On the Minimax Regret in Online Ranking with Top-k Feedback

Mingyuan Zhang*
University of Pennsylvania
Philadelphia, PA 19104
myz@seas.upenn.edu

Ambuj Tewari
University of Michigan, Ann Arbor
Ann Arbor, MI 48109
tewaria@umich.edu

Abstract

In online ranking, a learning algorithm sequentially ranks a set of items and receives feedback on its ranking in the form of relevance scores. Since obtaining relevance scores typically involves human annotation, it is of great interest to consider a partial feedback setting where feedback is restricted to the top-k items in the rankings. Chaudhuri and Tewari [2017] developed a framework to analyze online ranking algorithms with top-k feedback. A key element in their work was the use of techniques from partial monitoring. In this work, we further investigate online ranking with top-k feedback and solve 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 and for the following ranking performance measures: Pairwise Loss, Discounted Cumulative Gain, and Precision@n. In addition, we give an efficient algorithm that achieves the minimax regret rate for Precision@n.

1 Introduction

Ranking problems arise in many applications including search engines, recommendation systems, and online advertising (see, e.g., the book by Liu [2011]). The output space in ranking problems consists of permutations of objects. Given the 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), and Precision@n (P@n). Many ranking 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 for 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 work, make excellent starting points for developing more sophisticated online algorithms that can deal with *non-stationarity*. Non-stationarity is a major issue in ranking problems since user preferences can easily change over time.

The basic full feedback setting assumes that the relevance scores, typically obtained via human annotation, provide the correct feedback for each item in the ranking. Since the output in ranking problems is a permutation over a potentially large set of objects, it becomes practically impossible to get full feedback from human annotators. Therefore, researchers have looked into *weak supervision* or *partial feedback* settings where the correct relevance score is only partially revealed to the learning algorithm. For example, Chaudhuri and Tewari [2017] developed a model for online ranking with a particular case of partial feedback called *top-k feedback*. In this model, the online ranking problem is cast as an online partial monitoring game between a learner and an oblivious adversary (who

*This work is based on a dissertation submitted by MZ under AT's supervision in partial fulfillment of the requirements for the degree of Bachelor of Science (Honors Statistics) at the University of Michigan.

Preprint. Under review.

generates a sequence of outcomes before the game begins), played over T rounds.[2] 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 limited feedback at the end of each round: only the relevance scores of the top-k ranked objects are revealed to the learner. Here, $k \leq m$ (in practice $k \ll m$) 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 the upper bounds of the 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). Our focus in this work will be on the non-contextual setting where 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 by 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 adversary's move is uniquely determined by the feedback) nor bandit feedback (where the loss is determined by the feedback); the model falls under the framework of *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 $\Omega(T)$, 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 settings with degenerate actions which are never optimal themselves, but can provide useful information. Lattimore and Szepesvári [2019a,b] finished the characterization of the minimax regret for all partial monitoring games.

Our contributions: We establish the minimax regret rates for all values of k, i.e., $1 \leq k \leq m$ and for ranking measures PL, DCG, and P@n. We obtain these results by showing that the properties of global observability and local observability hold in the appropriate cases. In addition, we provide an algorithm based on the NEIGHBORHOODWATCH2 algorithm of Lattimore and Szepesvári [2019a]. Our algorithm achieves the minimax rate for P@n and has per-round time complexity polynomial in m (for any fixed n).

2 Notations and Problem Setup

We defer all proofs to the appendix.

Our problem formulation follows that of Chaudhuri and Tewari [2017]. 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 $R \in \{0, 1\}^m$ represents relevance for each object. $R(i)$, i-th component of R, is the relevance for object i. Sometimes, relevance values can be multi-graded, i.e., can take values other than 0 or 1. However, in this work, we only study binary relevance.

The learner can choose from $m!$ actions $\{\sigma \mid \sigma : [m] \to [m] \text{ is bijective}\}$ while the adversary can choose from 2^m outcomes $\{0, 1\}^m$. 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 consider an *oblivious* adversary who chooses all the relevance vectors R_t ahead of the game (but they are not revealed to the learner

[2] Some other problems that can be cast as partial monitoring games are multi-armed bandits [Auer et al., 2002], and dynamic pricing [Kleinberg and Leighton, 2003].

at that point).[3] 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 the learner's moves σ_t) defined as the difference in the realized loss and the loss of the best-fixed action in hindsight:

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

When the ranking measure is a gain, we can negate the gain function so that it becomes a loss function. The worst-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 worst-case regret over all strategies of the learner:

$$\mathcal{R}^*_T(G) = \inf_{\pi}\max_{R_1,...,R_T}\mathcal{R}_T(\pi, R_1, ...R_T)\,, \tag{2}$$

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

3 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 if and only if $\sigma^{-1}(i) > \sigma^{-1}(j)$ 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 work are the following.

