Rank Aggregation and Prediction with Item Features

Kai-Yang Chiang
Dept of Computer Science
UT Austin

Cho-Jui Hsieh
Dept of Statistics and Computer Science
UC Davis

Inderjit S. Dhillon
Dept of Computer Science
UT Austin

Abstract

We study the problem of rank aggregation with features, where both pairwise comparisons and item features are available to help the rank aggregation task. Observing that traditional rank aggregation methods disregard features, while models adapted from learning-to-rank task are sensitive to feature noise, we propose a general model to learn a total ranking by balancing between comparisons and feature information jointly. As a result, our proposed model takes advantage of item features and is also robust to noise. More importantly, we study the effectiveness of item features in our model and show that given sufficiently informative features, the sample complexity of our model can be asymptotically lower than models based only on comparisons for deriving an accurate ranking. The results theoretically justify that our model can achieve efficient learning by leveraging item feature information. In addition, we show that the proposed model can also be extended to two other related problems—online rank aggregation and rank prediction of new items. Finally, experiments show that our model is more effective and robust compared to existing methods on both synthetic and real datasets.

1 Introduction

Ranking is a fundamental problem in machine learning. Given n items with partial ranking information, the goal of rank aggregation is to obtain a full ranking that is consistent with most of the partial rankings. One classical setting is to consider pairwise comparisons, where each partial ranking gives a list of pairwise preferences. This pairwise rank aggregation problem has been shown to be important in real-world applications including ranking for sports teams [19] and recommender systems [12].

While a number of pairwise rank aggregation methods have been proposed, many of them learn a ranking based solely on item comparisons. Nevertheless, in many real-world applications, knowledge about items is also provided, and such knowledge is believed to be related to the rank of the items as well. For instance, when ranking sports teams, attributes of each team, such as its coach and budget, can all be potential factors that affect its rank beside competition history. Therefore, in this paper we focus on pairwise rank aggregation *with item features*. Our goal is to derive a better total ranking based on both pairwise comparisons and feature information.

Though most rank aggregation methods do not take item features into consideration, several models designed for other purposes can in fact be adapted to rank aggregation if features are provided. For example, Rank-SVM [16] also trains a model based on both comparisons and features and can be applied to this problem under a transductive setting. However, since such learning-to-rank models are originally designed to rank a list of unranked items based on their attributes, they tend to learn a ranking that heavily depends on item features, and as a result, they could perform poorly if features are noisy or partially corrupted, even if the comparison information is clean.

Motivated by the fact that current methods for rank aggregation are either ignorant of features or sensitive to noisy features, we propose a novel model that better learns ranking scores from features and comparisons *simultaneously*. Furthermore, we formally analyze the effect of features and provide sample complexity guarantees of our model. In particular, with informative features, we show that our model only requires $o(n)$ comparisons—i.e. sublinear in the number of items—to obtain an accurate ranking. We emphasize that since $\Omega(n)$ is the sample complexity lower bound for any aggregation method based only on com-

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

parisons [11, 28], our result suggests that such $\Omega(n)$ barrier can be asymptotically overcome by taking advantage of reasonably good features. Finally, we provide extensions of our model to show that by utilizing item features, the proposed framework can improve performance not only on rank aggregation but also on two related problems—online rank aggregation and rank prediction for new items. Our contributions can be summarized as follows:

- We propose a new model for rank aggregation where the ranking is estimated by balancing both feature and pairwise comparison information.
- We formally define feature quality in the context of ranking and show that given reasonably good features, the sample complexity can be improved from linear to sublinear using our model.
- We extend our model to online rank aggregation and rank prediction of new items, and show that our framework can be useful in these tasks by leveraging feature information.

2 Problem Setup and Related Work

Problem Setup of Rank Aggregation with Features. Let $\mathbf{s} \in \mathbb{R}^n$ be a (true) score vector for a set of n items, where s_i is the ranking score of item i. An item i is regarded as better (i.e. has a higher rank) than item j if $s_i > s_j$. Let $\mathbf{e}$ be the all-one vector and $Y = \mathbf{e}\mathbf{s}^T - \mathbf{s}\mathbf{e}^T$ where each $Y_{ij} = s_j - s_i$ is the score difference between item i and item j. A pairwise comparison P_{ij} is observed under two scenarios: [1]

- For each comparison, the *value* of the score difference is revealed: $P_{ij} = Y_{ij}$.
- For each comparison, only the *sign* of the rank difference is revealed: $P_{ij} = \text{sgn}(Y_{ij})$.

Let m be the number of observed comparisons, with the set of indices $\mathcal{S}_I = \{(i_t, j_t)\}_{t=1}^m$. Let $\mathbf{x}_i \in \mathbb{R}^d$ be the feature of item i. The item features can be assembled into a matrix $X \in \mathbb{R}^{n\times d}$, where the i-th row of X is $\mathbf{x}_i^T$. With these notations, the problem of rank aggregation with features can be stated as follows:

> Given n items, a set of observed comparisons $\{P_{ij} \mid (i,j) \in \mathcal{S}_I\}$ and a feature matrix X, the goal is to "recover" the ranking of $\mathbf{s}$.

Evaluation Metric for Recovery. Generally, it is impossible to recover the exact score of $\mathbf{s}$ with only pairwise comparisons due to an identifiability issue. [2] However, in ranking applications, only the *relative order* between items matters, not the exact score. Thus, we use the standard Kendall's Tau metric to measure the ranking distance between π and $\mathbf{s}$:

$$D_{k\tau}(\pi, \mathbf{s}) = \frac{1}{N}\sum_{s_i<s_j} \mathbf{1}(\pi_i > \pi_j), \quad N = \sum \mathbf{1}(s_i < s_j).$$

Therefore, to argue the effectiveness of an algorithm, the most ideal scenario is to show that its output π exactly recovers the ranking of $\mathbf{s}$ (i.e. $D_{k\tau}(\pi, \mathbf{s}) = 0$) given only a small number of observed comparisons. However, such a goal is still too prohibitive since in certain scenarios $\Omega(n^2)$ clean comparisons are still required to achieve exact recovery [15]. To provide non-trivial results, one popular metric is to consider "ϵ-recovery" as an approximate recovery scheme instead [15, 24, 28], where the goal is relaxed to derive an "ϵ-accurate" ranking π such that $D_{k\tau}(\pi, \mathbf{s}) < \epsilon$ for any chosen tolerance $\epsilon > 0$. We will name several interesting results of ϵ-recovery in the next subsection.

Related Work. Pairwise rank aggregation has received much attention in many areas. Popular approaches include direct learning on ranking scores [19, 28], probabilistic models [4, 20] and Markov Chain heuristics [22]. Traditional aggregation methods are designed to learn a total ranking based only on item-to-item comparisons and do not consider item features.

Several feature-based ranking models can also be adapted to rank aggregation if features are present. A major class of such models is learning-to-rank models, which were originally built to rank a list of new items. As examples, Rank-SVM [16], RankNet [6] and RankRLS [23] all train a model with both features and pairwise comparisons and can be applied to this problem under a transductive setting. However, compared to our model, these models have no theoretical guarantees when adapted to rank aggregation, and they can indeed fail to recover the underlying ranking both in theory and practice (see Section 6 in detail). A recently proposed model for dyad ranking [27] also learns a ranking from both features and item-specific scores, whose motivation is similar to our model. However, the authors do not provide theoretical guarantees of this model on dyad rank aggregation either.

One highlight of our model is its improved sample complexity guarantee. Sample complexity analysis for rank aggregation has received more attention recently, where the goal is to study the number of comparisons required to guarantee the derived ranking to be accurate. For example, Gleich and Lim [12] propose a matrix completion approach to recover the partially observed matrix of Y, where exact recovery is guaranteed with high probability given an observation of $O(n \log^2 n)$ random samples. Their approach, however, is applicable only if score differences are observed. For the more practical case $P_{ij} = \text{sgn}(Y_{ij})$, recovery is much more challenging. For example, Jamieson and Nowak [15] show that if the ranking score of an item

[1] In general, P_{ij} can further contains some noise. We will consider such a scenario in detail in Section 4.

[2] For example, if $P_{ij} = \text{sgn}(Y_{ij})$, any $\pi = (\mathbf{s}+c)/c'$ with $c, c' > 0$ generates the identical P_{ij} as $\mathbf{s}$ does.

embeds the Euclidean distance to a reference point in a vector space, any algorithm needs $\Omega(n^2)$ comparisons to recover the exact ranking. This result is somehow pessimistic since it implies almost all pairwise comparisons are needed for exact recovery. However, non-trivial complexity can indeed be achieved if we consider approximate recovery instead. For example, Radinsky and Ailon [24] show that $O(n)$ comparisons on average suffice if ϵ-recovery is considered. Wauthier et al. [28] also provide two algorithms that achieve $O(n)$ sample complexity for ϵ-recovery and further show that the bound is tight. Compared to those works, we show that the sample complexity for ϵ-recovery can be further improved to be sublinear in n by carefully incorporating informative item features using the proposed model.

The use of features in our proposed model also shares similarities to recent works on incorporating side information in matrix completion [7, 8], but here the goal is ranking recovery instead.

3 RABF Model with Item Features

We propose a rank aggregation model where the ranking is learned by balancing between pairwise comparisons and feature information. In our framework, the ranking score of an item i is modeled as $\mu_i \mathbf{w}^T \mathbf{x}_i + r_i$, which can be interpreted as that the score is jointly estimated by two parts, one is contributed from features and one is from pure comparisons respectively, with balancing parameters μ_i controlling the importance of two information. We then solve for $\mathbf{w}$ and $\mathbf{r}$ by fitting observed pairwise comparisons as follows:

$$\min_{\mathbf{r}\in\mathbb{R}^n, \mathbf{w}\in\mathbb{R}^d} \sum_{(i,j)\in\mathcal{S}_I} \ell\big((r_j - r_i) + \mathbf{w}^T(\mu_j \mathbf{x}_j - \mu_i \mathbf{x}_i), P_{ij}\big) + \lambda(\|\mathbf{w}\|^2 + \|\mathbf{r}\|^2), \quad (1)$$

where $\|\cdot\|$ denotes the vector $\ell 2$ norm, and ℓ is some convex surrogate loss function. The underlying ranking is estimated by the ranking of the vector $D_\mu X \mathbf{w}^* + \mathbf{r}^*$ where $D_\mu = \mathrm{diag}([\mu_1, \dots \mu_n])$. Problem (1) is convex and can be efficiently solved by transforming it to an ERM objective. Details are left in Appendix A due to space limitations.

