$$\begin{aligned} C_i \cap C_j = \{p \in \Delta :& \mathbb{E}[R(\sigma_i(1))] \geq ... \geq \mathbb{E}[R(\sigma_i(k))] \geq ... \\ & \geq \mathbb{E}[R(\sigma_i(k'))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))], \\ & \mathbb{E}[R(\sigma_j(1))] \geq ... \geq \mathbb{E}[R(\sigma_j(k))] \geq ... \\ & \geq \mathbb{E}[R(\sigma_j(k'))] \geq ... \geq \mathbb{E}[R(\sigma_j(m))]\} \end{aligned}$$

Since

$$\mathbb{E}[R(\sigma_i(k))] = \mathbb{E}[R(\sigma_j(k'))] = \mathbb{E}[R(a)]$$

and

$$\mathbb{E}[R(\sigma_i(k'))] = \mathbb{E}[R(\sigma_j(k))] = \mathbb{E}[R(b)]$$

It follows that $C_i \cap C_j$ has a constraint

$$\mathbb{E}[R(\sigma_i(k))] = ... = \mathbb{E}[R(\sigma_i(k'))]$$

Since $k' - k > 1$, $\mathbb{E}[R(\sigma_i(k))] = ... = \mathbb{E}[R(\sigma_i(k'))]$ has at least two equalities. Then $C_i \cap C_j$ has at least three equality constraints (including $\mathbf{1}^T p = 1$), which shows $C_i \cap C_j$ has dimension less than $(2^m - 2)$. Therefore, $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair.

Now consider case 2. If σ_i and σ_j differ in more than two positions, then there are at least two pairs of objects such that for each pair, the relative order of the two objects in σ_i is different from that in σ_j. Applying argument for case 1 to case 2 shows $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair. □

Remark 2. From Lemma 8, neighboring action pair $\{\sigma_i, \sigma_j\}$ has the form: $\sigma_i(k) = a, \sigma_i(k+1) = b, \sigma_j(k) = b, \sigma_j(k+1) = a$ and $\sigma_i(l) = \sigma_j(l), \forall l \neq k, k+1$, for objects a and b. Using the definition of SL, we can see that $l_i - l_j$ contains 2^{m-1} nonzero entries, of which 2^{m-2} entries are 1 and 2^{m-2} entries are -1. Moreover, if $R_s(a) = 1$ and $R_s(b) = 0$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is -1. If $R_s(a) = 0$ and $R_s(b) = 1$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is 1. If $R_s(a) = R_s(b)$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is 0.

Once we know what is a neighboring action pair, we need to characterize the corresponding neighborhood action set.

Lemma 9. For neighboring action pair $\{\sigma_i, \sigma_j\}$, the neighborhood action set is $N^+_{i,j} = \{i, j\}$, so $\oplus_{k \in N^+_{i,j}} \text{Col}(S^T_k) = \text{Col}(S^T_i) \oplus \text{Col}(S^T_j)$.

Proof. By definition of neighborhood action set, $N^+_{i,j} = \{k : 1 \leq k \leq m!, C_i \cap C_j \subseteq C_k\}$. Bartók et al. (2014) mentions that if $N^+_{i,j}$ contains some other action σ_k, then either $C_k = C_i$, $C_k = C_j$, or $C_k = C_i \cap C_j$. From Lemma 7, for SL each of learner's actions is

Pareto-optimal, so $\dim(C_k) = 2^m - 1$. This shows $C_k \neq C_i \cap C_j$. To see $C_k \neq C_i$, assume for contraction that $C_k = C_i$. Then this means that both actions σ_i and σ_k are optimal under p, $\forall p \in C_k$, which implies $0 = l_i \cdot p - l_k \cdot p = p \cdot (l_i - l_k)$ for all $p \in C_k$. For SL, $l_i \neq l_k$ for $i \neq k$. Then $p \cdot (l_i - l_k) = 0$ for all $p \in C_k$ would impose another equality constraint on C_k, so $\dim(C_k) \leq 2^m - 2$. We know $\dim(C_k) = 2^m - 1$, a contraction. This shows $C_k \neq C_i$. Similarly, we have $C_k \neq C_j$. Therefore, $N^+_{i,j} = \{i, j\}$ and $\oplus_{k \in N^+_{i,j}} \text{Col}(S^T_k) = \text{Col}(S^T_i) \oplus \text{Col}(S^T_j)$. □

We are ready to prove Theorem 3. For readers' convenience, let us recall that Theorem 3 says for SL the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$. The proof will contain two parts corresponding to the case when local observability fails and the case when local observability holds.

Proof. **Part 1:** We first prove the local observability fails for $k = 1, ..., m-2$. It suffices to show the local observability fails for $k = m-2$ because top k feedback has strictly more information than top k' feedback does for $k' < k$. (In other words, if top k' feedback is locally observable, then top k feedback must be, as one can always throw away the extra information.)

Note that for the signal matrix when $k = m-2$, each row has exactly 4 ones and each column has exactly 1 one.

Consider two actions $\sigma_1 = 1, 2, 3, ..., m-2, m-1, m$ and $\sigma_2 = 1, 2, 3, ..., m-2, m, m-1$. That is, σ_1 gives object i rank i for $1 \leq i \leq m$. σ_2 gives object i rank i for $1 \leq i \leq m-2$, object m rank $m-1$ and object $m-1$ rank m. By Lemma 8, σ_1 and σ_2 are neighboring actions.