Pairwise Loss (PL) and Sum Loss (SL)

$$PL(\sigma, R) = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{1}(\sigma^{-1}(i) < \sigma^{-1}(j))\mathbb{1}(R(i) < R(j))\,. \tag{3}$$

$$SL(\sigma, R) = \sum_{i=1}^{m}\sigma^{-1}(i)R(i)\,. \tag{4}$$

It has been shown that the regret under PL is equal to the regret under SL (Section 2 of Ailon [2014]). Therefore, we can study SL instead of PL. The minimax regret rate for PL is the same as that for SL:

$$\sum_{t=1}^{T} PL(\sigma_t, R_t) - \sum_{t=1}^{T} PL(\sigma, R_t) = \sum_{t=1}^{T} SL(\sigma_t, R_t) - \sum_{t=1}^{T} SL(\sigma, R_t)\,. \tag{5}$$

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)

$$DCG(\sigma, R) = \sum_{i=1}^{m}\frac{R(i)}{\log_2(1+\sigma^{-1}(i))}\,. \tag{6}$$

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

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

[3] Note that a stochastic adversary who draws i.i.d. relevance vectors R_t is a special case of an oblivious adversary.

Precision@n Gain (P@n)

$$P@n(\sigma, R) = \sum_{i=1}^{m} \mathbb{1}(\sigma^{-1}(i) \leq n) R(i) \,. \tag{7}$$

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

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

Remark 1. *There are reasons why we are interested in this linear (in R) form of ranking measures. The algorithms that establish upper bounds for the minimax regret rates require the construction of an unbiased estimator of the difference vector between two loss vectors that two different actions incur [Bartók et al., 2014, Lattimore and Szepesvári, 2019a]. The nonlinear (in R) form of ranking measures (including AUC, NDCG, and AP) makes such a construction extremely hard (see details in Chaudhuri and Tewari [2017]).*

4 Summary of Results

We first summarize our main results before delving into technical details. We remind the reader that we assume that the adversary is oblivious and the time horizon is T. We operate under a non-contextual top-k feedback model with m objects.[4]

1. The minimax regret rate for PL, SL and DCG is $\Theta(T^{2/3})$ for $k = 1, 2, ..., m-2$ and is $\Theta(T^{1/2})$ for $k = m-1, m$.
2. The minimax regret rate for P@n is $\Theta(T^{1/2})$ for $1 \leq k \leq m$.
3. The minimax regret rate for P@n can be achieved by an efficient algorithm that needs only $O(\text{poly}(m))$ rounds and has per-round time complexity $O(\text{poly}(m))$.

5 Finite Partial Monitoring Games

Our results on minimax regret rates are developed based on the theory for general finite partial monitoring games developed by Bartók et al. [2014], Lattimore and Szepesvári [2019a,b]. 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 Szepesvári [2019a]. For the sake of easy understanding, we adapt the definitions and notation to our setting.

5.1 A Quick Review of Finite Partial Monitoring Games

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 are 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, $\Delta = \{p \in \mathbb{R}^{2^m} : p \geq 0, \mathbf{1}^\top p = 1\}$ where the inequality between vector is to be interpreted component-wise. Elements of Δ can be treated as

[4] 'Non-contextual' means that there is no side information/features/context associated with the m objects being ranked.

opponent strategies as they are distributions of overall 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 an expected loss not greater than that of any other learner's actions under p.*

Identifying opponent strategies under which an action is optimal 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 the properties of different cells, we can classify corresponding actions as follows.

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 $(2^m - 1)$ dimensions, 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 duplicates of 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 feedback. Thus removing one of them might lead to a loss of information that the learner could have received.

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 to 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 feedback 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., 2000]. Unfortunately, this approach will not work in this setting. There are easy counterexamples (see Exhibit 1 in Appendix H of Lattimore and Szepesvári [2019a]). 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 an unbiased estimator for loss vector difference.

Definition 6 (Global observability)**.** *A pair of actions σ_i and σ_j is called **globally observable** if $l_i - l_j \in \oplus_{1 \leq k' \leq m!} Col(S^\top_{k'})$, where Col refers to column space. The **global observability** condition holds if every pair of neighboring actions is globally observable.*

Definition 7 (Local observability)**.** *A pair of neighboring actions σ_i and σ_j is called **locally observable** if $l_i - l_j \in \oplus_{k' \in N^+_{i,j}} Col(S^\top_{k'})$. 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 feedback from all actions, while local observability means that it can be estimated using just feedback from the neighborhood action set. Clearly, local observability is a stronger condition, and it implies global observability.

We note that the above two definitions are given in Bartók et al. [2014]. Later, when Lattimore and Szepesvári [2019a] extended Bartók et al. [2014]'s work, they proposed different (but equivalent) definitions of *global observability* and *local observability*. We reproduce as follows.

Definition 8 (Alternative definitions of global observability and local observability)**.** *Let $\Sigma = \{\sigma \mid \sigma : [m] \to [m] \text{is bijective}\}$. 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*

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

*A pair of actions σ_i and σ_j are **locally observable** if in addition to the above 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 neighboring 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 (Definition 8) are equivalent to the original definitions of global observability and local observability (Definition 6 and Definition 7), respectively.*

To explain intuitively, 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. Since the two sets of definitions are equivalent, we choose to use the one that is more convenient in the context.

5.2 Classification Theorem for Finite Partial Monitoring Games

To make this work self-contained, 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], Theorem 2 in Lattimore and Szepesvári [2019a], and Theorem 9 in Lattimore and Szepesvári [2019b]] The minimax regret rate of partial monitoring game $G = (L, H)$ satisfies*

$$R^*_T(G) = \begin{cases} 0, & \text{if } G \text{ has no pairs of neighboring actions}; \\ \Theta(T^{1/2}), & \text{if } G \text{ is locally observable and has neighboring actions}; \\ \Theta(T^{2/3}), & \text{if } G \text{ is globally observable, but not locally observable}; \\ \Omega(T), & \text{otherwise}. \end{cases}$$

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 Szepesvári, 2019a,b] contribute to this theorem. In particular, Bartók et al. [2014] summarizes 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). This is important for us since, as we shall see later, the game for P@n contains duplicate actions. Fortunately, Lattimore and Szepesvári [2019a] filled this gap in the literature. Later, Lattimore and Szepesvári [2019b] eliminated the logarithmic dependence on T for locally observable games that appeared in earlier work.