The choice of parameters μ_i is crucial in our model. Ideally, it can be set based on feature quality of the item i. However, in reality, feature quality of each item is usually unknown a priori. In this case, one can simply treat each feature equally by replacing all μ_i to be a single parameter μ, and the resulting formulation becomes equivalent to the form:

$$\min_{\mathbf{r},\mathbf{w}} \sum_{(i,j)\in\mathcal{S}_I} \ell\big((r_j - r_i) + \mathbf{w}^T(\mathbf{x}_j - \mathbf{x}_i), P_{ij}\big) + \lambda_w \|\mathbf{w}\|^2 + \lambda_r \|\mathbf{r}\|^2, \quad (2)$$

where $\lambda_w = \lambda/\mu^2$ and $\lambda_r = \lambda$. As a result, we only need to set (λ_w, λ_r) instead of μ_i, which can be determined using cross validation in practice. Since this scenario is more practical, we focus on problem (2) throughout the rest of the paper. We name the proposed model (2) "Rank Aggregation by Balancing Feature", or RABF.

Connections to methods without item features. If $\lambda_w = \infty$, $\mathbf{w}$ will be forced to zero, and RABF becomes an aggregation method without using any feature information. For example, the least-square based method proposed by [19] is to consider $P_{ij} = Y_{ij}$, ℓ to be squared loss, with $\lambda_r = 0$. The "ranking-SVM" model [28] is to consider $P_{ij} = \mathrm{sgn}(Y_{ij})$ and ℓ to be hinge loss. Probabilistic models like Bradley-Terry [4] and its variant [20] can also be described in this form, where the loss comes from minimizing negative log-likelihood objectives. These methods, however, disregard informative features even if they are provided.

Connections to learning to rank. If $\lambda_r = \infty$, RABF becomes a model where ranking scores are estimated by a *linear function of features*. For example, RABF objective is equivalent to the objectives of learning-to-rank models such as Rank-SVM [16], RankNet [6] and RankRLS [23] by setting ℓ to be hinge, logistic and squared loss respectively. One can think of this case as adapting learning-to-rank models Rank-SVM/RankNet to rank aggregation with features under a transductive setting, where training and testing items are the same. However, since ranking scores come from a linear function of features, these models can only guarantee recovery of the true ranking **only if s is in the column space of the feature matrix** X. If features are only partially correlated to $\mathbf{s}$, the ranking obtained from these models can be inaccurate. In real-world problems, features are usually noisy and far from linear, so adapting Rank-SVM often results in poor performance even if most comparisons are observed. This issue may be resolved by mapping features to a high dimensional space using kernels, but we will see that empirically RABF outperforms kernel Rank-SVM, suggesting that RABF is better than adapting learning-to-rank models.

In brief, by balancing between (λ_w, λ_r), RABF is more effective compared to these two classes of existing models, as it not only leverages feature information but is also more robust to noisy features.

4 Sample Complexity Analysis

In this section, we theoretically justify the usefulness of features in RABF model. We formally quantify the quality of features in Section 4.3 and show that given reasonably good features, RABF only requires sublinear number of clean comparisons to achieve ϵ-recovery

in Section 4.4. We then further generalize the results to noisy comparison case in Section 4.5. These results suggest that RABF is more efficient in learning accurate ranking by leveraging item features. All proofs of theorems and lemmas can be found in Appendix B.

4.1 Preliminaries

We consider the equivalent hard-constraint form of the original RABF formulation (2) as follows:

$$\min_{\theta} \sum_{(i,j)\in\mathcal{S}_I} \ell(\theta^T \bar{\mathbf{x}}_{ij}, P_{ij}), \text{s.t. } \|\mathbf{w}\| \le \mathcal{W}, \|\mathbf{r}\| \le \mathcal{R}, \quad (3)$$

where $\theta = [\mathbf{w}; \mathbf{r}]$, $\bar{\mathbf{x}}_{ij} = [\mathbf{x}_j - \mathbf{x}_i; \mathbf{e}_j - \mathbf{e}_i]$ where $\mathbf{e}_t$ denotes the unit vector on the t-th axis. Let the set of feasible θ defined as $\Theta = \{\theta = [\mathbf{w}; \mathbf{r}] \mid \|\mathbf{w}\| \le \mathcal{W}, \|\mathbf{r}\| \le \mathcal{R}\}$, and the set of functions $F_\Theta = \{f : \bar{\mathbf{x}} \to \theta^T \bar{\mathbf{x}} \mid \theta \in \Theta\}$. Let θ^* be the optimal solution of problem (3) and $\pi^* = X\mathbf{w}^* + \mathbf{r}^*$ be the output ranking scores. We also assume that the underlying scores are bounded, i.e. $\|\mathbf{s}\|_\infty \le \mathcal{T}$, and $d = O(1)$ as feature dimension does not grow as a function of n.

For any feasible $\theta \in \Theta$ and its corresponding ranking score $\pi = X\mathbf{w} + \mathbf{r}$, its Kendall's Tau distance to $\mathbf{s}$ can be expressed as the following expected risk quantity:

$$D_{k\tau}(\pi, \mathbf{s}) \equiv R(f) = \mathbb{E}_{(i,j)}\big[\mathbf{1}(\text{sgn}(f(\bar{\mathbf{x}}_{ij})) \ne \text{sgn}(Y_{ij}))\big].$$

Since optimizing the non-convex 0-1 loss is hard, the "ℓ-risk" defined on a convex surrogate ℓ is usually considered instead. For the case where comparisons are score differences, the ℓ-risk can be defined by:

$$R_\ell(f) = \mathbb{E}_{(i,j)}\big[\ell(f(\bar{\mathbf{x}}_{ij}), Y_{ij})\big],$$
$$\hat{R}_\ell(f) = \frac{1}{m}\textstyle\sum_{(i,j)\in\mathcal{S}_I} \ell(f(\bar{\mathbf{x}}_{ij}), Y_{ij}).$$

The term Y_{ij} is replaced by $\text{sgn}(Y_{ij})$ if comparisons are only the sign of score differences. Note that if observed comparisons are noiseless, our RABF model is to find θ^* parameterizing $f^* = \arg\min_{f\in F_\Theta} \hat{R}_\ell(f)$. For clarity, we will first focus on the noiseless comparison case from now on. We will generalize the results to the noisy comparison case in Section 4.5.

4.2 Sampling with Replacement

In our analysis, we consider that each $(i, j) \in \mathcal{S}_I$ is sampled from the distribution $\{1 \dots n\} \times \{1 \dots n\}$ uniformly i.i.d., i.e. randomly sample m comparisons with replacement. It may appear that the sampling with replacement model is unsuitable for analysis as entries in $\mathcal{S}_I$ could be repetitive. However, it turns out that we can bound the probability of RABF failing to attain ϵ-recovery when $\mathcal{S}_I$ is sampled from the collection of sets of size m by the sampling with replacement model:

Proposition 1 (Reduction of Sampling Models)**.** *The probability that RABF fails on the model where the set of observed comparisons is uniformly sampled from the collection of sets of size m is no greater than the probability that RABF fails on the model where m comparisons are sampled independently with replacement.*

Here, the failure event is defined as the output ranking fails to ϵ-recover the true ranking (i.e. $D_{k\tau}(\pi^*, \mathbf{s}) \ge \epsilon$). This proposition facilitates us to focus on the sampling with replacement model in the following discussion.

4.3 Measuring the Quality of Features

We now quantify the quality of features using Rademacher model complexity, a learning theoretic tool to measure the complexity of a function class. We show that "good features" will lead to a lower model complexity, and as a result, a good ranking can be guaranteed with fewer comparisons. We begin with the following lemma to bound the expected ℓ-risk:

Lemma 1 (Bound of excess risk [2])**.** *Let ℓ be a loss function with Lipschitz constant L_ℓ bounded by $\mathcal{B}$, and δ be a constant where $0 < \delta < 1$. Then, with probability at least $1 - \delta$, for all $f \in F_\Theta$ we have:*

$$R_\ell(f) \le \hat{R}_\ell(f) + 2L_\ell \mathbb{E}_{\mathcal{S}_I}\big[\mathfrak{R}(F_\Theta)\big] + \mathcal{B}\sqrt{\frac{\log 1/\delta}{2m}},$$

where $\mathfrak{R}(F_\Theta) := \mathbb{E}_\sigma\big[\sup_{f\in F_\Theta} \frac{1}{m}\sum \sigma_t f(\bar{\mathbf{x}}_{i_t j_t})\big]$ is the Rademacher complexity of the function class F_Θ.

Also, we introduce some definitions on features X used in the analysis. A feature matrix X is said to be γ-close if $\gamma \le \min_i \|\mathbf{x}_i\|/\mathcal{X}$, where $\mathcal{X} = \max_i \|\mathbf{x}_i\|$. Let $X = U\Sigma V^T$ be the reduced SVD of X, and $U_\mu(V_\mu)$ be the left (right) singular vectors with singular values at least $\mu\sigma_1$. With above notations, the following lemma further relates feature quality to model complexity:

Lemma 2 (Connection Between Model Complexity and Features)**.** *Let features X be γ-close and $\mu \in (0, 1]$ be a constant. By setting constraints in (3) to be:*

$$\mathcal{W} = \frac{\sqrt{d}}{(\mu\gamma\mathcal{X}\sqrt{n})}\|\mathbf{d}\| \quad \text{and} \quad \mathcal{R} = \|\mathbf{r}\|, \quad (4)$$

where $\mathbf{d} = U_\mu U_\mu^T \mathbf{s}$ and $\mathbf{r} = \mathbf{s} - \mathbf{d}$, the expected Rademacher complexity is bounded by:

$$\mathbb{E}_{\mathcal{S}_I}\big[\mathfrak{R}(F_\Theta)\big] \le \left(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}}\|\mathbf{d}\| + \|\mathbf{r}\|\right)\sqrt{\frac{2}{m}}. \quad (5)$$

As an explanation, $\mathbf{d}$ is the projection of $\mathbf{s}$ on the important part of feature space U_μ, and $\mathbf{r}$ is the residual that is not covered by U_μ. Since $d = O(1)$, as n goes large, the second term in (5) dominates the model complexity to be $O(\|\mathbf{r}\|/\sqrt{m})$. As we will see shortly, a smaller model complexity will lead to better guarantee, therefore a feature set can be regarded as good if the resulting $\|\mathbf{r}\|$ is small. Such a measurement matches the intuition of good features because smaller $\|\mathbf{r}\|$ can be accomplished if a large portion of $\mathbf{s}$ lies on U_μ, i.e. much of the underlying ranking information is contained in the informative part of feature space.

4.4 Guarantees for Noiseless Comparisons

With the above lemmas, we can now derive the theorems which guarantee the Kendall's Tau distance between the ranking from RABF model and the true ranking for the noiseless comparison case.

Theorem 1 (Guarantee for $P_{ij} = Y_{ij}$)**.** *Let $\delta \in (0,1)$ be a constant. Suppose the following assumptions hold:*

a. We observe m clean pairwise comparisons $P_{ij} = Y_{ij}$ under the sampling with replacement model.
b. Feature matrix X is γ-close with bounded $\mathcal{X}$.
c. The convex surrogate loss function ℓ is bounded for each P_{ij}, with $\ell(x,x) = 0$.