Inspired by observations from Remark 2, we form 2^{m-2} groups of 4 relevance vectors such that within each group, the relevance vectors only differ at object $m-1$ and m. Correspondingly, we divide the vector $l_1 - l_2$ into 2^{m-2} groups. Then each group is $[0\ 1\ -1\ 0]$. For signal matrices S_1 and S_2, we can also form 2^{m-2} groups of 4 columns accordingly. For $k = m-2$, the signal matrix is of size $2^{m-2} \times 2^m$, and in this case, σ_1 and σ_2 have

the same signal matrix $S = S_1 = S_2$ because σ_1 and σ_2 have exactly the same feedback no matter what the relevance vector is. Now in each group, there are only two types of rows of S, namely $[0\ 0\ 0\ 0]$ and $[1\ 1\ 1\ 1]$. Table 4.1 shows $l_1 - l_2$ and two types of rows of S for each group. It is clear that $l_1 - l_2 \notin \text{Col}(S^T)$. This shows the local observability fails for $k = m - 2$.

	$R(m-1)=0$ $R(m)=0$	$R(m-1)=0$ $R(m)=1$	$R(m-1)=1$ $R(m)=0$	$R(m-1)=1$ $R(m)=1$
$l_1 - l_2$	0	1	-1	0
rows of S	1 0	1 0	1 0	1 0

Table 4.1: SL, Part 1, within the group, $R(c)$ is the same for all $c \neq m-1, m$.

Part 2: We then prove the local observability holds for $k = m-1, m$. Again, it suffices to show the local observability holds for $k = m - 1$. (Note that for $k = m$, the game has bandit feedback, and thus is locally observable as in Section 2.1 of Bartók et al. (2014).)

Note that for the signal matrix when $k = m - 1$, each row has exactly 2 ones and each column has exactly 1 one.

Consider neighboring action pair $\{\sigma_i, \sigma_j\}$. Let $\{a, b\}$ be a pair of objects as in Remark 2. We proceed similarly as in Part 1. We form 2^{m-2} groups of 4 relevance vectors such that within each group, the relevance vectors only differ at object a and b. Correspondingly, we divide the vector $l_a - l_b$ into 2^{m-2} groups. Then each group is $[0\ 1\ -1\ 0]$. For signal matrices S_i and S_j, we can also form 2^{m-2} groups of 4 columns accordingly. Then there are two cases:

(1) Neither a nor b is ranked last by σ_i or σ_j, so the relevance for a and the relevance for b are both revealed through feedback. Concatenate S_i and S_j by row and denote the resultant matrix by S. S is of size $2^m \times 2^m$. Now in each group, there are only five types of rows of S, as shown in Table 4.2. It is clear that the piece $[0\ 1\ -1\ 0]$ is in the row space of S. In this case, $l_i - l_j \in \text{Col}(S_i^T) \oplus \text{Col}(S_j^T)$.

(2) Either a or b is ranked last by σ_i or σ_j, so only one of the relevance for a and the

relevance for b is revealed through feedback. Concatenate S_i and S_j by row and denote the resultant matrix by S. S is of size $2^m \times 2^m$. Now in each group, there are only five types of rows of S, as shown in Table 4.3. The piece $[0\ 1\ -1\ 0]$ is in the row space of S because $[0\ 1\ -1\ 0] = 2[1\ 1\ 0\ 0] + [0\ 0\ 1\ 1] - 2[1\ 0\ 1\ 0] - [0\ 1\ 0\ 1]$. In this case, $l_i - l_j \in \text{Col}(S_i^T) \oplus \text{Col}(S_j^T)$.

In either case, we have $l_i - l_j \in \text{Col}(S_i^T) \oplus \text{Col}(S_j^T)$, so $\{\sigma_i, \sigma_j\}$ is locally observable.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	1	-1	0
rows of S	1	0	0	0
	0	1	0	0
	0	0	1	0
	0	0	0	1
	0	0	0	0

Table 4.2: SL, Part 2 (1), within the group, $R(c)$ is the same for all $c \neq a, b$.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	1	-1	0
rows of S	1	1	0	0
	0	0	1	1
	1	0	1	0
	0	1	0	1
	0	0	0	0

Table 4.3: SL, Part 2 (2), within the group, $R(c)$ is the same for all $c \neq a, b$.

Hence the local observability holds. □

4.3 Proofs for Theorem 4 in Section 3.2

In this section, the loss function is negated DCG unless otherwise stated.

Lemma 10. For negated DCG, each of the leaner's actions σ_i is Pareto-optimal.

Proof. Let $f(\sigma) = [\frac{1}{\log_2(1+\sigma^{-1}(1))}, ..., \frac{1}{\log_2(1+\sigma^{-1}(m))}]$. Note that negated DCG can be written as $-DCG(\sigma, R) = -f(\sigma) \cdot R$. For $p \in \Delta$, $l_i \cdot p = -\sum_{j=1}^{2^m} p_j (f(\sigma_i) \cdot R_j) = -f(\sigma_i) \cdot \sum_{j=1}^{2^m} p_j R_j = -f(\sigma_i) \cdot \mathbb{E}[R]$, where the expectation is taken with respect to p. Note that $-f(\sigma_i)$ is an element-wise strictly increasing transformation of σ_i^{-1}. That is, $\sigma_i^{-1}(k') > \sigma_i^{-1}(l')$ implies $-\frac{1}{\log_2(1+\sigma_i^{-1}(k'))} > -\frac{1}{\log_2(1+\sigma_i^{-1}(l'))}$. Then similarly as in SL, $l_i \cdot p$ is minimized when $\mathbb{E}[R(\sigma_i(1))] \geq \mathbb{E}[R(\sigma_i(2))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))]$. Therefore, the cell of σ_i is $C_i = \{p \in \Delta : \mathbf{1}^T p = 1, \mathbb{E}[R(\sigma_i(1))] \geq \mathbb{E}[R(\sigma_i(2))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))]\}$. C_i has only one equality constraint, and hence has dimension $(2^m - 1)$. This shows σ_i is Pareto-optimal. □