Intuitively, a game is 'trivial' (zero regret) if no learning is needed. This only happens when there is one Pareto-optimal action or all Pareto-optimal actions are duplicates. A game is 'easy' ($T^{1/2}$ regret) if a learner needs to play an action with a small price in order to gain information (for loss

vector difference estimation). A game is 'hard' ($T^{2/3}$ regret) if a learner needs to play an action with a heavy price in order to gain information. A game is 'hopeless' (linear regret) if no information can be obtained to determine optimal actions (for example, two actions with the same feedback).

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

6 Minimax Regret Rates for PL, SL and DCG

We first show the minimax regret rate for a family of ranking loss measures that satisfy the following assumption.

Assumption 1 (Strict increasing property)**.** *The ranking loss measure $RL(\sigma, R)$ can be expressed in the form $f(\sigma) \cdot R$ where $f : \mathbb{R}^m \to \mathbb{R}^m$, is composed of m copies of a univariate strictly increasing scalar-valued function $f^s : \mathbb{R} \to \mathbb{R}$, that is, $\sigma^{-1}(i) > \sigma^{-1}(j)$ implies $f^s(\sigma^{-1}(i)) > f^s(\sigma^{-1}(j))$.*

As mentioned in Section 3, SL satisfies Assumption 1 with $f(\sigma) = \sigma^{-1} = [\sigma^{-1}(1), ..., \sigma^{-1}(m)]$. The negated DCG, which is a ranking loss measure, also satisfies Assumption 1 with $f(\sigma) = [-\frac{1}{\log_2(1+\sigma^{-1}(1))}, ..., -\frac{1}{\log_2(1+\sigma^{-1}(m))}]$.

In the following, assume the ranking loss measure RL satisfies Assumption 1 unless otherwise stated. We first identify the classification of actions and their corresponding cell decomposition.

Lemma 3 (Classification of actions for RL satisfying Assumption 1)**.** *For RL that satisfies Assumption 1, each of the learner's actions σ_i is Pareto-optimal.*

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

Lemma 4 (Neighboring action pair for RL satisfying Assumption 1)**.** *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.*

Remark 2. *From Lemma 4, 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. By properties of RL, we can calculate the R_t entry of $l_i - l_j$ (the entry that corresponds to R_t) as $RL(\sigma_i, R_t) - RL(\sigma_j, R_t) = R_t \cdot (f(\sigma_i) - f(\sigma_j)) = R_t(a) \cdot (f^s(\sigma_i^{-1}(a)) - f^s(\sigma_j^{-1}(a))) + R_t(b) \cdot (f^s(\sigma_i^{-1}(b)) - f^s(\sigma_j^{-1}(b))) = R_t(a) \cdot (f^s(k') - f^s(k'+1)) + R_t(b) \cdot (f^s(k'+1) - f^s(k'))$. We can see that $l_i - l_j$ contains 2^{m-1} nonzero entries, of which 2^{m-2} entries are $f^s(k') - f^s(k'+1)$ and 2^{m-2} entries are $f^s(k'+1) - f^s(k')$. Moreover, if $R_t(a) = 1$ and $R_t(b) = 0$ for the R_t relevance, then the R_t entry of $l_i - l_j$ is $f^s(k') - f^s(k'+1)$. If $R_t(a) = 0$ and $R_t(b) = 1$ for the R_t relevance, then the R_t entry of $l_i - l_j$ is $f^s(k'+1) - f^s(k')$. If $R_t(a) = R_t(b)$ for the R_t relevance, then the R_t entry of $l_i - l_j$ is 0.*

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

Lemma 5 (Neighborhood action set for RL satisfying Assumption 1)**.** *For neighboring action pair $\{\sigma_i, \sigma_j\}$, the neighborhood action set is $N_{i,j}^+ = \{i, j\}$, so $\oplus_{k \in N_{i,j}^+} Col(S_k^\top) = Col(S_i^\top) \oplus Col(S_j^\top)$.*

We are now ready to state the first important theorem in this work.

Theorem 6 (Local observability for RL satisfying Assumption 1)**.** *Under top-k feedback model with m objects, with respect to loss matrix L and feedback matrix H for RL satisfying Assumption 1, the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$.*

Since SL satisfies Assumption 1, we have the following corollary from Theorem 6.

Corollary 7 (Local observability for SL)**.** *With respect to loss matrix L and feedback matrix H for SL, the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$.*

Minimax regret rates for SL and PL. 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 6 and chaining with Theorem 2, we immediately have the minimax regret for SL:

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

By Equation (5), PL has exactly the same minimax regret rates as SL.

Discussion. Corollary 7 shows that 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 the failure of local observability only for $k = 1$.

As for the time complexity, Chaudhuri and Tewari [2017] provided an efficient (polynomial of m time) algorithm for PL and SL for values of k when global observability holds, so we have an 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 a large value of k is not interesting in practice, so we do not pursue this question.