Then by setting $\mathcal{W}$ and $\mathcal{R}$ as (4), with probability at least $1-\delta$, the optimal π^* from problem (3) satisfies:

$$D_{k\tau}(\pi^*, \mathbf{s}) \le O\left((\sqrt{d} + \|\mathbf{r}\|)\sqrt{\frac{1}{m}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{m}}\right).$$

Theorem 2 (Guarantee for $P_{ij} = \mathrm{sgn}(Y_{ij})$)**.** *Let $\delta \in (0,1)$ be a constant. Suppose the following assumptions hold:*

a. We observe m clean comparisons $P_{ij} = sgn(Y_{ij})$ under the sampling with replacement model.
b. Feature matrix X is γ-close with bounded $\mathcal{X}$.
c. The convex surrogate loss ℓ is bounded for each P_{ij}.

Then by setting $\mathcal{W}$ and $\mathcal{R}$ as (4), with probability at least $1-\delta$, the optimal π^* from problem (3) satisfies:

$$D_{k\tau}(\pi^*, \mathbf{s}) \le O\big(\hat{R}_\ell(f^*) - R_\ell^*\big) + O\left((\sqrt{d} + \|\mathbf{r}\|)\sqrt{\frac{1}{m}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{m}}\right),$$

where $R_\ell^* = \inf_f R_\ell(f)$. The sample complexity of RABF can thus be derived as follows.

Corollary 1 (Sample Complexity for Noiseless Comparisons)**.** *Given any $\epsilon > 0$ and suppose assumptions a-c in Theorem 1 hold. Then with sufficiently large n, $O(\|\mathbf{r}\|^2/\epsilon^2)$ comparisons are sufficient for RABF to guarantee an ϵ-accurate ranking.*

The same sample complexity can be derived for the case $P_{ij} = \mathrm{sgn}(Y_{ij})$ provided that $\hat{R}_\ell(f^*) - R_\ell^* = O(\epsilon)$. Corollary 1 suggests that for a better feature set (i.e. smaller $\|\mathbf{r}\|$), fewer comparisons are required to achieve an ϵ-accurate ranking. In particular, if feature quality is sufficiently good such that $\|\mathbf{r}\|^2 = o(n)$, the sample complexity becomes only *sublinear* to the number of items. To show the scenario is realistic, we provide two concrete instances for such a scenario as follows.

Theorem 3 (Example Scenarios for Sublinear Sample Complexity)**.** *Let $X^\star \in \mathbb{R}^{n\times d}$ be a feature set where $\mathbf{s} \in col(X^\star)$ and $d = O(1)$. Suppose now $O(\log n)$ items are corrupted in either of the following scenarios:*

1. Each corrupted item i has perturbed feature $\mathbf{x}_i^\star + \Delta\mathbf{x}_i$, where $\Delta\mathbf{x}_i$ are Subgaussian variables, $\|\Delta\mathbf{x}_i\|_\infty \le \xi$ with a constant ξ.
2. Each corrupted item i has shuffled feature $\mathbf{x}_j^\star$ from another corrupted item j.

Then, given such a corrupted feature matrix X, $O(\log n)$ comparisons are sufficient for RABF to guarantee an ϵ-accurate ranking.

Method	good features ($\|\mathbf{r}\|^2 = o(n)$)	bad features ($\|\mathbf{r}\|^2 = O(n)$)
RABF	$o(n)$	$O(n)$
Comparison only	$O(n)$ (and also $\Omega(n)$)	
Rank-SVM	Cannot recover unless $\mathbf{s} \in \mathrm{col}(X)$	

Table 1: Sample complexity of various methods. We see that RABF is the only one that not only always recovers the ranking with enough samples but also achieve sublinear complexity provided good features.

Comparisons with Other Models. We highlight the strength of our result by comparing to other methods. First, for methods without features, it has been shown that *any algorithm based only on comparisons* requires at least $\Omega(n)$ comparisons for ϵ-recovery [11, 28]. Compared to them, our RABF model has sample complexity at most $O(n)$ since $\|\mathbf{r}\| \le \|\mathbf{s}\| = O(\sqrt{n})$, suggesting that RABF is at least as good as any method based purely on comparisons. Note that it is reasonable to meet the $\Omega(n)$ lower bound even if given features, as in an extreme case where X is a random matrix, the given information is same as the case where only comparisons are given. However, in practice, features are expected to be informative, and our results show that we can asymptotically improve sample complexity by leveraging informative features using RABF model.

On the other hand, methods adapted from learning-to-rank usually cannot even guarantee recoverability if $\mathbf{s}$ does not perfectly lie in the feature space (see discussions in Section 3). Thus, given a general feature set, the true ranking $\mathbf{s}$ may be infeasible, in which case the recovery may not be attained even if all n^2 clean comparisons are observed. Compared to them, true ranking is always feasible in RABF given *any* feature set and an ϵ-accurate ranking is guaranteed with $O(n)$ comparisons. The above comparisons are summarized in Table 1 and will also be empirically supported in Section 6.1.

4.5 Guarantees for Noisy Comparisons

So far, our analysis focuses on the case where comparisons are clean. We now further show that RABF can also achieve efficient learning even if comparisons are noisy. We consider a standard "flip-sign model" [5, 28] where each observed comparison may be corrupted by flip-sign noise as $P_{ij} = -\mathrm{sgn}(Y_{ij})$ with probability ρ_c or remain clean as $P_{ij} = \mathrm{sgn}(Y_{ij})$ otherwise, where $\rho_c \in [0, 0.5)$ is the comparison noise level. Then, the following theorem shows that we can still obtain an

accurate ranking efficiently from the RABF model:

Theorem 4 (Sample Complexity for Noisy Comparisons)**.** *Let X be a γ-close feature set, and each P_{ij} is now observed under the flip-sign model with $\rho_c \in [0, 0.5)$. Then by solving RABF model with squared loss, $O(\|\mathbf{r}\|^2/((1-2\rho_c)^2\epsilon^2))$ comparisons suffice to guarantee an ϵ-accurate ranking.*

The theorem shows that in noisy comparison case, RABF can achieve ϵ-recovery with the same order of sample complexity (w.r.t. n) as in noiseless case, and the extra price to pay is a $1/(1-2\rho_c)^2$ factor. Thus, given a fixed noise level ρ_c, sublinear sample complexity can still be achieved provided informative enough features (i.e. $\|\mathbf{r}\|^2 = o(n)$). It suggests that by leveraging features, RABF model can also learn the underlying ranking efficiently in noisy comparison case.

5 Extensions

In this section, we turn our attention to another two related problems—online rank aggregation and rank prediction of new items. Though settings and goals of these problems seem to be different from rank aggregation, we show that by extending the RABF model, we can also approach these problems more effectively by leveraging item features.

5.1 Online Rank Aggregation with Features

The online rank aggregation problem is widely considered in modern rating systems, e.g. Glicko [13] and TrueSkill [14]. The problem can be stated as follows. Given n items and a feature matrix X, the learner can only observe a single comparison P_{ij} at each time stamp $t : 1 \le t \le T$ and is asked to output an estimate ranking of $\mathbf{s}$ at time T. The problem is at least as hard as (batch) rank aggregation, since by a reduction, it can be shown that $\Omega(n)$ is still the lower bound for any method based only on online comparisons. However, it is not clear if such a $\Omega(n)$ barrier can be also tackled by making use of features as in the batch setting.

We propose an online extension of RABF to achieve sublinear sample complexity for online rank aggregation. The core concept is to solve the RABF model by performing a stochastic gradient update on $(\mathbf{w}, \mathbf{r})$ for each P_{ij} observed at time t. By doing so, we can further prove that online-RABF only requires $O(\|\mathbf{r}\|^2/\epsilon^2)$ online comparisons to output an ϵ-accurate ranking, which again implies that given good features such that $\|\mathbf{r}\|^2 = o(n)$, a sublinear number of samples suffices. See Appendix C for detailed algorithm and theoretical analysis and Section 6.2 for experimental results.

5.2 Rank Prediction of New Items

We now consider another different task—rank prediction of new items. Suppose in the training phase we are given the feature matrix X of n items and a set of comparisons between these items, and in the testing phase, we are given another new item where only its feature $\mathbf{x}_{\text{new}}$ is available. The task is to predict the rank of that new item among the seen items given in the training phase.

	$P_{ij} = Y_{ij}$	$P_{ij} = \text{sgn}(Y_{ij})$
Comparisons only	LS, SVP	MLE, BRE
Features and comparisons	RABF-SQ* MR	RABF-LOG*, RankNet RankSVM, Rank-KSVM

Table 2: Setting of each rank aggregation method. Starred methods are instances of our RABF model (2).

As an advantage of leveraging feature information, our model could be extended to this problem by first deriving a ranking of training items using RABF model and predicting the ranking score of the new item simply by $\mathbf{w}^T\mathbf{x}_{\text{new}}$. The rank of $\mathbf{w}^T\mathbf{x}_{\text{new}}$ in sorted $X\mathbf{w}+\mathbf{r}$ will be the predicted rank of the new item. This can be viewed as treating the new item having 0 comparison score (so $r_{\text{new}} = 0$) as a priori and only using its feature score to decide its ranking score.

Readers may notice that feature-based models like Rank-SVM could also be adapted to this problem in a similar way, and it is natural to ask what is the advantage of adapting RABF rather than other feature-based models. Indeed, it may appear that RABF is no better than other feature-based methods for rank prediction since there is no comparison information for new items that RABF can leverage. However, interestingly, we found that RABF outperforms Rank-SVM if features are noisy (see Section 6.2). To explain the result, note that the rank of new item is decided by the rank of its predicted score among scores of training items, so how accurate the model recovers the ranking of training items also influences the performance. Thus, since RABF better recovers the ranking of training items when features are noisy, it will also rank new items more accurately in such cases.

6 Experiments

We first conduct experiments on rank aggregation on both synthetic and real datasets in Section 6.1. We show that proposed RABF model is effective in two aspects: 1) it is more robust to noisy information, and 2) it needs fewer comparisons to obtain a good ranking. In Section 6.2, we conduct experiments on online rank aggregation and rank prediction for new items, showing that the extensions of RABF also improve performance on these problems by leveraging item features.

6.1 Experiments on rank aggregation

Experiment setup. We select two representatives of the RABF model. For $P_{ij} = Y_{ij}$, we consider RABF-

SQ as $\ell(t, y) = (t - y)^2$ to be squared loss, and for $P_{ij} = \text{sgn}(Y_{ij})$, we consider RABF-LOG as $\ell(t, y) = \log(1+e^{-ty})$ to be logistic loss. We compare our methods with other methods including: aggregation with least squares (LS) [19] and nuclear norm minimization (SVP) [12], a variant of Bradley-Terry model (MLE) [20], Balanced Rank Estimation (BRE) [28], method of manifold regularization (MR) [9], and Rank-SVM (Rank-SVM) [16], kernel Rank-SVM (Rank-KSVM), RankNet (RankNet) [6] adapted to rank aggregation. Settings of each method are summarized in Table 2. Due to limited space, we only display the plots for $P_{ij} = \text{sgn}(Y_{ij})$. Results for $P_{ij} = Y_{ij}$ are similar and can be seen in Appendix E. Parameters of each model are selected via cross validation with parameter set $\{10^k\}_{k=-2}^{3}$ and all results are averaged with 10 trials.