Lemma 11. A pair of actions σ_i and σ_j is a neighboring action pair if and only if there is exactly one pair of objects $\{a, b\}$, whose positions differ in σ_i and σ_j, such that a is placed just before b in σ_i, and b is placed just before a in σ_j.

Proof. This follows from the proof for Lemma 8. □

Remark 3. From Lemma 11, neighboring action pair $\{\sigma_i, \sigma_j\}$ has the form: $\sigma_i(k') = a, \sigma_i(k'+1) = b, \sigma_j(k') = b, \sigma_j(k'+1) = a$ for some k', and $\sigma_i(l) = \sigma_j(l), \forall l \neq k', k'+1$, for objects a and b. Using the definition of negated DCG, we can see that $l_i - l_j$ contains 2^{m-1} nonzero entries, of which 2^{m-2} entries are $-\frac{1}{\log_2(1+k'+1)} + \frac{1}{\log_2(1+k')}$ and 2^{m-2} entries are $-\frac{1}{\log_2(1+k')} + \frac{1}{\log_2(1+k'+1)}$. Moreover, if $R_s(a) = 1$ and $R_s(b) = 0$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is $-\frac{1}{\log_2(1+k')} + \frac{1}{\log_2(1+k'+1)}$. If $R_s(a) = 0$ and $R_s(b) = 1$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is $-\frac{1}{\log_2(1+k'+1)} + \frac{1}{\log_2(1+k')}$. If $R_s(a) = R_s(b)$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is 0.

Note that Remark 3 is different from Remark 2. For negated DCG, the loss vector difference between a neighboring action pair is specific to that pair (the dependence on k' as in Remark 3). For SL, the loss vector difference between a neighboring action pair is

almost the same (up to sign) for all neighboring action pairs. We will see this difference, however, does not affect our analysis.

Lemma 12. For neighboring action pair $\{\sigma_i, \sigma_j\}$, the neighborhood action set is $N^+_{i,j} = \{i, j\}$, so $\oplus_{k \in N^+_{i,j}} \text{Col}(S^T_k) = \text{Col}(S^T_i) \oplus \text{Col}(S^T_j)$.

Proof. This follows from the proof for Lemma 9. □

We now prove Theorem 4. For readers' convenience, let us recall that Theorem 4 says for DCG the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$. The proof will contain two parts corresponding to the case when local observability fails and the case when local observability holds.

Proof. **Part 1:** We first prove the local observability fails for $k = 1, ..., m-2$. It suffices to show the local observability fails for $k = m-2$ because top k feedback has strictly more information than top k' feedback does for $k' < k$.

Note that for the signal matrix when $k = m-2$, each row has exactly 4 ones and each column has exactly 1 one.

Consider two actions $\sigma_1 = 1, 2, 3, ..., m-2, m-1, m$ and $\sigma_2 = 1, 2, 3, ..., m-2, m, m-1$. That is, σ_1 gives object i rank i for $1 \leq i \leq m$. σ_2 gives object i rank i for $1 \leq i \leq m-2$, object m rank $m-1$ and object $m-1$ rank m. By Lemma 11, σ_1 and σ_2 are neighboring actions.

Inspired by observations from Remark 3, we form 2^{m-2} groups of 4 relevance vectors such that within each group, the relevance vectors only differ at object $m-1$ and m. Correspondingly, we divide the vector $l_1 - l_2$ into 2^{m-2} groups. Then each group is $[0,\ -\frac{1}{\log_2(m+1)} + \frac{1}{\log_2(m)},\ -\frac{1}{\log_2(m)} + \frac{1}{\log_2(m+1)},\ 0]$ (see Remark 3). For signal matrices S_1 and S_2, we can also form 2^{m-2} groups of 4 columns accordingly. For $k = m-2$, the signal matrix is of size $2^{m-2} \times 2^m$, and in this case, σ_1 and σ_2 have the same signal matrix $S = S_1 = S_2$ because σ_1 and σ_2 have exactly the same feedback no matter what the relevance vector is. Now in each group, there are only two types of rows of S, namely $[0\ 0\ 0\ 0]$

and $[1\ 1\ 1\ 1]$. Table 4.4 shows $l_1 - l_2$ and two types of rows of S for each group. It is clear that $l_1 - l_2 \notin \text{Col}(S^T)$. This shows the local observability fails for $k = m - 2$.

	$R(m-1)=0$ $R(m)=0$	$R(m-1)=0$ $R(m)=1$	$R(m-1)=1$ $R(m)=0$	$R(m-1)=1$ $R(m)=1$
$l_1 - l_2$	0	$-\frac{1}{\log_2(m+1)} + \frac{1}{\log_2(m)}$	$-\frac{1}{\log_2(m)} + \frac{1}{\log_2(m+1)}$	0
rows of S	1 0	1 0	1 0	1 0

Table 4.4: DCG, Part 1, within the group, $R(c)$ is the same for all $c \neq m-1, m$.