Similarly, since negated DCG also satisfies Assumption 1, we have the following corollary from Theorem 6.

Corollary 8 (Local observability for DCG)**.** *With respect to loss matrix L and feedback matrix H for DCG, the local observability fails for $k = 1, ..., m-2$ and holds for $k = m-1, m$.*

Minimax regret rate for DCG. Corollary 10 of Chaudhuri and Tewari [2017] showed that for DCG, the minimax regret rate is $O(T^{2/3})$ for $1 \leq k \leq m$. Combining with our Corollary 8 and chaining with Theorem 2, we immediately have the minimax regret for DCG:

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

Discussion. Corollary 8 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 $k = m-1$.

7 Minimax Regret Rate for P@n

The negated P@n does not satisfy Assumption 1 because f^s is not strictly increasing (see Eq. (7)), so Theorem 6 does not apply to negated P@n.

In the following, the ranking loss measure is negated P@n unless otherwise stated. To establish a minimax regret rate for P@n, we first need to identify the classification of actions and their corresponding cell decomposition.

Lemma 9 (Classification of actions for P@n)**.** *For negated P@n, each of the learner's actions σ_i is Pareto-optimal.*

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

Lemma 10 (Neighboring action pairs for P@n)**.** *For negated P@n, a pair of learner's actions $\{\sigma_i, \sigma_j\}$ is a neighboring action pair if and only 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 $A_i = \{a : \sigma_i^{-1}(a) \leq n\}$, $B_i = \{b : \sigma_i^{-1}(b) > n\}$, and A_j and B_j are defined similarly.*

Remark 3. *From Lemma 10, 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 10. Using the definition of negated P@n, 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.*

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

Lemma 11 (Neiborhood action set for P@n)**.** *For neighboring action pair $\{\sigma_i, \sigma_j\}$, the neighborhood action set is $N^+_{i,j} = \{k : 1 \le k \le m!, l_k = l_i \text{ or } l_k = l_j\}$.*

Remark 4. *Negated P@n says that it only matters the way of partitioning m objects into 2 sets A and B as in Lemma 9. 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 are prepared to state the local observability theorem for P@n. The proof uses the same technique as that in the proof of Theorem 6.

Theorem 12 (Local observability for P@n)**.** *For fixed n such that $1 \le n \le m$, with respect to loss matrix L and feedback matrix H for P@n, the local observability holds for all $1 \le k \le m$.*

Minimax regret rate for P@n. Note that this game contains many duplicate actions (but no degenerate actions) 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, Lattimore and Szepesvári [2019a,b] have proved that locally observable games enjoy $\Theta(T^{1/2})$ minimax regret, regardless of the existence of duplicate actions. This shows the minimax regret for P@n is

$$R^*_T = \Theta(T^{1/2}) \quad \text{for } 1 \le k \le m\,.$$

Discussion. We note that Chaudhuri and Tewari [2017] only showed $O(T^{2/3})$ 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.

8 Efficient Algorithm for Obtaining Minimax Regret Rate for P@n

Lattimore and Szepesvári [2019a] showed 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.[5] In particular, they showed that NEIGHBORHOOD-WATCH2 achieves $\Theta(T^{1/2})$ minimax regret when there are no degenerate actions.[6] 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 $O(\mathrm{poly}(m))$ time per round and obtains a minimax regret rate of $\Theta(T^{1/2})$. 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 of 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.

Lemma 10 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 duplicated 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 any 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.

[5] $\widetilde{\Theta}(\cdot)$ indicates growth up to logarithmic factors.

[6] This result appeared in the arXiv version of Lattimore and Szepesvári [2019a]. See `https://arxiv.org/abs/1805.09247` for Lattimore and Szepesvári [2018].

Algorithm 1 NEIGHBORHOODWATCH2

1: **Input** L, H, η, γ
2: **for** $t = 1, \ldots, T$ **do**
3: For $a, k \in \mathcal{C}$, let
4: $$Q_{tka} = \mathbb{1}_{\mathcal{A}}(k) \frac{\mathbb{1}_{N_k \cap \mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{Z}_{ska}\right)}{\sum_{b \in N_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \widetilde{Z}_{skb}\right)} + \mathbb{1}_{\mathcal{D}}(k) \frac{\mathbb{1}_{\mathcal{A}}(a)}{\mid \mathcal{A} \mid}$$
5: Find distribution $\widetilde{P}_t$ such that $\widetilde{P}_t^\top = \widetilde{P}_t^\top Q_t$
6: Compute $P_t = (1-\gamma)\widetilde{P}_t + \frac{\gamma}{|\mathcal{C}|}\mathbf{1}$
7: Sample $A_t \sim P_t$ and receive feedback Φ_t
8: Compute loss-difference estimators for each $k \in \mathcal{A}$ and $a \in N_k \cap \mathcal{A}$:
9: $\widehat{Z}_{tka} = \frac{\widetilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}}$,
10: $\beta_{tka} = \eta V^2 \sum_{b \in N_{ak}^+} \frac{\widetilde{P}_{tk}^2}{P_{tb}}$, and
11: $\widetilde{Z}_{tka} = \widehat{Z}_{tka} - \beta_{tka}$
12: **end for**

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 track of a distribution to sample from $n\binom{m}{n}$ (a polynomial of m for any fixed n) actions, instead of sampling from $m!$ actions.

