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:

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

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:

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

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:

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

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.,

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

where $\Sigma_\mu^\dagger = \text{diag}(1/\sigma_1, 1/\sigma_2 \cdots 1/\sigma_{\bar{d}})$. 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:*

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

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:

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

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

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

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:

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

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

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 $\hat{R}_\ell(f^*) \le \hat{R}_\ell(f_{\hat{\theta}})$. Nevertheless, $\hat{\theta}$ still provides an instance to show $R^* = 0$. Thus, by applying Lemma 4 in this case, we have:

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

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

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

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 $\hat{R}_\ell(f^*) \le \hat{R}_\ell(f_{\hat{\theta}})$, 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:

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

□

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:

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

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)):

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

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:

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

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:*

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

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:

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

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:

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

□

Proof of Lemma 5

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

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

Therefore, problem (14) can be rewritten as:

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

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:

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

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:

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 $\pi^{(T)} = X\mathbf{w}^{(T)} + \mathbf{r}^{(T)}$

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

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:

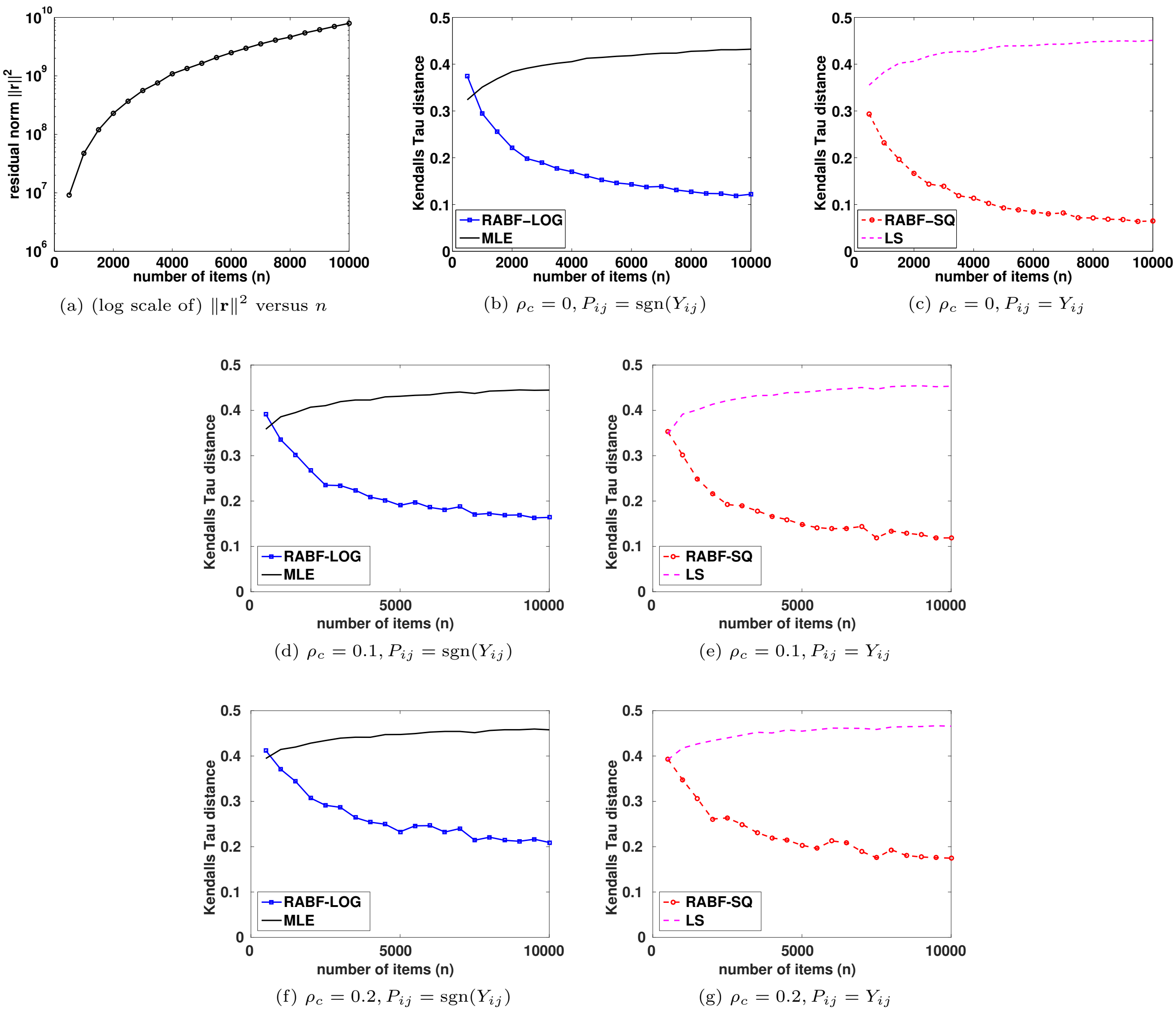
Theorem 6. *Suppose assumptions b, c in Theorem 1 hold, ℓ is strongly convex and twice differentiable, and n is sufficiently large. Then by running Algorithm 1 with appropriate setting of (λ_w, λ_r), with high probability, its output score vector $\pi^{(T)}$ satisfies:*