Part 2: We then prove the local observability holds for $k = m-1, m$. Again, it suffices to show the local observability holds for $k = m - 1$. (Note that for $k = m$, the game has bandit feedback, and thus is locally observable as in Section 2.1 of Bartók et al. (2014).)

Note that for the signal matrix when $k = m - 1$, each row has exactly 2 ones and each column has exactly 1 one.

Consider neighboring action pair $\{\sigma_i, \sigma_j\}$. Let $\{a, b\}$ be a pair of objects as in Remark 3. We proceed similarly as in Part 1. We form 2^{m-2} groups of 4 relevance vectors such that within each group, the relevance vectors only differ at object a and b. Correspondingly, we divide the vector $l_a - l_b$ into 2^{m-2} groups. Then each group is $[0,\ -\frac{1}{\log_2(m+1)} + \frac{1}{\log_2(m)},\ -\frac{1}{\log_2(m)} + \frac{1}{\log_2(m+1)},\ 0]$. For signal matrices S_i and S_j, we can also form 2^{m-2} groups of 4 columns accordingly. Then there are two cases:

(1) Neither a nor b is ranked last by σ_i or σ_j, so the relevance for a and the relevance for b are both revealed through feedback. Concatenate S_i and S_j by row and denote the resultant matrix by S. S is of size $2^m \times 2^m$. Now in each group, there are only five types of rows of S, as shown in Table 4.5. It is clear that the piece $[0,\ -\frac{1}{\log_2(m+1)} + \frac{1}{\log_2(m)},\ -\frac{1}{\log_2(m)} + \frac{1}{\log_2(m+1)},\ 0]$ is in the row space of S. In this case, $l_i - l_j \in \text{Col}(S_i^T) \oplus \text{Col}(S_j^T)$.

(2) Either a or b is ranked last by σ_i or σ_j, so only one of the relevance for a and the relevance for b is revealed through feedback. Concatenate S_i and S_j by row and denote the resultant matrix by S. S is of size $2^m \times 2^m$. Now in each group, there are only five types of rows of S, as shown in Table 4.6. The piece $[0,\ -\frac{1}{\log_2(m+1)} + \frac{1}{\log_2(m)},\ -\frac{1}{\log_2(m)} + \frac{1}{\log_2(m+1)},\ 0]$

is in the row space of S because $[0,\ -\frac{1}{\log_2(m+1)}+\frac{1}{\log_2(m)},\ -\frac{1}{\log_2(m)}+\frac{1}{\log_2(m+1)},\ 0] = (-\frac{1}{\log_2(m+1)}+\frac{1}{\log_2(m)})[1\ 1\ 0\ 0] - (-\frac{1}{\log_2(m+1)}+\frac{1}{\log_2(m)})[1\ 0\ 1\ 0]$. In this case, $l_i - l_j \in \text{Col}(S_i^T) \oplus \text{Col}(S_j^T)$.

In either case, we have $l_i - l_j \in \text{Col}(S_i^T) \oplus \text{Col}(S_j^T)$, so $\{\sigma_i, \sigma_j\}$ is locally observable.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	$-\frac{1}{\log_2(m+1)}+\frac{1}{\log_2(m)}$	$-\frac{1}{\log_2(m)}+\frac{1}{\log_2(m+1)}$	0
rows of S	1	0	0	0
	0	1	0	0
	0	0	1	0
	0	0	0	1
	0	0	0	0

Table 4.5: DCG, Part 2 (1), within the group, $R(c)$ is the same for all $c \neq a, b$.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	$-\frac{1}{\log_2(m+1)}+\frac{1}{\log_2(m)}$	$-\frac{1}{\log_2(m)}+\frac{1}{\log_2(m+1)}$	0
rows of S	1	1	0	0
	0	0	1	1
	1	0	1	0
	0	1	0	1
	0	0	0	0

Table 4.6: DCG, Part 2 (2), within the group, $R(c)$ is the same for all $c \neq a, b$.

Hence the local observability holds. □

4.4 Proofs for Theorem 5 in Section 3.3

In this section, the loss function is negated P@n unless otherwise stated.

Lemma 13. For negated P@n, each of learner's actions σ_i is Pareto-optimal.

Proof. The negated P@n is defined as $-P@n(\sigma, R) = -f(\sigma)\cdot R$ where $f(\sigma) = [\mathbb{1}(\sigma^{-1}(1) \leq n), ..., \mathbb{1}(\sigma^{-1}(m) \leq n)]$. For any $p \in \Delta$, we have $l_i \cdot p = -\sum_{j=1}^{2^m} p_j(f(\sigma_i)\cdot R_j) = -f(\sigma_i)\cdot(\sum_{j=1}^{2^m} p_j R_j) = -f(\sigma_i)\cdot\mathbb{E}[R]$, where the expectation is taken with respect to

p. Let $A_i = \{a : \mathbb{1}(\sigma_i^{-1}(a) \leq n) = 1\}$ and $B_i = \{b : \mathbb{1}(\sigma_i^{-1}(b) \leq n) = 0\}$ be subsets of $\{1, 2, ..., m\}$. A_i is the set of objects contributing to the loss while B_i is the set of objects not contributing to the loss. Then $l_i \cdot p$ is minimized when the expected relevances of objects are such that $\mathbb{E}[R(a)] \geq \mathbb{E}[R(b)]$ for all $a \in A_i, b \in B_i$. Therefore, $C_i = \{p \in \Delta : \mathbf{1}^T p = 1, \mathbb{E}[R(a)] \geq \mathbb{E}[R(b)], \forall a \in A_i, \forall b \in B_i\}$. C_i has only one equality constraint and hence has dimension $(2^m - 1)$. This shows action σ_i is Pareto-optimal. □

Next, we characterize neighboring action pairs for negated P@n.