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 $\mid \mathcal{A} \mid = \binom{m}{n}$ and $\mathcal{A}$ contain 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 of a and a's neighbors. To make this section self-contained, we include the algorithm NEIGHBORHOODWATCH2 in Lattimore and Szepesvári [2019a] with some changes so that it is consistent with our notations. See Algorithm 1.[7]

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 $\max_{\sigma \in \Sigma, s \in \mathcal{H}} \mid v^{ab}(\sigma, s) \mid$. The following lemma shows $\|v^{ab}\|_\infty \leq 4$ for suitable choice of v^{ab}, so V can be upper bounded by 4.

Lemma 13 (Upper bound for $\|v^{ab}\|_\infty$)**.** *For each pair of neighboring actions a, b for the ranking loss measure (negated) P@n, there exists a function $v^{ab} : \Sigma \times \mathcal{H} \to \mathbb{R}$ such that Definition 8 is satisfied and moreover, $\|v^{ab}\|_\infty = \max_{\sigma \in \Sigma, s \in \mathcal{H}} \mid v^{ab}(\sigma, s) \mid$ can be upper bounded by 4.*

Below, we provide an upper bound on the regret when running Algorithm 1 for the top-k feedback model with P@n.

Theorem 14 (Lattimore and Szepesvári [2018, 2019a])**.** *Let $K = \mid \mathcal{C} \mid = n\binom{m}{n}$. For top-1 feedback model with P@n, suppose Algorithm 1 is run on $G = (L, H)$ with $\eta = \frac{1}{V}\sqrt{\log(K)/T}$ and $\gamma = \eta K V$. Then*

$$\mathbb{E}[\mathcal{R}_T] \leq O\Big(\frac{KV}{\epsilon_G}\sqrt{T \log(K)}\Big),$$

where ϵ_G is a constant specific to the game G, not depending on T, and $\frac{V}{\epsilon_G} \leq 8m$. Using a loose bound $\binom{m}{n} \leq m^n$, we have

$$\mathbb{E}[\mathcal{R}_T] \leq O\Big(n m^{n+1}\sqrt{T(\log(n) + n\log(m))}\Big).$$

So the regret is upper bounded by $O(\mathrm{poly}(m) \cdot T^{1/2})$ for any fixed n. Moreover, the time complexity in each round is $O(\mathrm{poly}(m))$. Since top-k (for $k > 1$) feedback contains strictly more information than top-1 feedback does, this result applies to the general top-k feedback model with P@n as well.

[7] The original NEIGHBORHOODWATCH2 algorithm, which this algorithm is based on, requires a REDISTRIBUTE function at line 6 to handle degenerate actions. Since there are no degenerate actions in this game, we leave it out.

Before proving this theorem, the following lemma shows that $\frac{1}{\epsilon_G}$ can be defined as $2m$ in this setting.

Lemma 15 (Lemma 6 in Bartók et al. [2014], and Lemma 5 in Lattimore and Szepesvári [2019a])**.** *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 taken as $2m$.

Proof of Theorem 14.

Proof. It is easy to see from Algorithm 1 that the time complexity in each round is $O(\mathrm{poly}(K)) = O(\mathrm{poly}(m))$. Lemma 15 shows $\frac{1}{\epsilon_G} = 2m$. From Lemma 13, we have $V = \max_{a,b}\|v^{ab}\|_\infty \leq 4$. Then $\frac{V}{\epsilon_G} \leq 8m$. The remaining proof follows Lattimore and Szepesvári [2018, 2019a]. □

Discussion. We have provided an upper bound on the regret in terms of m, n, T. Yet, doing the same for a lower bound is tricky (a known lower bound is $\Omega(T^{1/2})$), so we leave it as future work. Lattimore and Szepesvári [2020] has some discussion on lower bounds for locally observable partial monitoring games in terms of the number of actions and the number of outcomes. We refer interested readers to check for further details.

The regret bound for NEIGHBORHOODWATCH2 involves a game-dependent constant $\frac{1}{\epsilon_G}$ which can be arbitrarily large [Lattimore and Szepesvári, 2019a], even though in our setting, we can upper bound it by $2m$. In the follow-up work, Lattimore and Szepesvári [2019b] established new regret bounds that are independent of arbitrarily large game-dependent constants. However, their approach is non-constructive. Then, Lattimore and Szepesvári [2020] provided an algorithm so that for locally observable games without degenerate actions (as is the case in our current setting), its regret upper bound matches the best-known information-theoretical upper bound shown in Lattimore and Szepesvári [2019b]. When running their algorithm with our current setting, the regret upper bound is $O(K^{3/2}\sqrt{T\log(K)})$, where $K = n\binom{m}{n}$. Comparing it with our bound $O(mK\sqrt{T\log(K)})$ shown in Theorem 14, we can see that our bound is better in our current setting.

Recently, new algorithms based on mirror-decent and information-directed sampling have been proposed for partial monitoring games [Tsuchiya et al., 2020, Lattimore and György, 2021, Lattimore, 2022]. They could potentially be applied in our current setting as well.

9 Conclusion

In this work, we have successfully closed one of the most interesting open questions proposed by Chaudhuri and Tewari [2017]: we have established a full characterization of minimax regret rates with top-k feedback model for all k for ranking measures Pairwise Loss (PL), Discounted Cumulative Gain (DCG) and Precision@n Gain (P@n).