Synthetic datasets. Our synthetic datasets were created as follows. We generated a true ranking score vector $\mathbf{s} \in \mathbb{R}^n$ and uniformly sampled m clean comparisons P_{ij} from Y. We also constructed a feature matrix $X^\star \in \mathbb{R}^{n\times 50}$ whose top-30 singular vectors span $\mathbf{s}$. We then added noise to comparisons and features as follows. For each observed P_{ij}, we flipped its sign as a noisy comparison with probability ρ_c. For feature matrix $X^\star$, we select each row to be a corrupted item with probability ρ_f, and all selected rows were randomly shuffled to form a noisy feature set X.

First, we compare all methods under various feature quality. We fix $n = 1000$, $m = 5n$, $\rho_c = 0.1$ and vary ρ_f from 0 to 1. We apply each method to estimate a ranking and plot its Kendall's Tau to $\mathbf{s}$ in Figure 1a. In Figure 1a, we can see that when ρ_f is small, Rank-SVM is more effective than methods without features since X contains much information of $\mathbf{s}$. However, as ρ_f increases, performance of Rank-SVM quickly drops since features become misleading, while methods based only on comparisons will not be influenced. Rank-KSVM uses nonlinear Gaussian kernel to avoid fitting a linear combination of poor features, but when feature quality is good, it works worse than Rank-SVM because of overfitting. On the other hand, both RABF-LOG and RABF-SQ are the all-time winners under different quality of features. They make use of good features when ρ_f is small and are also robust to bad features by learning a ranking mainly from comparisons when ρ_f is large. The results show that RABF combines advantages of two classes of methods.

We next compare all methods under different comparison quality. We fix $n = 1000$, $m = 5n$, $\rho_f = 0.25$ and vary ρ_c from 0 to 0.5. The results of different methods are shown in Figure 1b. We observe that RABF-LOG and RABF-SQ also perform the best among all methods under each noise level. This shows that RABF is also less sensitive to noisy comparisons.

Finally, we also conduct experiments to show that given good features, RABF requires much fewer (i.e. sublinear) number of samples to achieve ϵ-recovery as an empirically support of theoretical results. The results are left in Appendix D due to limited space.

Real-world datasets. We now show the effectiveness of RABF on real-world datasets where features are typically noisy. We first consider the Forbes ranking of the world's biggest public companies, in which experts ranked the top-2000 global companies in 2014 based on mixed performance factors. For each company, we also collected its features $\mathbf{x}_i \in \mathbb{R}^{152}$ correlated to its rank, such as its country, industry, and financial indices. To conduct the experiment, we randomly sampled m clean comparisons from underlying true ranking with various m, and applied each method to estimate the ranking list given feature set X and m comparisons. The results are shown in Figure 1c. We can observe that RABF achieves the best Kendall's Tau given the same number of comparisons.

Finally, we consider an application of ranking sports teams. We consider an NBA matchup dataset [3], where 30 teams had $m = 1144$ matchups in 2008-2009 regular season. For each team, a 13-dimensional feature vector is also collected from the team's performance last season, such as the total points, assists and rebounds the team made. Apparently, these features should be only partially correlated to team ranking since each team may make some upgrades during the off-season. The goal is to take both matchup results and team features to produce a good ranking of teams.

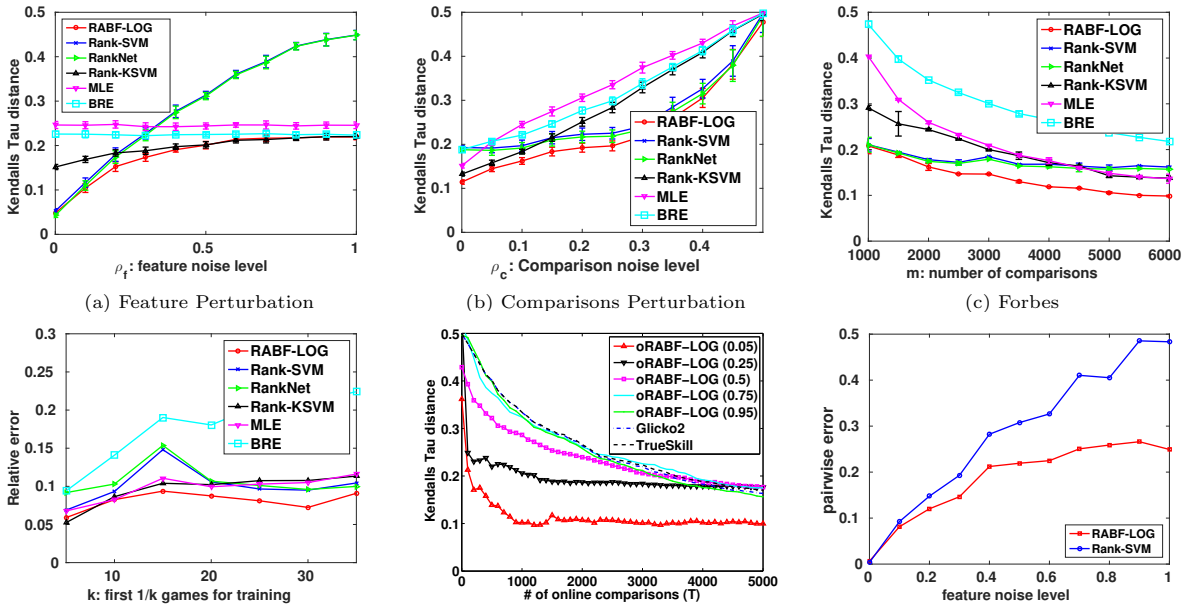
The experiment is conducted as follows. We take the first m/k games in the season as training comparisons to derive a team ranking π and evaluate the ranking by using π to predict winning teams on the remaining $m(k-1)/k$ games. For each remaining game, we simply predict the team with the higher rank wins, and a good ranking should result in higher accuracy (or lower error rate). Note that the best ranking π_{opt} for prediction is the ranking based on teams' winning percentage in the remaining games. Thus, we evaluate the ranking π using the following relative error criterion:

$$\text{Rel-err}(\pi) = \text{Acc achieved by } \pi_{opt} - \text{Acc achieved by } \pi.$$

The results are shown in Figure 1d. We see that RABF generally achieves lower error compared to others. In particular, RABF-LOG performs the best when the training games are few (i.e. larger k), suggesting that a good ranking can be derived with fewer comparisons.

6.2 Experiments for online rank aggregation and rank prediction of new items

We now show that the extensions of RABF model also improve performance on online rank aggregation and



(a) Feature Perturbation (b) Comparisons Perturbation (c) Forbes

(d) NBA (e) Online rank aggregation (f) Rank prediction of new items

Figure 1: Performance of proposed RABF models on rank aggregation (1a∼1d), online rank aggregation (1e) and rank prediction of new items (1f). These experiments show the effectiveness of the RABF model for various ranking problems by leveraging item features.

rank prediction of new items. Details of the extensions of RABF are stated in Section 5.

Online rank aggregation. We first compare the online extension of RABF-LOG (oRABF-LOG) with two state-of-the-art online rank aggregation methods, Glicko2 [13] and TrueSkill [14]. We consider synthetic datasets where $n = 1000$, $\rho_c = 0.1$, with ρ_f from 0.05 to 0.95. [3] We online update all methods given a comparison at each time and plot their performance versus number of online comparisons T in Figure 1e. We first observe that the performance of Glicko2 and TrueSkill are almost identical as Kendall's Tau drops linearly when T increases. On the other hand, oRABF-LOG performs at least as good as Glicko2 and TrueSkill even when features are noisy ($\rho_f \geq 0.75$), and the performance is significantly improved when features become informative. In particular, with reasonably good features ($\rho_f \leq 0.25$), oRABF-LOG is able to output a ranking with bounded $D_{k\tau}$ after only sublinear number of online comparisons are seen. This experiment shows the effectiveness and sublinear sample complexity of online RABF model described in Section 5.1.

Rank prediction of new items. Finally, we show that the RABF model is also useful for rank prediction of new items as stated in Section 5.2. We consider synthetic datasets where $n = 1000$, $m = 10n$, $\rho_c = 0$ and ρ_f from 0 to 1. For each ρ_f, we randomly select an item i (with feature $\mathbf{x}_i = \mathbf{x}_{\text{new}}$) as the testing new item, and the other 999 items as given items for training. We train both RABF-LOG and Rank-SVM on seen items, use the obtained models to predict score of the testing item (which is $\mathbf{w}^T\mathbf{x}_{\text{new}}$) and evaluate the prediction using the following pairwise error metric:

$$1/(n-1)\textstyle\sum_{j\neq i}\mathbf{1}\big[\text{sgn}(s_j - s_i) \neq \text{sgn}(\mathbf{w}^T\mathbf{x}_j + r_j - \mathbf{w}^T\mathbf{x}_{\text{new}})\big].$$

We repeat the procedure 100 times for each ρ_f and plot the average pairwise error in Figure 1f. We observe that while RABF-LOG and Rank-SVM perform similarly with good features, RABF-LOG gives better prediction when features become noisy. As explained in Section 5.2, the performance will be influenced not only by the prediction of the new item but also by the recovered ranking of seen items. Thus, RABF achieves lower error rate when features are noisy because it can estimate the ranking of seen items more accurately in such a case (also see Figure 1a for support).

7 Conclusions

We propose a new RABF model to exploit item features for rank aggregation problem. The effect of features is analyzed and an improved sample complexity of the model is derived with the aid of features. The model is also shown to be advantageous on online rank aggregation and predicting rank of new items. The effectiveness of the proposed model is also empirically supported by several experiments. These results show the usefulness of item features under the context of ranking in both theory and practice.

[3] Note that Glicko2 and TrueSkill will not be influenced by ρ_f as they only take pairwise comparisons into account.

Acknowledgement

This research was supported by NSF grants CCF-1320746, IIS-1546452 and CCF-1564000.

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138 – 156, 2006.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[3] A. Barzilai. Basketballvalue.com. http://basketballvalue.com/index.php, 2012.

[4] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pages 324–345, 1952.

[5] M. Braverman and E. Mossel. Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 268–276, 2008.

[6] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

[7] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon. Matrix completion with noisy side information. In *NIPS*, pages 3447–3455, 2015.

[8] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon. Robust principal component analysis with side information. In *ICML*, pages 2291–2299, 2016.

[9] F. Diaz. Regularizing query-based retrieval scores. *Information Retrieval*, 10, 2007.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[11] J. Giesen, E. Schuberth, and M. Stojaković. Approximate sorting. *Fundamenta Informaticae*, 90(1-2):67–72, 2009.

[12] D. F. Gleich and L.-H. Lim. Rank aggregation via nuclear norm minimization. In *KDD*, 2011.