Lemma 14. For negated P@n, a pair of learner's actions $\{\sigma_i, \sigma_j\}$ is a neighboring action pair if there is exactly one pair of objects $\{a, b\}$ such that $a \in A_i$, $a \in B_j$, $b \in B_i$, and $b \in A_j$, where A_i, A_j, B_i, B_j are defined as in Lemma 13.

Proof. For the "if" part, assume the condition (*there is exactly ...*) holds. From Lemma 13, action σ_i is Pareto-optimal and its cell is $C_i = \{p \in \Delta : \mathbb{E}[R(x)] \geq \mathbb{E}[R(y)], \forall x \in A_i, y \in B_i\}$. Action σ_j is also Pareto-optimal and its cell is $C_j = \{p \in \Delta : \mathbb{E}[R(x)] \geq \mathbb{E}[R(y)], \forall x \in A_j, y \in B_j\}$. Then $C_i \cap C_j = \{p \in \Delta : \mathbb{E}[R(a)] = \mathbb{E}[R(b)]$ and $\mathbb{E}[R(x)] \geq \mathbb{E}[R(y)], \forall x \in A_i, y \in B_i$ and $\mathbb{E}[R(z)] \geq \mathbb{E}[R(w)], \forall z \in A_j, w \in B_j\}$. $C_i \cap C_j$ has only two equality constraints (counting $\mathbf{1}^T p = 1$), and hence it has dimension $(2^m - 2)$. Therefore, $\{\sigma_i, \sigma_j\}$ is a neighboring action pair.

For the "only if" part, assume the condition (*there is exactly ...*) does not hold. Note that for negated P@n, $|A_i| = n$ and $|B_i| = m - n$ for all action σ_i. There are two cases. 1. $|A_i \setminus A_j| = 0$. 2. $|A_i \setminus A_j| > 1$.

For the first case, if $|A_i \setminus A_j| = 0$, then $A_i = A_j$ and $B_i = B_j$. Then $C_i \cap C_j = C_i$ has dimension $(2^m - 1)$ because σ_i is Pareto-optimal by Lemma 13. Thus, in this case, $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair.

For the second case, if $|A_i \setminus A_j| > 1$, then there are at least two pair of objects $\{a, b\}$ and $\{a', b'\}$ such that $a, a' \in A_i$, $a, a' \in B_j$, $b, b' \in B_i$, and $b, b' \in A_j$. Following the arguments in the "if" part, it is easy to show that $C_i \cap C_j$ has at least three equality constraints (counting

$\mathbf{1}^T p = 1$), and hence it has dimension less than $(2^m - 2)$. Thus, in this case, $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair. □

Remark 4. From Lemma 14, for neighboring action pair $\{\sigma_i, \sigma_j\}$, we know there is exactly one pair of objects $\{a, b\}$ such that $a \in A_i$, $a \in B_j$, $b \in B_i$, and $b \in A_j$, where A_i, A_j, B_i, B_j are defined as in Lemma 13. Using definition of negated P@n, we can see $l_i - l_j$ contains 2^{m-1} nonzero entries, of which 2^{m-2} entries are 1 and 2^{m-2} entries are -1. Moreover, if $R_s(a) = 1$ and $R_s(b) = 0$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is -1. If $R_s(a) = 0$ and $R_s(b) = 1$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is 1. If $R_s(a) = R_s(b)$ for the s-th ($1 \leq s \leq 2^m$) relevance, then the s-th entry of $l_i - l_j$ is 0.

Then we characterize neighborhood action set for a neighboring action pair.

Lemma 15. For neighboring action pair $\{\sigma_i, \sigma_j\}$, the neighborhood action set is $N^+_{i,j} = \{k : 1 \leq k \leq m!, l_k = l_i \text{ or } l_k = l_j\}$.

Proof. By definition of neighborhood action set, $N^+_{i,j} = \{k : 1 \leq k \leq m!, C_i \cap C_j \subseteq C_k\}$. Bartók et al. (2014) mentions that if $N^+_{i,j}$ contains some other action σ_k, then either $C_k = C_i$, $C_k = C_j$, or $C_k = C_i \cap C_j$. From Lemma 13, every action is Pareto-optimal for negated P@n, so $\dim(C_k) = 2^m - 1$. Hence $C_k \neq C_i \cap C_j$. If $C_k = C_i$, then both actions σ_i and σ_k are optimal under p, $\forall p \in C_k$, which implies $0 = l_i \cdot p - l_k \cdot p = p \cdot (l_i - l_k)$ for all $p \in C_k$. Since C_k has dimension $(2^m - 1)$, $p \cdot (l_i - l_k) = 0$ cannot impose an equality constraint on C_k. Therefore, $l_i = l_k$. Similarly, if $C_k = C_j$, then $l_j = l_k$. This shows $N^+_{i,j} = \{k : 1 \leq k \leq 2^m, l_k = l_i \text{ or } l_k = l_j\}$. □

Remark 5. Negated P@n says that it only matters the way of partitioning m objects into 2 sets A and B as in Lemma 13. For a fixed partition A and B, we can permute objects within A and within B, and all such permutations give the same loss vector and the same cell. Thus, there are duplicate actions in P@n, but no degenerate actions.