For PL and DCG, we have improved the results in Chaudhuri and Tewari [2017] and ruled out the possibility of better regret for values of k that are practically interesting. For P@n, which is widely used in learning to rank community, we have shown a surprisingly good regret of $\Theta(T^{1/2})$ for all k, which improved the original regret $O(T^{2/3})$ in Chaudhuri and Tewari [2017]. Moreover, we have provided an efficient algorithm that achieves this regret rate.

References

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*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 29–37. JMLR.org, 2014.

András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theor. Comput. Sci.*, 473:77–99, 2013.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Electron. Colloquium Comput. Complex.*, TR00-068, 2000.

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

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Math. Oper. Res.*, 31(3):562–580, 2006.

Sougata Chaudhuri and Ambuj Tewari. Online learning to rank with top-k feedback. *J. Mach. Learn. Res.*, 18:103:1–103:50, 2017.

Robert D. Kleinberg and Frank Thomson Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Symposium on Foundations of Computer Science 2003)*, pages 594–605. IEEE Computer Society, 2003.

Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1547–1575. PMLR, 2022.

Tor Lattimore and András György. Mirror descent and the information ratio. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2965–2992. PMLR, 2021.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. *CoRR*, abs/1805.09247, 2018.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory, ALT 2019*, volume 98 of *Proceedings of Machine Learning Research*, pages 529–556. PMLR, 2019a.

Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory, COLT 2019*, volume 99 of *Proceedings of Machine Learning Research*, pages 2111–2139. PMLR, 2019b.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 2488–2515. PMLR, 2020.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001*, volume 2111 of *Lecture Notes in Computer Science*, pages 208–223. Springer, 2001.

Daiki Suehiro, Kohei Hatano, Shuji Kijima, Eiji Takimoto, and Kiyohito Nagano. Online prediction under submodular constraints. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012*, volume 7568 of *Lecture Notes in Computer Science*, pages 260–274. Springer, 2012.

Taira Tsuchiya, Junya Honda, and Masashi Sugiyama. Analysis and design of thompson sampling for stochastic partial monitoring. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

A Proofs for Section 5

Proof of Lemma 1.

Proof. We first prove that 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 that a pair of actions σ_i and σ_j is globally observable if $l_i - l_j \in \oplus_{1 \le k' \le m!} \text{Col}(S_{k'}^\top)$. So we can write $l_i - l_j$ as

$$l_i - l_j = \sum_{k'=1}^{m!} \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^\top)_{l'} \,,$$

where $c_{k',l'} \in \mathbb{R}$ is constant, and $(S_{k'}^\top)_{l'}$ is the l'-th column of $S_{k'}^\top$. Define

$$f(\sigma_{k'}, H_{k',k''}) = \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^\top)_{k'',l'} \quad \text{for } 1 \le k'' \le 2^m \,, \tag{8}$$

where $(S_{k'}^\top)_{k'',l'}$ is the element in row k'' and column l' of $S_{k'}^\top$. 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.

Now, we prove the other direction. That is, the original definition of global observability (Definition 6) generalizes the alternative definition (Definition 8). It will follow that the original definition of local observability (Definition 7) generalizes the alternative definition (Definition 8).

Definition 8 says that a pair of actions σ_i and σ_j is globally observable if there exists a function $f : \Sigma \times \mathcal{H} \to \mathbb{R}$ such that

$$\sum_{k'=1}^{m!} f(\sigma_{k'}, H_{k',k''}) = L_{i,k''} - L_{j,k''} \quad \text{for all } 1 \le k'' \le 2^m \,. \tag{9}$$

Note that in this setting (online learning with top-k feedback), $\mid \Sigma \mid = m!$ and $\mid \mathcal{H} \mid = 2^k$, so function f has a finite domain. Using the definition of signal matrices (Definition 5), we see that in Eq. (8),

$$(S_{k'}^\top)_{k'',l'} = (S_{k'})_{l',k''} = \mathbb{1}(H_{k',k''} = s_{l'}) \,,$$

where $s_{l'}$ is the l'-th symbol in $\mathcal{H}$. So, among all $(S_{k'}^\top)_{k'',l'}$, $1 \le l' \le 2^m$, exactly one of them is 1. Therefore, we can let $c_{k',l'} = f(\sigma_{k'}, H_{k',k''})$ for l' such that $H_{k',k''} = s_{l'}$, and $c_{k',l'} = 0$ otherwise. Then, Eq. (9) can be written as

$$L_{i,k''} - L_{j,k''} = \sum_{k'=1}^{m!} \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^\top)_{k'',l'} \quad \text{for all } 1 \le k'' \le 2^m \,.$$

So,

$$l_i - l_j = \sum_{k'=1}^{m!} \begin{bmatrix} \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^\top)_{1,l'} \\ \dots \\ \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^\top)_{2^m,l'} \end{bmatrix} = \sum_{k'=1}^{m!} \sum_{l'=1}^{2^k} c_{k',l'} (S_{k'}^\top)_{l'} \,,$$

satisfying Definition 6. □

B Proofs for Section 6

Proof of Lemma 3.

Proof. 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. Since $f(\sigma_i)$ is an element-wise strictly increasing transformation

of σ_i^{-1}, 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 $C_i = \{p \in \Delta : \mathbf{1}^\top p = 1, \mathbb{E}[R(\sigma_i(1))] \geq \mathbb{E}[R(\sigma_i(2))] \geq ... \geq \mathbb{E}[R(\sigma_i(m))]\}$. Note that C_i is non-empty. C_i has only one equality constraint and hence has dimension $(2^m - 1)$. This shows σ_i is Pareto-optimal. □

Proof of Lemma 4.