[13] M. E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, 1999.

[14] R. Herbrich, T. Minka, and T. Graepel. Trueskill(tm): A bayesian skill rating system. In *NIPS*, pages 569–576. MIT Press, January 2007.

[15] K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. In *NIPS*, pages 2240–2248, 2011.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

[17] A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, 2009.

[18] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, pages 793 – 800, 2008.

[19] K. Masse. Statistical models applied to the rating of sports teams. *Master Thesis, Bluefield College*, 1997.

[20] A. K. Massimino and M. A. Davenport. One-bit matrix completion for pairwise comparison matrices. In *Proceedings of Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2013.

[21] N. Natarajan, A. Tewari, I. S. Dhillon, and P. Ravikumar. Learning with noisy labels. In *Neural Information Processing Systems (NIPS)*, pages 1196–1204, dec 2013.

[22] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *NIPS*, pages 2483–2491, 2012.

[23] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Boberg, and T. Salakoski. Learning to rank with pairwise regularized least-squares. In *SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, 2007.

[24] K. Radinsky and N. Ailon. Ranking from pairs and triplets: Information quality, evaluation methods and query complexity. In *WSDM*, 2011.

[25] B. Recht. A simpler approach to matrix completion. *JMLR*, 12:3413–3430, 2011.

[26] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math*, pages 1707–1739, 2009.

[27] D. Schäfer and E. Hüllermeier. Dyad ranking using a bilinear plackett-luce model. In *ECML-PKDD*, 2015.

[28] F. Wauthier, M. Jordan, and N. Jojic. Efficient ranking from pairwise comparisons. In *ICML*, pages 109–117, 2013.

Appendix A: Solving the Proposed Model

To solve the proposed problem (1) efficiently, we can rewrite the problem as follows. For each comparison P_{ij}, consider a corresponding vector $\tilde{\mathbf{x}}_{ij} \in \mathbb{R}^{n+d}$ defined by:

$$\tilde{\mathbf{x}}_{ij} = [(\mu_j \mathbf{x}_j - \mu_i \mathbf{x}_i); \mathbf{e}_j - \mathbf{e}_i],$$

where $\mathbf{e}_j$ ($\mathbf{e}_i$) is an n dimensional unit vector with only the j-th (i-th) position is one. Then the problem can be written compactly as:

$$\min_{\theta \in \mathbb{R}^{n+d}} \sum_{(i,j)\in\mathcal{S}_I} \ell(\theta^T \tilde{\mathbf{x}}_{ij}, P_{ij}) + \lambda \|\theta\|^2, \tag{6}$$

where $\theta = [\mathbf{w}; \mathbf{r}]$ is the parameter set we want to optimize. The problem now is in a standard empirical risk minimization (ERM) form, which can be solved efficiently using publicly available solvers (e.g. LIBLINEAR package [10] used in our experiments).

Appendix B: Proofs

Proof of Proposition 1

Proof (of Proposition 1). The argument and the proof of the proposition is quite standard in recovery literatures, e.g. [25]. We repeat the high-level idea of the proposition for completeness. Let Ω be the set of m comparisons, each of which is sampled independently from $\{1 \dots n\} \times \{1 \dots n\}$. Let Ω_t be the set of entries with cardinality t, uniformly sampled from the collection of sets of t unique comparisons. Let $\mathcal{F}(\Omega)$ and $\mathcal{F}(\Omega_m)$ be the event that the problem (2) fails to output an ϵ-accurate ranking given the comparison set Ω and Ω_m respectively. Then we have:

$$\begin{aligned}
\Pr(\mathcal{F}(\Omega)) &= \sum_{t=1}^{m} \Pr(\mathcal{F}(\Omega) \mid |\Omega| = t)) \Pr(|\Omega| = t) \\
&= \sum_{t=1}^{m} \Pr(\mathcal{F}(\Omega_t)) \Pr(|\Omega| = t) \\
&\geq \Pr(\mathcal{F}(\Omega_m)) \sum_{t=1}^{m} \Pr(|\Omega| = t) \\
&= \Pr(\mathcal{F}(\Omega_m)),
\end{aligned}$$

where the third inequality is because the failure probability will not increase as number of samples increases in Ω_t, i.e.

$$\text{for all } t_1 \leq t_2, \quad \Pr(\mathcal{F}(\Omega_{t_1})) \geq \Pr(\mathcal{F}(\Omega_{t_2})).$$

□

Proof of Lemma 2

First, we need the following preliminary lemma to bound the Rademacher complexity of class of linear functions.

Lemma 3 (Complexity Bound on Linear Function Class [18])**.** *Let F_W be a class of linear functions $\{\mathbf{x} \to \mathbf{w}^T\mathbf{x} \mid \|\mathbf{w}\| \leq \hat{\mathcal{W}}\}$, and each $\mathbf{x}$ is bounded by $\hat{\mathcal{X}}$. Then the Rademacher complexity of F_W is bounded by:*

$$\mathfrak{R}(F_W) \leq \hat{\mathcal{X}}\hat{\mathcal{W}}\sqrt{\frac{1}{m}}.$$

With this lemma, now we can present the proof of Lemma 2.

Proof (of Lemma 2). By the definition of the Rademacher complexity of function class F_Θ, we can rewrite $\mathfrak{R}(F_\Theta)$ as follows:

$$\begin{aligned}
\mathfrak{R}(F_\Theta) &= \mathbb{E}_\sigma\big[\sup_{\theta\in\Theta} \frac{1}{m}\sum_{t=1}^{m} \sigma_t \theta^T \bar{\mathbf{x}}_{i_t j_t}\big] \\
&= \mathbb{E}_\sigma\big[\sup_{\|\mathbf{w}\|\leq\mathcal{W}} \frac{1}{m}\sum_{t=1}^{m} \sigma_t \mathbf{w}^T(\mathbf{x}_{j_t} - \mathbf{x}_{i_t})\big] \\
&\quad + \mathbb{E}_\sigma\big[\sup_{\|\mathbf{r}\|\leq\mathcal{R}} \frac{1}{m}\sum_{t=1}^{m} \sigma_t \mathbf{r}^T(\mathbf{e}_{j_t} - \mathbf{e}_{i_t})\big],
\end{aligned} \tag{7}$$

which contains the complexity of two linear function classes. Since for any (i_t, j_t), $\|\mathbf{x}_{j_t} - \mathbf{x}_{i_t}\| \leq 2\mathcal{X}$ and $\|\mathbf{e}_{j_t} - \mathbf{e}_{i_t}\| \leq \sqrt{2}$, by applying Lemma 3 to each term in (7), we can upper bound the complexity of $\mathbb{E}_{\mathcal{S}_I}\big[\mathfrak{R}(F_\Theta)\big]$ by:

$$\mathbb{E}_{\mathcal{S}_I}\big[\mathfrak{R}(F_\Theta)\big] \leq (\sqrt{2}\mathcal{X}\mathcal{W} + \mathcal{R})\sqrt{\frac{2}{m}}. \tag{8}$$

We now further construct an appropriate setting of $\mathcal{W}$ and $\mathcal{R}$ as follows. Let $\mathbf{d} = U_\mu U_\mu^T \mathbf{s}$ be the projection of $\mathbf{s}$ on the subspace given by the orthogonal matrix U_μ. Consider $\hat{\mathbf{w}} = \arg\min_{\mathbf{d}=X\mathbf{w}} \|\mathbf{w}\|^2$. The minimum norm solution $\hat{\mathbf{w}}$ is given by the SVD of X, i.e.,

$$\hat{\mathbf{w}} = X^\dagger \mathbf{d} = V\Sigma^\dagger U^T \mathbf{d} = V_\mu \Sigma_\mu^\dagger U_\mu^T \mathbf{d}, \tag{9}$$

where $\Sigma_\mu^\dagger = \operatorname{diag}(1/\sigma_1, 1/\sigma_2 \cdots 1/\sigma_{\tilde{d}})$. Combining with the definition of U_μ, we have

$$\|\hat{\mathbf{w}}\| \leq \frac{1}{\mu\sigma_1}\|\mathbf{d}\|,$$

in which σ_1 can be further bounded as follows:

$$\sigma_1^2 = \|X\|_2^2 \geq \frac{\|X\|_F^2}{d} \geq \frac{n\gamma^2\mathcal{X}^2}{d}.$$

Therefore, we can upper bound $\|\hat{\mathbf{w}}\|$ by:

$$\|\hat{\mathbf{w}}\| \leq \frac{\sqrt{d}}{\mu\gamma\mathcal{X}\sqrt{n}}\|\mathbf{d}\|.$$

The lemma is therefore proved by plugging $\mathcal{W} = \|\hat{\mathbf{w}}\|$ and $\mathcal{R} = \|\mathbf{s} - \mathbf{d}\|$ into (8). □

Proof of Theorem 1

The following preliminary lemma is required in the proof to link ℓ-risk to excess risk of 0-1 loss:

Lemma 4 (Consistency of Excess Risk [1]). *Let ℓ be a convex surrogate loss function. Then there exists a strictly increasing function Ψ, $\Psi(0) = 0$, such that for all measurable f:*

$$R(f) - R^* \leq \Psi(R_\ell(f) - R_\ell^*),$$

where $R^* = \inf_f R(f)$ and $R_\ell^* = \inf_f R_\ell(f)$.

Now we can prove the Theorem as follows.

Proof (of Theorem 1). Consider the problem (3) with $P_{ij} = Y_{ij}$ where $\mathcal{W}$ and $\mathcal{R}$ are set to be (4). Let $f^*(\bar{\mathbf{x}}) = \theta^{*T}\bar{\mathbf{x}}$ where $\theta^* \in \Theta$ is the optimal solution of (3). From the construction in the proof of Lemma 2, $\hat{\theta} = [\hat{\mathbf{w}}, \mathbf{r}]$ is (one of) an optimal solution θ^* since $\hat{\theta}$ satisfies $\ell(f(\bar{\mathbf{x}}_{ij}), P_{ij}) = \ell(s_j - s_i, P_{ij}) = \ell(Y_{ij}, Y_{ij}) = 0$ for any (i, j). This suggests that $\hat{R}_\ell(f^*) = 0$ and apparently $R^* = R_\ell^* = 0$. Therefore, in this context, Lemma 4 becomes:

$$R(f^*) \leq \Psi(R_\ell(f^*)).$$

On the other hand, since $\ell(f^*(\bar{\mathbf{x}}_{ij}), P_{ij}) \leq \mathcal{B}$, the expected ℓ-risk of f^* can be bounded by Lemma 1 as:

$$R_\ell(f^*) \leq 2L_\ell \left(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}} \|\mathbf{d}\| + \|\mathbf{r}\| \right) \sqrt{\frac{2}{m}} + \mathcal{B}\sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{10}$$