We now prove Theorem 5. Let us recall that Theorem 5 says the local observability holds for all $1 \leq k \leq m$.

Proof. It suffices to show the local observability holds for $k = 1$ because there is strictly more information for game with $k > 1$ than that with $k = 1$.

Note that for the signal matrix when $k = 1$, each row has exactly 2^{m-1} ones and each column has exactly 1 one.

Consider neighboring action pair $\{\sigma_i, \sigma_j\}$. Let $\{a, b\}$ be a pair of objects as in Remark 4. We form 2^{m-2} groups of 4 relevance vectors such that within each group, the relevance vectors only differ at object a and b. Correspondingly, we divide the vector $l_a - l_b$ into 2^{m-2} groups. Then each group is $[0 \ -1 \ 1 \ 0]$. For signal matrices S_l where $l \in N^+_{i,j}$, we can also form 2^{m-2} groups of 4 columns accordingly. Then concatenate all signal matrices S_l where $l \in N^+_{i,j}$ by row and denote the resultant matrix by S. S is of size $2^{4n!(m-n)!} \times 2^m$. Now in each group, there are only five types of rows of S, as shown in Table 4.7. $[1\ 1\ 0\ 0]$ and $[0\ 0\ 1\ 1]$ correspond to the action σ_l with $l \in N^+_{i,j}$ that puts object a rank 1. $[1\ 0\ 1\ 0]$ and $[0\ 1\ 0\ 1]$ correspond to the action $\sigma_{l'}$ with $l' \in N^+_{i,j}$ that puts object b rank 1. The piece $[0 \ -1 \ 1 \ 0]$ is in the row space of S because $[0 \ -1 \ 1 \ 0] = -2[1\ 1\ 0\ 0] - [0\ 0\ 1\ 1] + 2[1\ 0\ 1\ 0] + [0\ 1\ 0\ 1]$. Therefore, $l_i - l_j \in \oplus_{l \in N^+_{i,j}} \text{Col}(S_l^T)$, so $\{\sigma_i, \sigma_j\}$ is locally observable and the local observability holds for P@n. □

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	-1	1	0
rows of S	1	1	0	0
	0	0	1	1
	1	0	1	0
	0	1	0	1
	0	0	0	0

Table 4.7: P@n, within the group, $R(c)$ is the same for all $c \neq a, b$.

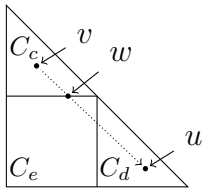

Figure 4.1: Illustrating proof for Lemma 16
, adopted from Lattimore and Szepesvari (2018).

4.5 Proofs for Theorem 6 in Section 3.4

In this section, the loss function is negated P@n unless otherwise stated. We consider top-1 feedback model as described in Section 3.4.

Lemma 16 (Lemma 2 in Lattimore and Szepesvari (2018), Lemma 6 in Bartók et al. (2014))**.** There exists a constant $\epsilon_G > 0$, depending only on G, such that for all $c, d \in \mathcal{A}$ and $u \in C_d$ there exists $e \in N_c \cap \mathcal{A}$ with

$$(l_c - l_d) \cdot u \leq \frac{1}{\epsilon_G}(l_c - l_e) \cdot u$$

Moreover, $\frac{1}{\epsilon_G}$ can be defined as $4m$.

Proof. We follow the proof for Lemma 2 in Lattimore and Szepesvari (2018) with some modifications to ensure $\frac{1}{\epsilon_G}$ is in the order of $poly(m)$.

Since $u \in C_d$, we have $(l_c - l_d) \cdot u \geq 0$. The result is trivial if c, d are neighbors or $(l_c - l_d) \cdot u = 0$.

Now assume c, d are not neighbors and $(l_c - l_d) \cdot u > 0$. Let v be the centroid of C_c. Consider the line segment connecting u and v. Then let w be the first point on this line segment for which there exists $e \in N_c \cap \mathcal{A}$ with $w \in C_e$ (see Figure 4.1). w is well-defined by the Jordan-Brouwer separation theorem, and e is well-defined because $\mathcal{A}$ is a duplicate-free set of Pareto-optimal classes.

Recall that each class $c' \in \mathcal{A}$ corresponds to a unique partition of $[m]$ into two subsets $A_{c'}$ and $B_{c'}$ such that only objects in $A_{c'}$ contribute to the calculation of negated P@n. For

each $c' \in \mathcal{A}$, we can define $f(c') = [\mathbb{1}(1 \in A_{c'}), ..., \mathbb{1}(m \in A_{c'})]$. Let $\mathbf{R} = [R_1, ..., R_{2^m}]$ collect all relevance vectors: the i-th column of $\mathbf{R}$ is the i-th relevance vector R_i. Then we can rewrite $(l_{c'} - l_{d'}) \cdot u'$ as

$$(l_{c'} - l_{d'}) \cdot u' = (-f(c') \cdot \mathbf{R} + f(d') \cdot \mathbf{R}) \cdot u' = (-f(c') + f(d')) \cdot \mathbf{R}u'$$

for all $c', d' \in \mathcal{A}$ and $u' \in \Delta$. Note that $\mathbf{R}u' = \mathbb{E}_{u'}[R]$ is the expected relevance vector under u'.