Proof. We first prove the "if" part.

From Lemma 3, we know each of learner's actions σ_i is Pareto-optimal and it has cell

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

Let $\sigma_i(k) = a$, $\sigma_i(k+1) = b$, $\sigma_j(k+1) = a$ and $\sigma_j(k) = b$ for some $k \in [m-1]$. For all $k' \in [m] \setminus \{k, k+1\}$, let $\sigma_i(k') = \sigma_j(k')$. Then

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

Therefore, there are two equalities in $C_i \cap C_j$: $\mathbf{1}^\top p = 1$ and $\sum_{j=1}^{2^m} p_j(R_j(\sigma_i(k)) - R_j(\sigma_i(k+1))) = 0$. These two equations define two non-parallel hyperplanes, so the intersection of the two hyperplanes has dimension $(2^m - 2)$. By Definition 4, σ_i and σ_j are neighboring actions.

We then 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 3, σ_i has cell

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

and σ_j has cell

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

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 $\mathbf{1}^\top p = 1$), 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 the argument for case 1 to case 2 shows $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair. □

Proof of Lemma 5.

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 3, for RL 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 RL, $l_i \neq l_k$ for $i \neq k$. Moreover, $l_i - l_k$ is independent of $\mathbf{1}$ for $i \neq k$ because each action is Pareto-optimal. 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 $\oplus_{k \in N^+_{i,j}} \text{Col}(S_k^\top) = \text{Col}(S_i^\top) \oplus \text{Col}(S_j^\top)$. □

Proof of Theorem 6.

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 for the top-k feedback model (defined according to Definition 5) 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 4, 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 $[0,\ f^s(m) - f^s(m-1),\ f^s(m-1) - f^s(m),\ 0]$ (see Remark 2). 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 1 shows $l_1 - l_2$ and two types of rows of S for each group. Since f^s is strictly increasing, it is clear that $l_1 - l_2 \notin \text{Col}(S^\top)$. 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	$f^s(m) - f^s(m-1)$	$f^s(m-1) - f^s(m)$	0
rows of S	1 0	1 0	1 0	1 0

Table 1: Part 1, within the group, $R(c)$ is the same for all $c \neq m-1, m$. See proof of Theorem 6 for more details.

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 $l_i - l_j$ into 2^{m-2} groups. Then each group is $[0,\ f^s(k'+1) - f^s(k'),\ f^s(k') - f^s(k'+1),\ 0]$. 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 2. It is clear that the piece $[0,\ f^s(k'+1) - f^s(k'),\ f^s(k') - f^s(k'+1),\ 0]$ is in the row space of S. In this case, $l_i - l_j \in \text{Col}(S_i^\top) \oplus \text{Col}(S_j^\top)$.
(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 3. The piece $[0,\ f^s(k'+1) - f^s(k'),\ f^s(k') - f^s(k'+1),\ 0]$ is in the row space of S because $[0,\ f^s(k'+1) - f^s(k'),\ f^s(k') - f^s(k'+1),\ 0] = (f^s(k'+1) - f^s(k'))[1\ 1\ 0\ 0] - (f^s(k'+1) - f^s(k'))[1\ 0\ 1\ 0]$. In this case, $l_i - l_j \in \text{Col}(S_i^\top) \oplus \text{Col}(S_j^\top)$.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	$f^s(k'+1) - f^s(k')$	$f^s(k') - f^s(k'+1)$	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 2: Part 2 (1), within the group, $R(c)$ is the same for all $c \neq a, b$. See proof of Theorem 6 for more details.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	0	$f^s(k'+1) - f^s(k')$	$f^s(k') - f^s(k'+1)$	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 3: Part 2 (2), within the group, $R(c)$ is the same for all $c \neq a, b$. See proof of Theorem 6 for more details.

In either case, we have $l_i - l_j \in \text{Col}(S_i^\top) \oplus \text{Col}(S_j^\top)$, so $\{\sigma_i, \sigma_j\}$ is locally observable. Hence the local observability holds, concluding the proof. □

C Proofs for Section 7

Proof of Lemma 9.

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, $C_i = \{p \in \Delta : \mathbf{1}^\top p = 1, \mathbb{E}[R(a)] \geq \mathbb{E}[R(b)], \forall a \in A_i, \forall b \in B_i\}$. C_i has only one equality constraint and hence has dimension $(2^m - 1)$. This shows action σ_i is Pareto-optimal. □

Proof of Lemma 10.

Proof. For the "if" part, assume the condition (*there is exactly* ...) holds. From Lemma 9, 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 $\mathbf{1}^\top p = 1$), 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, $\mid A_i \mid = n$ and $\mid B_i \mid = m - n$ for all action σ_i. There are two cases. 1. $\mid A_i \setminus A_j \mid = 0$. 2. $\mid A_i \setminus A_j \mid > 1$.

For the first case, if $\mid A_i \setminus A_j \mid = 0$, 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 9. Thus, in this case, $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair.

For the second case, if $\mid A_i \setminus A_j \mid > 1$, 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 $\mathbf{1}^\top p = 1$), and hence it has dimension less than $(2^m - 2)$. Thus, in this case, $\{\sigma_i, \sigma_j\}$ is not a neighboring action pair. □

Proof of Lemma 11.