Finally, let $L_\Psi = \Psi(\mathcal{B})$ be the (bounded) Lipschitz constant for Ψ. Then, by putting above two equations together, we can derive the Theorem as:

$$\begin{aligned}
&D_{k\tau}(\pi^*, \mathbf{s}) \\
=&R(f^*) \\
\leq&\Psi(R_\ell(f^*)) \\
\leq&L_\Psi \left(2L_\ell \big(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}} \|\mathbf{d}\| + \|\mathbf{r}\| \big) \sqrt{\frac{2}{m}} + \mathcal{B}\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \right) \\
=&O\left((\sqrt{d} + \|\mathbf{r}\|) \sqrt{\frac{1}{m}} \right) + O\left(\sqrt{\frac{\log 1/\delta}{m}} \right),
\end{aligned}$$

by the fact that $\|\mathbf{d}\| \leq \|\mathbf{s}\| = O(\sqrt{n})$. □

Proof of Theorem 2

Proof (of Theorem 2). Again, consider the problem (3) where $\mathcal{W}$ and $\mathcal{R}$ are set as (4), except that now $P_{ij} = \text{sgn}(Y_{ij})$ is observed instead. The instance $\hat{\theta} = [\hat{\mathbf{w}}; \mathbf{r}]$ (defined in the proof of Lemma 2) is still in the feasible solution set Θ, and thus its corresponding function $f_{\hat{\theta}}$ is also feasible in F_Θ. However, unlike the case $P_{ij} = Y_{ij}$ in Theorem 1, $\hat{\theta}$ is not necessarily the optimal solution of problem (3) for the case $P_{ij} = \text{sgn}(Y_{ij})$. Indeed, although $\hat{\theta}$ satisfies $X\hat{\mathbf{w}} + \mathbf{r} = \mathbf{s}$, it may exist another $\theta^* \in \Theta$ such that $\hat{R}_\ell(f^*) \leq \hat{R}_\ell(f_{\hat{\theta}})$. Nevertheless, $\hat{\theta}$ still provides an instance to show $R^* = 0$. Thus, by applying Lemma 4 in this case, we have:

$$R(f^*) \leq \Psi(R_\ell(f^*) - R_\ell^*). \tag{11}$$

Using Lemma 1, the quantity $R_\ell(f^*) - R_\ell^*$ can be further bounded by:

$$\begin{aligned}
&R_\ell(f^*) - R_\ell^* \\
\leq&\hat{R}_\ell(f^*) - R_\ell^* \\
&+ 2L_\ell \left(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}} \|\mathbf{d}\| + \|\mathbf{r}\| \right) \sqrt{\frac{2}{m}} + \mathcal{B}\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.
\end{aligned}$$

Note that here $\hat{R}_\ell(f^*) - R_\ell^*$ can amount a positive quantity, as f^* may still make the term $\ell(f^*(\bar{\mathbf{x}}_{ij}), P_{ij})$ non-zero in empirical ℓ-risk. However, such a quantity is expected to be extreme small since $\hat{R}_\ell(f^*) \leq \hat{R}_\ell(f_{\hat{\theta}})$, where $\hat{R}_\ell(f_{\hat{\theta}})$ is the ℓ-risk of the true ranking.

Finally, let L_Ψ be the Lipschitz constant for Ψ bounded by $\Psi(\mathcal{B})$. Then the Theorem follows by putting the above two equations together as:

$$\begin{aligned}
&D_{k\tau}(\pi^*, \mathbf{s}) \\
=&R(f^*) \\
\leq&\Psi(R_\ell(f^*) - R_\ell^*) \\
\leq&L_\Psi \Bigg(\hat{R}_\ell(f^*) - R_\ell^* \\
&+ 2L_\ell \big(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}} \|\mathbf{d}\| + \|\mathbf{r}\| \big) \sqrt{\frac{2}{m}} + \mathcal{B}\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \Bigg) \\
=&O\big(\hat{R}_\ell(f^*) - R_\ell^* \big) \\
&+ O\left((\sqrt{d} + \|\mathbf{r}\|) \sqrt{\frac{1}{m}} \right) + O\left(\sqrt{\frac{\log 1/\delta}{m}} \right).
\end{aligned}$$

□

Proof of Theorem 3

Proof (of Theorem 3). We prove the Theorem by showing that the residual norm $\|\mathbf{r}\| = O(\sqrt{\log n})$ with high probability, and thus, the claim will be proved by applying Theorem 1 and 2. To begin with, we consider the first scenario, where each corrupted feature

can be expressed as $\mathbf{x}_i^\star + \Delta\mathbf{x}_i$. The feature matrix X can thus be described as $X^\star + \Delta X$, where in ΔX there are $C' \log n$ rows to be non-zero. Let $\Delta X = U_\Delta \Sigma_\Delta V_\Delta^T$ be the reduced SVD of ΔX. Then the norm of the residual can be bounded by:

$$\begin{aligned}\|\mathbf{r}\| &\le \|U_\Delta U_\Delta^T \mathbf{s}\| \\ &= \|\Delta X V_\Delta \Sigma_\Delta^{-2} V_\Delta^T \Delta X^T \mathbf{s}\| \\ &\le \|\Delta X\|_2 \|\Sigma_\Delta^{-2}\|_2 \|\Delta X^T \mathbf{s}\| \end{aligned} \tag{12}$$

where the last term $\|\Delta X^T \mathbf{s}\| \le \sqrt{d} C' \xi \mathcal{T} \log n$. Now, to bound the first two terms, we need to bound the largest and smallest singular value of ΔX. Consider $\Delta X' \in \mathbb{R}^{C' \log n \times d}$ to be the truncated ΔX where only non-zero rows in ΔX are left. The spectrum of $\Delta X'$ is the same as ΔX. Moreover, its two norm can be bounded by:

$$\|\Delta X'\|_2 \le \|\xi E\|_2 \le \xi \sqrt{C' d \log n},$$

where $E \in \mathbb{R}^{C' \log n \times d}$ is the matrix with all entries are one. Also, using the result of [26], we can guarantee that with high probability $\sigma_d(\Delta X') \ge \Omega(\sqrt{\log n} - \sqrt{d})$, which suggests w.h.p.:

$$\|\Sigma_\Delta^{-2}\|_2 = \frac{1}{\sigma_d(\Delta X)^2} = \frac{1}{\sigma_d(\Delta X')^2} \le O(\frac{1}{\log n}).$$

Thus by substituting all above back to (12), we can conclude that $\|\mathbf{r}\| = O(\sqrt{\log n})$.

To prove the second case where $C' \log n$ items have shuffled features, note that we can still express the feature matrix $X = X^\star + \Delta X$, where now the row of ΔX follows:

$$\Delta\mathbf{x}_i = \begin{cases} \mathbf{x}_j - \mathbf{x}_i, & \text{if item } i \text{ is corrupted,} \\ 0, & \text{otherwise.} \end{cases}$$

We can further bound the infinity norm of $\Delta\mathbf{x}_i$ by $\|\Delta\mathbf{x}_i\|_\infty \le \|\mathbf{x}_j - \mathbf{x}_i\|_\infty \le \|\mathbf{x}_j - \mathbf{x}_i\| \le 2\mathcal{X}$. Now the claim is proved by applying $\xi = 2\mathcal{X}$ to the proof of scenario 1. □

Proof of Theorem 4

We will focus on proving the following theorem instead.

Theorem 5 (Kendall's Tau Guarantee for Noisy Comparisons from Flip-Sign Model)**.** *Let δ be any constant such that $0 < \delta < 1$. Suppose the following assumptions hold:*

a. We observe m noisy pairwise comparisons under the flip sign model (parameterized by some noise level $0 \le \rho_c < 0.5$).
b. Feature matrix X is γ-close with bounded $\mathcal{X}$.

Consider the following instance of RABF model (problem (3)):

$$\min_{\theta \in \mathbb{R}^{d+n}} \sum_{(i,j)\in\mathcal{S}_I} (\theta^T \bar{\mathbf{x}}_{ij} - P_{ij})^2, \quad P_{ij} \sim D_{\rho_c} \tag{13}$$
$$s.t. \quad \|\mathbf{w}\| \le (1-2\rho_c)\mathcal{W}, \quad \|\mathbf{r}\| \le (1-2\rho_c)\mathcal{R},$$

where $\mathcal{W}$ and $\mathcal{R}$ are set to be (4), and the distribution D_{ρ_c} is defined by:

$$\begin{aligned}&\Pr(P_{ij} = +1 \mid sgn(Y_{ij}) = -1) \\ =&\Pr(P_{ij} = -1 \mid sgn(Y_{ij}) = +1) \\ =&\rho_c,\end{aligned}$$

which describes the flip sign model. Then with probability at least $1-\delta$, the optimal π^* of the problem satisfies:

$$\begin{aligned}&D_{k\tau}(\pi^*, \mathbf{s}) \\ \le& O\big(\min_{f\in F_\Theta} R_\ell(f) - R_\ell^*\big) \\ &+ O\bigg(\frac{1}{1-2\rho_c}\big(\sqrt{d} + \|\mathbf{r}\|\big)\sqrt{\frac{1}{m}}\bigg) + O\bigg(\sqrt{\frac{\log 1/\delta}{m}}\bigg).\end{aligned}$$

Theorem 4 follows directly from Theorem 5 provided that $\min_{f\in F_\Theta} R_\ell(f) - R_\ell^* = O(\epsilon)$. [4] Thus, proving Theorem 5 will suffice.

However, Theorem 5 is harder to conclude compared to Theorem 1 and 2. In particular, note that when comparisons are generated from flip-sign model, the solution of the RABF model (13) is no longer the minimizer of the problem $\min_{f\in F_\Theta} \hat{R}_\ell(f)$. It is because the definition of $\hat{R}_\ell(f)$ is on the clean distribution (i.e. $P_{ij} = \text{sgn}(Y_{ij})$), while in problem (13) each P_{ij} is sampled from noise distribution D_{ρ_c}. Thus, the optimizer of problem (13) is only the minimizer over empirical risk of noisy comparisons. We again use $\theta^*/f^*/\pi^*$ to denote the optimal parameter/function/corresponding score vector of problem (13). The challenge is hence to bound the risk of f^* with respect to the clean distribution, i.e. $R(f^*)$.

The high level idea of our proof is as follows. We first show that the problem (13) is equivalent to an ERM problem with some "unbiased estimator" for the loss over *clean* distribution [21] (stating in Lemma 5 introduced shortly), and the two optimal solutions will be only different with a $(1-2\rho_c)$ factor. We then apply the result in [21] to guarantee the risk of the optimum of the equivalent problem with respect to the clean distribution, which concludes the proof.

[4] Similar to the discussion in the proof of Theorem 2, such a condition will be satisfied in nature for a sufficiently expressive F_Θ.