Now, using twice $(l_c - l_e) \cdot w = 0$, we calculate

$$\begin{aligned}
(l_c - l_e) \cdot u &= (l_c - l_e) \cdot (u - w) \\
&= (-f(c) + f(e)) \cdot \mathbf{R}(u - w) \\
&= \frac{\|\mathbf{R}(u - w)\|_2}{\|\mathbf{R}(w - v)\|_2} (-f(c) + f(e)) \cdot \mathbf{R}(w - v) \\
&= \frac{\|\mathbf{R}(u - w)\|_2}{\|\mathbf{R}(w - v)\|_2} (l_c - l_e) \cdot (w - v) \\
&= \frac{\|\mathbf{R}(u - w)\|_2}{\|\mathbf{R}(w - v)\|_2} (l_e - l_c) \cdot v > 0
\end{aligned} \tag{4.2}$$

where the third equality uses $w \neq v$ is a point of the line segment connecting v and u, so that $w - v$ and $u - w$ are parallel and have the same direction. Note that $(l_e - l_c) \cdot v > 0$ because c, e are different Pareto-optimal classes and v is the centroid of C_c. $\|\mathbf{R}(w - v)\|_2 = \|\mathbb{E}_w[R] - \mathbb{E}_v[R]\|_2 > 0$ because otherwise, $\mathbb{E}_w[R] = \mathbb{E}_v[R]$ would imply $(l_c - l_e) \cdot v = (-f(c) + f(e)) \cdot \mathbb{E}_v[R] = (-f(c) + f(e)) \cdot \mathbb{E}_w[R] = 0$, contradicting v is the centroid of C_c. To see $\|\mathbf{R}(u - w)\|_2 > 0$, we recalculate $(l_c - l_e) \cdot u$ in another way

$$(l_c - l_e) \cdot u = (l_c - l_e) \cdot (u - w) = \frac{\|u - w\|_2}{\|w - v\|_2} (l_c - l_e) \cdot (w - v) = \frac{\|u - w\|_2}{\|w - v\|_2} (l_e - l_c) \cdot v > 0 \tag{4.3}$$

The inequality in Equation 4.3 holds because $\|u - w\|_2 > 0$ and $\|w - v\|_2 > 0$ (see Figure

4.1). Therefore, $\|\mathbf{R}(u-w)\|_2 > 0$ in Equation 4.2 also holds.

Let $v_{c'}$ be the centroid of $C_{c'}$ for any $c' \in \mathcal{A}$. Then we have

$$\begin{aligned}
\frac{(l_c - l_d)\cdot u}{(l_c - l_e)\cdot u} &= \frac{(l_c - l_d)\cdot (w + u - w)}{(l_c - l_e)\cdot u} \\
&\overset{\text{(a)}}{\leq} \frac{(l_c - l_e)\cdot w + (l_c - l_d)\cdot (u - w)}{(l_c - l_e)\cdot u} \\
&\overset{\text{(b)}}{=} \frac{(l_c - l_d)\cdot (u - w)}{(l_c - l_e)\cdot u} \\
&= \frac{(-f(c) + f(d))\cdot \mathbf{R}(u-w)}{(-f(c) + f(e))\cdot \mathbf{R}u} \\
&\overset{\text{(c)}}{=} \frac{\|\mathbf{R}(w-v)\|_2 \,(-f(c) + f(d))\cdot \mathbf{R}(u-w)}{\|\mathbf{R}(u-w)\|_2 \,(l_e - l_c)\cdot v} \\
&\overset{\text{(d)}}{\leq} \frac{\|\mathbf{R}(w-v)\|_2 \,\|-f(c) + f(d)\|_2}{(l_e - l_c)\cdot v} \\
&= \frac{\|\mathbb{E}_w[R] - \mathbb{E}_v[R]\|_2 \,\|-f(c) + f(d)\|_2}{(l_e - l_c)\cdot v} \\
&\overset{\text{(e)}}{\leq} \frac{2m}{\min_{c'\in\mathcal{C}} \min_{d'\in N_{c'}} (l_{d'} - l_{c'})\cdot v_{c'}}
\end{aligned}$$

where (a) follows since $(l_c - l_d)\cdot w < 0 = (l_c - l_e)\cdot w$, (b) follows since $(l_c - l_e)\cdot w = 0$, (c) follows by Equation 4.2, (d) follows by Cauchy-Schwarz. Note that $0 \preceq \mathbb{E}[R] \preceq 1$, we can bound $\|\mathbb{E}_w[R] - \mathbb{E}_v[R]\|_2$ by $\sqrt{m}$. Since both $f(c)$ and $f(d)$ are binary vectors, we can bound $\|-f(c) + f(d)\|_2$ by $2\sqrt{m}$. Then (e) follows since v is the centroid of C_c and $(l_e - l_c)\cdot v \geq \min_{c'\in\mathcal{C}} \min_{d'\in N_{c'}} (l_{d'} - l_{c'})\cdot v_{c'}$.