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 9, 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 $N^+_{i,j} = \{k : 1 \leq k \leq m!, l_k = l_i \text{ or } l_k = l_j\}$. □

Proof of Theorem 12.

Proof. It suffices to show the local observability holds for $k = 1$ because there is strictly more information for the 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 3. 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 $l_i - l_j$ 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 $2n!(m-n)!$ [8] signal matrices S_l where $l \in N^+_{i,j}$ by row and denote the resultant matrix by S. S is of size $4n!(m-n)! \times 2^m$. Now in each group, there are only five types of rows of S, as shown in Table 4. $[1\ 1\ 0\ 0]$ 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, $l_i - l_j \in \oplus_{l \in N^+_{i,j}} \text{Col}(S_l^\top)$, so $\{\sigma_i, \sigma_j\}$ is locally observable and the local observability holds for P@n. □

[8] See Remark 4 for how this number is calculated.

	$R(a)=0$ $R(b)=0$	$R(a)=0$ $R(b)=1$	$R(a)=1$ $R(b)=0$	$R(a)=1$ $R(b)=1$
$l_i - l_j$	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: P@n, within the group, $R(c)$ is the same for all $c \neq a, b$. See proof of Theorem 12 for more details.

D Proofs for Section 8

In this section, the loss function is negated P@n unless otherwise stated. We consider top-1 feedback model as described in Section 8.

Proof of Lemma 13.

Proof. Lemma 11 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 local observability (Definition 7), we can write $l_a - l_b$ as

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

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

$$v^{ab}(\sigma_{k'}, H_{k',k''}) = \left[c_{k',1}(S^\top_{k'})_{k'',1} + c_{k',2}(S^\top_{k'})_{k'',2} \right], \quad \text{for } 1 \leq k'' \leq 2^m, \tag{10}$$

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

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

Now back to Equation (10), $(S^\top_{k'})_{k'',1}$ and $(S^\top_{k'})_{k'',2}$ are binary for all k', k''. From the proof for Theorem 12, we can choose $c_{k',1}$ and $c_{k',2}$ such that $\mid c_{k',1} \mid \leq 2$ and $\mid c_{k',2} \mid \leq 2$ for all k'. Then it follows that $\|v^{ab}\|_\infty = \max_{\sigma \in \Sigma, s \in \mathcal{H}} \mid v^{ab}(\sigma, s) \mid \leq 4$, completing the proof. □

Proof of Lemma 15.

Proof. We follow the proof for Lemma 5 in Lattimore and Szepesvári [2019a] with some modifications to ensure $\frac{1}{\epsilon_G}$ is in the order of $\text{poly}(m)$.

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 $(l_c - l_d) \cdot u = 0$.

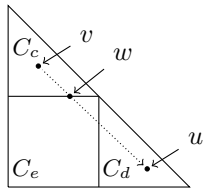

Figure 1: Illustrating proof for Lemma 15, adopted from Lattimore and Szepesvári [2019a].

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 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

$$\begin{aligned}
(l_c - l_e) \cdot u &= (l_c - l_e) \cdot (u - w) \\
&= (-f(c) + f(e)) \cdot \mathbf{R}(u - w) \\
&= \frac{\|\mathbf{R}(u - w)\|_2}{\|\mathbf{R}(w - v)\|_2} (-f(c) + f(e)) \cdot \mathbf{R}(w - v) \\
&= \frac{\|\mathbf{R}(u - w)\|_2}{\|\mathbf{R}(w - v)\|_2} (l_c - l_e) \cdot (w - v) \\
&= \frac{\|\mathbf{R}(u - w)\|_2}{\|\mathbf{R}(w - v)\|_2} (l_e - l_c) \cdot v > 0
\end{aligned} \tag{11}$$

where the third equality uses that $w \neq v$ 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 that 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

$$\begin{aligned}
(l_c - l_e) \cdot u &= (l_c - l_e) \cdot (u - w) \\
&= \frac{\|u - w\|_2}{\|w - v\|_2} (l_c - l_e) \cdot (w - v) \\
&= \frac{\|u - w\|_2}{\|w - v\|_2} (l_e - l_c) \cdot v > 0 \,.
\end{aligned} \tag{12}$$

The inequality in Equation (12) holds because $\|u - w\|_2 > 0$ (since $(l_c - l_d) \cdot u > 0$) and $\|w - v\|_2 > 0$. Therefore, $\|\mathbf{R}(u - w)\|_2 > 0$ in Equation (11) also holds.

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

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

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 (11), (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 $\sqrt{m}$. Then (e) follows since v is the centroid of C_c and $(l_e - l_c)\cdot v \geq \min_{c'\in\mathcal{A}} \min_{d'\in N_{c'}\cap\mathcal{A}} (l_{d'} - l_{c'})\cdot v_{c'}$ ($v_{c'}$ is the centroid of $C_{c'}$).

Finally, we want to find a lower bound for

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

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 (by the symmetry property of the centroid $v_{c'}$ of $C_{c'}$). Along with observations from Remark 3, we have

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

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

$$\frac{m}{\min_{c'\in\mathcal{A}} \min_{d'\in N_{c'}\cap\mathcal{A}} (l_{d'} - l_{c'})\cdot v_{c'}} \leq 2m := \frac{1}{\epsilon_G}\,. \tag{13}$$

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