Before presenting the proof, we introduce a lemma which shows that the problem (13) is equivalent to another ERM problem with an unbiased estimator of squared loss with noisy labels (see Section 3 in [21] for more details):

Lemma 5 (Equivalence of Problem (13) with Unbiased Estimator)**.** *The problem* (13) *is equivalent to the following optimization problem:*

$$\min_{\tilde{\theta}=[\tilde{\mathbf{w}};\tilde{\mathbf{r}}]\in\mathbb{R}^{d+n}} \sum_{(i,j)\in\mathcal{S}_I} \tilde{\ell}(\tilde{\theta}^T\bar{\mathbf{x}}_{ij}, P_{ij}), \quad s.t. \quad \|\tilde{\mathbf{w}}\| \le \mathcal{W}, \quad \|\tilde{\mathbf{r}}\| \le \mathcal{R}, \tag{14}$$

where $\tilde{\ell}(t,y)$ is an unbiased estimator of squared loss from noisy comparisons defined by:

$$\tilde{\ell}(t,y) = \frac{(1-\rho_c)(t-y)^2 - \rho_c(t+y)^2}{1-2\rho_c}.$$

Furthermore, the optimal solution of the problem (14)*, denoted as $\tilde{\theta}^*$, satisfies:*

$$\theta^* = (1-2\rho_c)\tilde{\theta}^* \tag{15}$$

where θ^* is the optimal solution of the problem (13)*.*

The proof of Lemma 5 will be shown in next subsection for completeness. Now, with this lemma, we are ready to present the proof of Theorem 5 as follows.

Proof (of Theorem 5). Let $\tilde{\theta}^*/\tilde{f}^*/\tilde{\pi}^*$ denote the optimal parameter/function/corresponding ranking of problem (14). Then from Theorem 3 of [21], we can guarantee that with probability at least $1-\delta$, the risk of $\tilde{f}^*$ w.r.t. clean distribution is bounded by:

$$R_\ell(\tilde{f}^*) \le \min_{f\in F_\Theta} R_\ell(f) + \frac{8L_\ell}{1-2\rho_c}\mathbb{E}_{\mathcal{S}_I}\big[\mathfrak{R}(F_\Theta)\big] + 2\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{16}$$

However, since $\theta^* = (1-2\rho_c)\tilde{\theta}^*$ from Lemma 5, we know that the ranking scores of all items in π^* are only scaled by a $1-2\rho_c$ factor with respect to $\tilde{\pi}^*$ and furthermore, the *ranking order* will still remain same as $\tilde{\pi}^*$. This implies that $R(f^*) = D_{k\tau}(\pi^*, \mathbf{s}) = D_{k\tau}(\tilde{\pi}^*, \mathbf{s}) = R(\tilde{f}^*)$. Finally, by applying Lemma 2, Lemma 4 to (16), the claim of Theorem 5 can be obtained as:

$$\begin{aligned}
&D_{k\tau}(\pi^*, \mathbf{s}) \\
=&R(\tilde{f}^*) \\
\le&\Psi(R_\ell(\tilde{f}^*) - R_\ell^*) \\
\le&L_\Psi\bigg(\min_{f\in F_\Theta} R_\ell(f) - R_\ell^* \\
&+ \frac{8L_\ell}{1-2\rho_c}\Big(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}}\|\mathbf{d}\| + \|\mathbf{r}\|\Big)\sqrt{\frac{2}{m}} + 2\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\bigg) \\
=&O\Big(\min_{f\in F_\Theta} R_\ell(f) - R_\ell^*\Big) \\
&+ O\bigg(\frac{1}{1-2\rho_c}\big(\sqrt{d} + \|\mathbf{r}\|\big)\sqrt{\frac{1}{m}}\bigg) + O\bigg(\sqrt{\frac{\log 1/\delta}{m}}\bigg).
\end{aligned}$$

□

Proof of Lemma 5

Proof (of Lemma 5). First off, we rewrite the unbiased estimator of squared loss $\tilde{\ell}(t,y)$ as:

$$\begin{aligned}
\tilde{\ell}(t,y) &= t^2 - \frac{2t}{1-2\rho_c}y + y^2 \\
&= \bigg(t - \frac{y}{1-2\rho_c}\bigg)^2 + \bigg(y^2 - \frac{1}{1-2\rho_c}y^2\bigg).
\end{aligned}$$

Therefore, problem (14) can be rewritten as:

$$\begin{aligned}
&\min_{\tilde{\theta}\in\mathbb{R}^{d+n}} \sum_{(i,j)\in\mathcal{S}_I} \tilde{\ell}(\tilde{\theta}^T\bar{\mathbf{x}}_{ij}, P_{ij}) \\
\equiv&\min_{\tilde{\theta}\in\mathbb{R}^{d+n}} \sum_{(i,j)\in\mathcal{S}_I} \bigg(\tilde{\theta}^T\bar{\mathbf{x}}_{ij} - \frac{P_{ij}}{1-2\rho_c}\bigg)^2 \\
\equiv&\min_{\tilde{\mathbf{w}},\tilde{\mathbf{r}}} \sum_{(i,j)\in\mathcal{S}_I} \bigg(\tilde{\mathbf{w}}^T(\mathbf{x}_j - \mathbf{x}_i) + (\tilde{r}_j - \tilde{r}_i) - \frac{P_{ij}}{1-2\rho_c}\bigg)^2, \\
&\text{s.t.} \quad \|\tilde{\mathbf{w}}\| \le \mathcal{W}, \quad \|\tilde{\mathbf{r}}\| \le \mathcal{R}.
\end{aligned} \tag{17}$$

Now define two new variables as:

$$\begin{aligned}
\mathbf{w} &= (1-2\rho_c)\tilde{\mathbf{w}} \\
\mathbf{r} &= (1-2\rho_c)\tilde{\mathbf{r}}
\end{aligned} \tag{18}$$

and substitute (18) to the problem (17). We can further derive an equivalent optimization problem w.r.t. $\mathbf{w}$ and $\mathbf{r}$ as:

$$\begin{aligned}
&\min_{\mathbf{w},\mathbf{r}} \sum_{(i,j)\in\mathcal{S}_I} \bigg(\mathbf{w}^T(\mathbf{x}_j - \mathbf{x}_i) + (r_j - r_i) - P_{ij}\bigg)^2 \\
\equiv&\min_{\theta} \sum_{(i,j)\in\mathcal{S}_I} (\theta^T\bar{\mathbf{x}}_{ij} - P_{ij})^2, \\
&\text{s.t.} \quad \|\mathbf{w}\| \le (1-2\rho_c)\mathcal{W}, \quad \|\mathbf{r}\| \le (1-2\rho_c)\mathcal{R},
\end{aligned}$$

which is the problem (13) as claimed. In addition, from (18), the optimal solutions between two problems satisfy:

$$\theta^* = [\mathbf{w}^*, \mathbf{r}^*] = (1 - 2\rho_c)[\tilde{\mathbf{w}}^*, \tilde{\mathbf{r}}^*] = (1 - 2\rho_c)\tilde{\theta}^*$$

and the proof is thus completed. □

Proof of Theorem 6

Proof (of Theorem 6). First, note that the frequently used accumulated regret bound for online learning cannot be directly applied here, since we want to bound the excess risk achieved by the final model $\theta^{(T)}$. Therefore, in this proof we use guarantee from SGD convergence for our online-to-batch conversion. Consider Algorithm 1 as a SGD algorithm that solves the problem $\min_{f \in F_\Theta} R_\ell(f)$. Then, with a strongly convex, twice differentiable ℓ, a standard SGD convergence analysis (e.g. [17]) tells us that:

$$R_\ell(f^{(T)}) - R_\ell(f^*) \leq \frac{\hat{C} L_\ell}{2T}$$

with some constant $\hat{C}$. Now consider the batch problem (6), with m observations to be online comparisons Algorithm 1 observed (so that $m = T$). The problem shares the same f^* with Algorithm 1, and furthermore, its equivalent hard constraint problem in form (3) will satisfy equation (10). This means that we can guarantee with high probability,

$$\begin{aligned} R_\ell(f^{(T)}) \leq & 2L_\ell \left(\frac{\sqrt{2d}}{\mu\gamma\sqrt{n}} \|\mathbf{d}\| + \|\mathbf{r}\| \right) \sqrt{\frac{2}{T}} \\ & + \mathcal{B}\sqrt{\frac{\log \frac{1}{\delta}}{2T}} + \frac{\hat{C} L_\ell}{2T}, \end{aligned}$$

and the Theorem can be derived by following the same procedure below equation (10) in the proof of Theorem 1. □

Appendix C: Details of Online Rank Aggregation with Features

As introduced in Section 5.1, we could extend our RABF model to online rank aggregation by solving RABF formulation using SGD. Specifically, for each pairwise comparison P_{ij} observed at time t, we perform a SGD update on model parameters $(\mathbf{w}, \mathbf{r})$ with the following update rule:

$$\begin{aligned} \mathbf{w}^{(t+1)} \leftarrow & \mathbf{w}^{(t)} \\ & - \eta \left(\frac{\partial \ell({\mathbf{w}^{(t)}}^T (\mathbf{x}_j - \mathbf{x}_i) + r_j - r_i, P_{ij})}{\partial \mathbf{w}} + \lambda_w \mathbf{w}^{(t)} \right) \\ \mathbf{r}^{(t+1)} \leftarrow & \mathbf{r}^{(t)} \\ & - \eta \left(\frac{\partial \ell({\mathbf{w}^{(t)}}^T (\mathbf{x}_j - \mathbf{x}_i) + r_j - r_i, P_{ij})}{\partial \mathbf{r}} + \lambda_r \mathbf{r}^{(t)} \right) \end{aligned} \tag{19}$$

The procedure of the online-RABF algorithm is shown in Algorithm 1. The following Theorem provides a guarantee on the output of the score vector $\pi^{(T)}$ from online-RABF algorithm:

Algorithm 1 Online RABF (oRABF)

Input: feature matrix X, parameters (λ_w, λ_r), step size η.
$\mathbf{w}^{(0)} \leftarrow 0$, $\mathbf{r}^{(0)} \leftarrow 0$.
for $t = 1, 2, \cdots T$ **do**
  Update $\mathbf{w}^{(t+1)}, \mathbf{r}^{(t+1)}$ using rule (19) based on the given the observed P_{ij} at time t.
end for
return $\pi^{(T)} = X\mathbf{w}^{(T)} + \mathbf{r}^{(T)}$

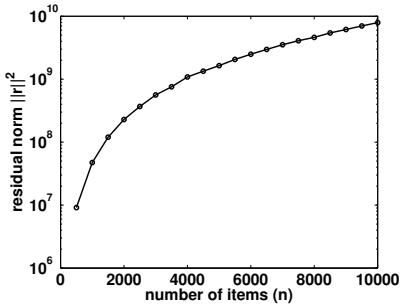
Theorem 6. *Suppose assumptions b, c in Theorem 1 hold, ℓ is strongly convex and twice differentiable, and n is sufficiently large. Then by running Algorithm 1 with appropriate setting of (λ_w, λ_r), with high probability, its output score vector $\pi^{(T)}$ satisfies:*

$$D_{k\tau}(\pi^{(T)}, \mathbf{s}) \leq O(\sqrt{\frac{\|\mathbf{r}\|^2}{T}}).$$

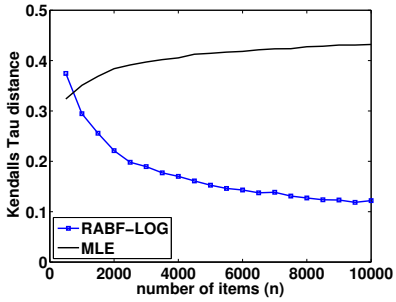
A similar result can also be proved for $P_{ij} = \text{sgn}(Y_{ij})$. As a consequence, Algorithm 1 only needs $O(\|\mathbf{r}\|^2/\epsilon^2)$ online updates to guarantee an ϵ-accurate ranking, which again implies that given good features such that $\|\mathbf{r}\|^2 = o(n)$, sublinear number of samples is sufficient. The result shows that the sublinear sample complexity is also achievable by online RABF as in batch setting. The proof of Theorem 6 can be found in Appendix B.