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

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

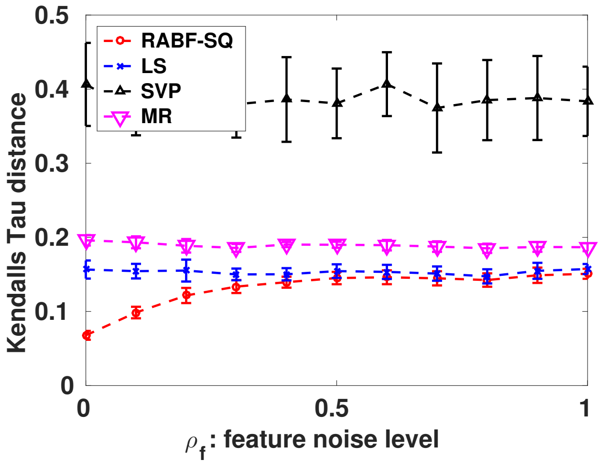
(g) $\rho_c = 0.2, P_{ij} = Y_{ij}$

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∼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.

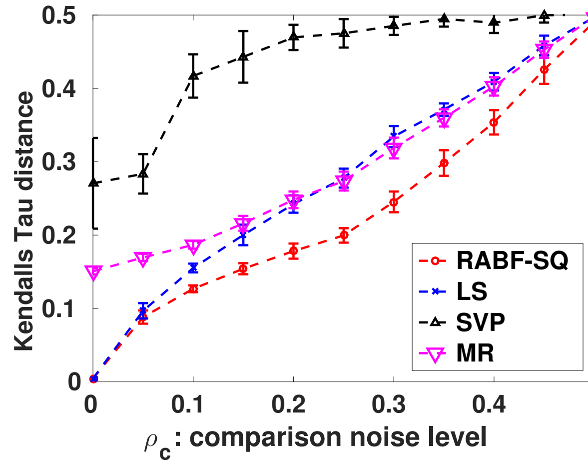
$\rho_f = 50 \log n/n$, so there are $O(\log n)$ items having corrupted features in X by construction. We first sample $m = 50 \log n$ clean pairwise comparisons ($\rho_c = 0$) and apply the proposed methods (RABF-LOG and RABF-SQ) and methods without features (MLE and LS) to recover the ranking. The 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 $P_{ij} = Y_{ij}$.

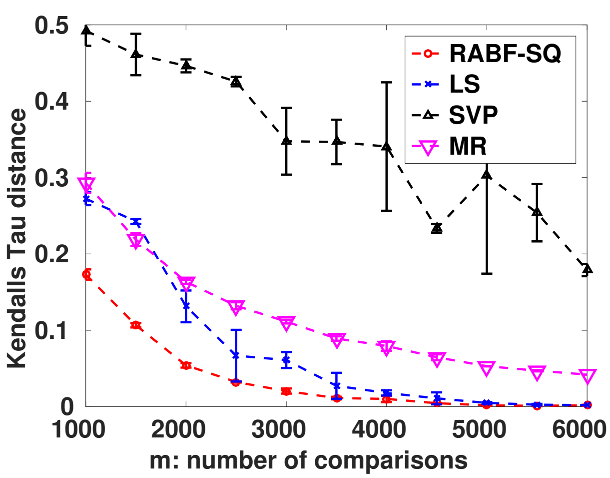


(a) Feature perturbation

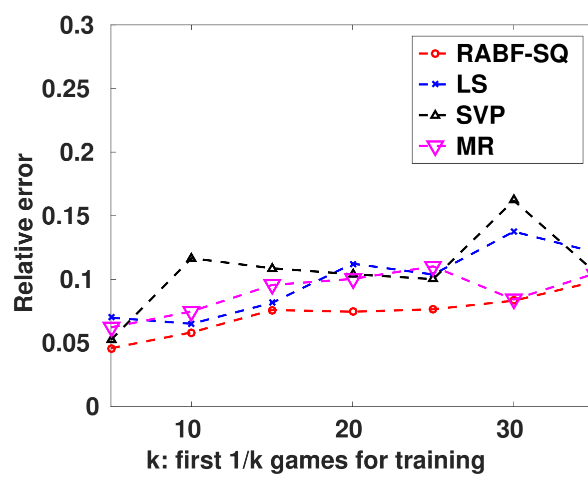


(b) Comparisons perturbation

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.