Finally, we want to find a lower bound for

$$\min_{c'\in\mathcal{C}} \min_{d'\in N_{c'}} (l_{d'} - l_{c'})\cdot v_{c'} = \min_{c'\in\mathcal{C}} \min_{d'\in N_{c'}} (-f(d') + f(c'))\cdot \mathbb{E}_{v_{c'}}[R]$$

Note that for any $c' \in \mathcal{A}$, $\mathbb{E}_{v_{c'}}[R(i)] = 1$ if object $i \in A_{c'}$ and $\frac{1}{2}$ otherwise. Along with observations from Remark 4, we have

$$(-f(d') + f(c'))\cdot \mathbb{E}_{v_{c'}}[R] = \frac{1}{2}$$

for all $c' \in \mathcal{A}$ and $d' \in N_{c'}$. Therefore, we can bound

$$\frac{2m}{\min_{c' \in \mathcal{C}} \min_{d' \in N_{c'}} (l_{d'} - l_{c'}) \cdot v_{c'}} \leq 4m := \frac{1}{\epsilon_G} \tag{4.4}$$

$\frac{1}{\epsilon_G}$ is clearly a polymonial of m. □

Lemma 17. For each pair of neighboring actions a, b, there exists a function $v^{ab} : \Sigma \times \mathcal{H} \to \mathbb{R}$ such that Definition 8 is satisfied and moreover, $\|v^{ab}\|_\infty = \max_{\sigma \in \Sigma, s \in \mathcal{H}} |v^{ab}(\sigma, s)|$ can be upper bounded by 4.

Proof. Lemma 15 shows for neighboring action pair $\{a, b\}$, the neighborhood action set is $N^+_{a,b} = \{k : 1 \leq k \leq m!, l_k = l_a \text{ or } l_k = l_b\}$ where l_a and l_b are loss vectors of actions a and b respectively.

For top-1 feedback model, each signal matrix $S_{k'}$ is 2 by 2^m. By definition of locally observability (Definition 7), we can write $l_a - l_b$ as

$$l_a - l_b = \sum_{k' \in N^+_{a,b}} \left[c_{k',1} (S^T_{k'})_1 + c_{k',2} (S^T_{k'})_2 \right]$$

where $c_{k',l'} \in \mathbb{R}$ is constant, and $(S^T_{k'})_{l'}$ is the l'-th column of $S^T_{k'}$, for $l' = 1, 2$. Define

$$v^{ab}(\sigma_{k'}, H_{k',k''}) = \left[c_{k',1} (S^T_{k'})_{k'',1} + c_{k',2} (S^T_{k'})_{k'',2} \right], \quad \text{for } 1 \leq k'' \leq 2^m \tag{4.5}$$

where $(S^T_{k'})_{k'',l'}$ is the element in row k'' and column l' of $S^T_{k'}$, for $l' = 1, 2$. Then $l_i - l_j$ can also be written as

$$l_i - l_j = \sum_{k' \in N^+_{a,b}} \begin{bmatrix} v^{ab}(\sigma_{k'}, H_{k',1}) \\ \dots \\ v^{ab}(\sigma_{k'}, H_{k',2^m}) \end{bmatrix}$$

Now back to Equation 4.5, $(S^T_{k'})_{k'',1}$ and $(S^T_{k'})_{k'',2}$ are binary for all k', k''. From the proof for Theorem 5, we can choose $c_{k',1}$ and $c_{k',2}$ such that $|c_{k',1}| \leq 2$ and $|c_{k',2}| \leq 2$ for all k'. Then it follows that $\|v^{ab}\|_\infty = \max_{\sigma \in \Sigma, s \in \mathcal{H}} |v^{ab}(\sigma, s)| \leq 4$, completing the proof. □

Proof for Theorem 6.

Proof. It is easy to see that the time complexity in each round is only $O(poly(K)) = O(poly(m))$. Lemma 16 shows $\frac{1}{\epsilon_G} = 4m$. From Lemma 17, we have $V = \max_{a,b}\|v^{ab}\|_\infty \leq 4$. Then $\frac{V}{\epsilon_G} \leq 16m$ which is $O(poly(m))$. The remaining proof can be found in Lattimore and Szepesvari (2018). □

BIBLIOGRAPHY

Nir Ailon. Improved bounds for online learning over the permutahedron and other ranking polytopes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 29–37. JMLR.org, 2014. URL http://jmlr.org/proceedings/papers/v33/ailon14.html.

András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theor. Comput. Sci.*, 473:77–99, February 2013. ISSN 0304-3975. doi: 10.1016/j.tcs.2012.10.008. URL http://dx.doi.org/10.1016/j.tcs.2012.10.008.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1998. ISSN 0272-5428. doi: 10.1109/SFCS.1995.492488. URL http://ieeexplore.ieee.org/document/492488/.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014. doi: 10.1287/moor.2014.0663. URL https://doi.org/10.1287/moor.2014.0663.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006. doi: 10.1287/moor.1060.0206. URL https://doi.org/10.1287/moor.1060.0206.

Sougata Chaudhuri and Ambuj Tewari. Online learning to rank with top-k feedback. *Journal of Machine Learning Research*, 18(103):1–50, 2017. URL http://jmlr.org/papers/v18/16-285.html.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '03, pages 594–, Washington,

DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-2040-5. URL http://dl.acm.org/citation.cfm?id=946243.946352.

Tor Lattimore and Csaba Szepesvari. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. *arXiv.org*, May 2018.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011. ISBN 978-3-642-14266-6. doi: 10.1007/978-3-642-14267-3. URL https://doi.org/10.1007/978-3-642-14267-3.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss (extended abstract). In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 208–223, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Daiki Suehiro, Kohei Hatano, Shuji Kijima, Eiji Takimoto, and Kiyohito Nagano. Online prediction under submodular constraints. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 260–274, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34106-9.