Appendix D: Empirical Justification of Sublinear Sample Complexity

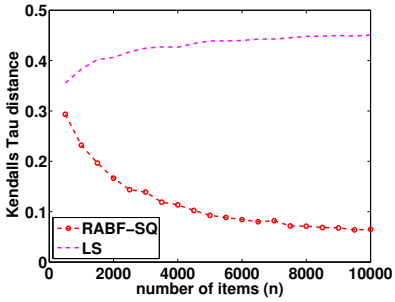
In this experiment, we show that sample complexity of RABF can be sublinear given sufficiently good features for both noiseless and noisy comparison cases. We consider synthetic datasets generated by the procedure described in Section 6. We generate several true score vectors $\mathbf{s} \in \mathbb{R}^n$ with n from 500 to 10000. For each n, we further generate a perturbed feature matrix X with



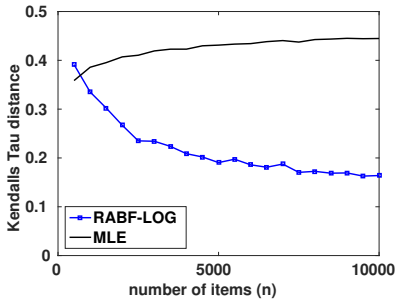
(a) (log scale of) $\|\mathbf{r}\|^2$ versus n



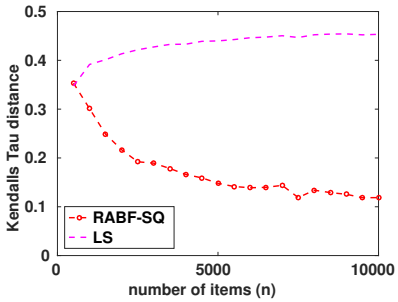
(b) $\rho_c = 0, P_{ij} = \text{sgn}(Y_{ij})$



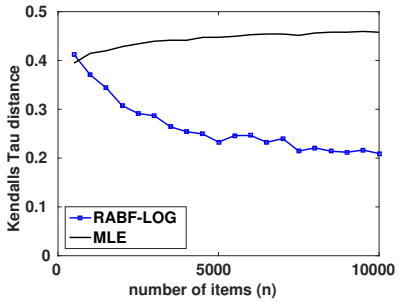
(c) $\rho_c = 0, P_{ij} = Y_{ij}$



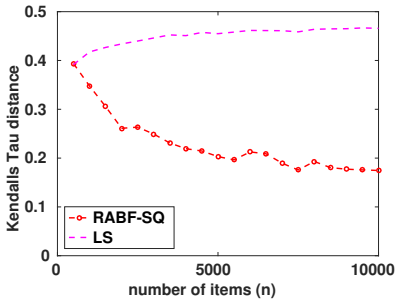
(d) $\rho_c = 0.1, P_{ij} = \text{sgn}(Y_{ij})$



(e) $\rho_c = 0.1, P_{ij} = Y_{ij}$



(f) $\rho_c = 0.2, P_{ij} = \text{sgn}(Y_{ij})$



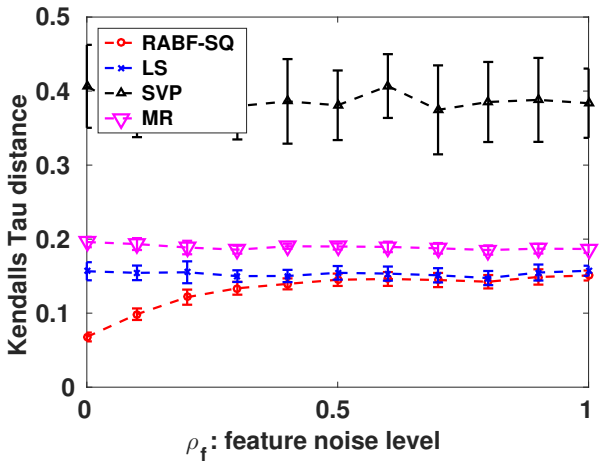
(g) $\rho_c = 0.2, P_{ij} = Y_{ij}$

Figure 2: A synthetic experiment where $O(\log n)$ item features are corrupted. Figure 2a shows that the feature quality is good as $\|\mathbf{r}\|^2$ grows under the order of $\log n$. Figure 2b$\sim$2g show that for our RABF model, $O(\log n)$ comparisons suffice to output an ϵ-accurate ranking with bounded $D_{k\tau}$, while for methods without features $D_{k\tau}$ becomes unbounded as n increases. In addition, the argument holds regardless of whether comparisons are clean ($\rho_c = 0$) or noisy ($\rho_c = 0.1, 0.2$). The results empirically support the fact that RABF is able to leverage informative features to achieve faster learning (i.e. sublinear sample complexity) as shown in theory.

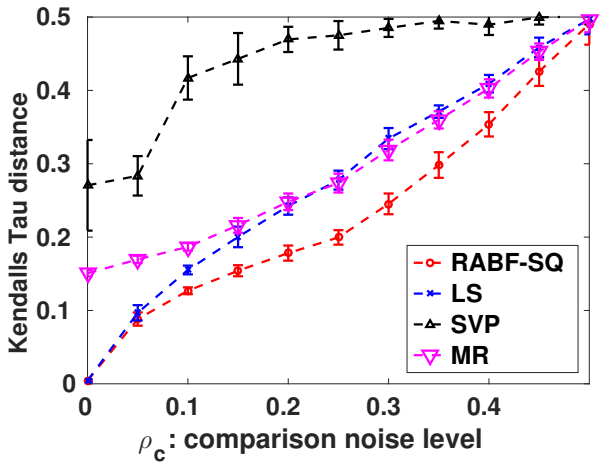
$\rho_f = 50 \log n/n$, so there are $O(\log n)$ items having corrupted features in X by construction. We first sample $m = 50 \log n$ clean pairwise comparisons ($\rho_c = 0$) and apply the proposed methods (RABF-LOG and RABF-SQ) and methods without features (MLE and LS) to recover the ranking. The results are shown in Figure 2. In Figure 2a, we observe that $\|\mathbf{r}\|^2$ grows $O(\log n)$ in this scenario. Hence, from Corollary 1, $m = O(\log n)$ should suffice for our model RABF to guarantee an ϵ-accurate ranking with bounded $D_{k\tau}$. This is indeed true as suggested in Figure 2c and 2b, where Kendall's Tau of the rankings from RABF-SQ and RABF-LOG do not grow with n provided $O(\log n)$ comparisons. As a comparison, both LS and MLE fail to output good rankings (i.e. bounded $D_{k\tau}$) with only $O(\log n)$ comparisons as n goes large. Furthermore, we redo the same experiment except that now the sampled comparisons changed to be noisy ($\rho_c = 0.1$ and 0.2). The results are shown in Figure 2e to 2f. From these figures, we can observe that $O(\log n)$ samples are still sufficient for RABF to guarantee a ranking with bounded $D_{k\tau}$ for noisy comparisons case. These experiments empirically confirm the fact that by making use of informative features, RABF is able to produce an ϵ-accurate ranking with only sublinear number of (either clean or noisy) comparisons.

Appendix E: Experiments of Rank Aggregation Methods for $P_{ij} = Y_{ij}$

Here we show the experimental results of rank aggregation methods for $P_{ij} = Y_{ij}$, where the detailed experiment setup is described in Section 6.1. Figure 3a and 3b are results on synthetic datasets where we perturb features and comparisons and compare the robustness of each model. Figure 4a and 4b are results on Forbes and NBA datasets as real-world applications. Similar to the results for $P_{ij} = \text{sgn}(Y_{ij})$, here we see RABF-SQ also outperforms other existing methods, showing the effectiveness of our model for rank aggregation task for the case $P_{ij} = Y_{ij}$.

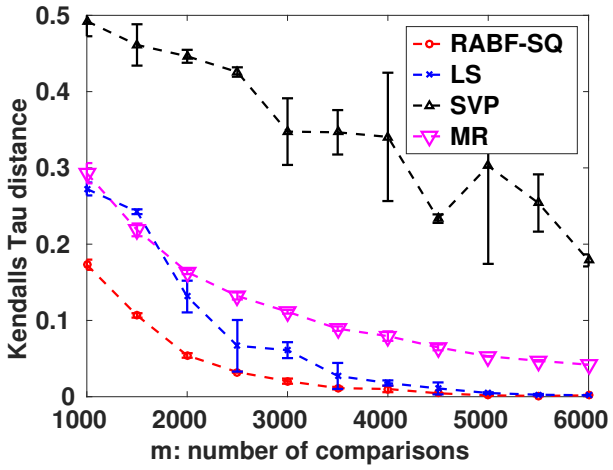


(a) Feature perturbation

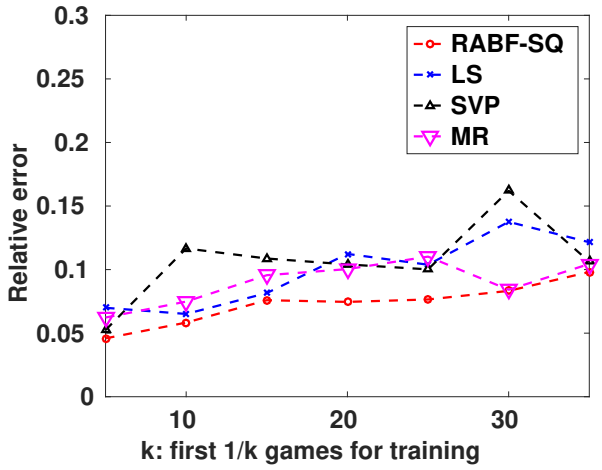


(b) Comparisons perturbation

Figure 3: Performance of rank aggregation methods for $P_{ij} = Y_{ij}$ on synthetic datasets. Similar to Figure 1a and 1b, RABF-SQ performs the best under different feature and comparison noise levels.



(a) Forbes



(b) NBA

Figure 4: Performance of rank aggregation methods for $P_{ij} = Y_{ij}$ on real-world datasets. Similar to Figure 1c and 1d, here we see that RABF-SQ model has smaller sample complexity in real-world applications compared to other methods.