Ranking with Features: Algorithm and A Graph Theoretic Analysis

Aadirupa Saha∗, Arun Rajkumar†

Abstract

We consider the problem of ranking a set of items from pairwise comparisons in the presence of features associated with the items. Recent works have established that $O(n \log(n))$ samples are needed to rank well when there is no feature information present. However, this might be sub-optimal in the presence of associated features. We introduce a new probabilistic preference model called feature-Bradley-Terry-Luce (f-BTL) model that generalizes the standard BTL model to incorporate feature information. We present a new least squares based algorithm called fBTL-LS which we show requires much lesser than $O(n \log(n))$ pairs to obtain a good ranking – precisely our new sample complexity bound is of $O(\alpha \log \alpha)$, where $\alpha$ denotes the number of ‘independent items’ of the set, in general $\alpha << n$. Our analysis is novel and makes use of tools from classical graph matching theory to provide tighter bounds that sheds light on the true complexity of the ranking problem, capturing the item dependencies in terms of their feature representations. This was not possible with earlier matrix completion based tools used for this problem. We also prove an information theoretic lower bound on the required sample complexity for recovering the underlying ranking, which essentially shows the tightness of our proposed algorithms. The efficacy of our proposed algorithms are validated through extensive experimental evaluations on a variety of synthetic and real world datasets.

1 Introduction

We consider the problem of ranking a set of items from pairwise comparisons when some side information about the items is available. Given a set of $n$ items and $m$ pairwise comparisons among them, the problem of ranking from pairwise comparisons is to recover an underlying ranking among the $n$ items. This is a well studied problem in several disciplines including statistics, operations research, theoretical computer science, social choice theory, machine learning, decision systems etc [Thurstone(1927), Bradley and Terry(1952), Luce(1959), Saaty(2008), Ailon et al.(2008)Ailon, Charikar, and Newman], [Braverman and Mossel(2008),

∗Indian Institute of Science. aadirupa@iisc.ac.in
†Conduent Labs, India. arun.rajkumar@conduent.com
Gleich and Lim(2011), Jamieson and Nowak(2011), Negahban et al.(2012)Negahban, Oh, and Shah], [Wauthier et al.(2013)Wauthier, Jordan, and Jojic, Busa-Fekete et al.(2014)Busa-Fekete, Hüllermeier, and Szörényi, Rajkumar and Agarwal(2014), Shah and Wainwright(2015), Borkar et al.(2016)Borkar, Karamchandani, and Mirani, Chen and Joachims(2016), Rajkumar and Agarwal(2016), Shah et al.(2016)Shah, Balakrishnan, Guntuboyina, and Wainwright, Niranjan and Rajkumar(2017)]. A typical approach to solve this problem is to assume that the comparisons are generated in a stochastic fashion according to a score based pairwise probability model such as the Bradley-Terry-Luce model [Bradley and Terry(1952) Luce(1959)] or the Thurstone model [Thurstone(1927)] and to develop algorithms [Gleich and Lim(2011)], [Negahban et al.(2012)Negahban, Oh, and Shah], [Rajkumar and Agarwal(2014)], [Borkar et al.(2016)Borkar, Karamchandani, and Mirani] that first estimate the score vector generating the comparisons and rank the items by simply sorting the estimated scores. However in practice these algorithms suffer from several shortcomings. Firstly, side information such as features or certain relationships among items is often available. For example, to rank a set of mobile phones, it is natural to use features such as cost, battery life, size etc. which influence the pairwise preferences of users in preferring one mobile over the other. However, most algorithms do not take into account the valuable feature information associated with items. Secondly, they fail to handle the case when new items get added to the existing set i.e., one cannot find the position of a new item in an already estimated ranking without collecting pairwise preferences of the newly added item with at least some items in the existing set. Finally, the sample complexity of previous approaches scale as $O(n \log(n))$ which may often be sub-optimal in the presence of features.

In this work, introduce the feature-Bradley–Terry–Luce (f-BTL) model of pairwise comparisons to tackle the problems listed above. The f-BTL model is a generalization of the standard BTL model where the probability of preferring one item over the other explicitly depends on their associated features. We propose $fBTL-LS$, a least squares based algorithm to recover the underlying true scores (and ranking) under the f-BTL model. The novelty of our approach is in the analysis of the sample complexity (the number of comparisons needed to achieve a fixed error) of our algorithm to recover a good ranking which we show is often smaller than $O(n \log(n))$, based on the structure of items dependencies. Our improved sample complexity guarantee is of $O(\alpha \log \alpha)$, where on an intuitive level, $\alpha$ denotes the number of the main (independent) items which influence the preference structures of the rest of $n - \alpha$ items in the set — This clearly shows a significant reduction in the number of comparisons needed, compared to the earlier known bound $O(n \log(n))$, especially when $\alpha \ll n$, which is often the case for various practical scenarios. The key ingredient which helps us achieve this is a relation graph that we define on the features. We apply ideas from classical graph matching theory on the relation graph and show how they help us prove our sample complexity bounds. We demonstrate the usefulness of our bounds by deriving sample complexity results for several special cases of relation graphs such as cliques, disconnected cliques, trees, star graphs, cycles etc. By explicitly modelling dependencies among features (rather than just considering them to lie in some low dimensional space), our bounds reveal the true
complexity of the problem. We believe that the graph theory based approach used in this work would be of wider use to the learning theory community. Furthermore, we also give a matching lower bound guarantee analyzing the minimal number of pairwise preferences (sample complexity) required for estimating a ‘good ranking’, which in fact establishes the optimality of our proposed algorithms. Our experiments on synthetic and real world preference data sets show that the proposed algorithm significantly outperforms existing algorithms.

1.1 Related Work

As mentioned earlier, ranking from pairwise comparisons has been studied extensively in various disciplines and it is not in the scope of this paper to review all the previous works. We review the most relevant work to the setting considered here. The work that is most related to our setting is that of [Niranjan and Rajkumar(2017)] which also uses features associated with items. They assume the features lie in some low dimensional space and use a matrix completion based approach to obtain a ranking. The low rank assumption is a global assumption on the features and might miss out completely on the exact dependencies on them. As we will see, the set of features in a low dimensional space might give rise to very different type of relation graphs which may lead to very different sample complexity bounds that our analysis will capture while theirs does not. [Gleich and Lim(2011)], [Borkar et al.(2016)] give least squares based algorithms. However, they do not consider feature information. [Negahban et al.(2012)] give least squares based algorithms. However, they do not consider feature information. [Negahban et al.(2012)] work in the pairwise ranking setting under different probabilistic models (including the BTL model). Again, none of these use features explicitly and hence (as we will see in the experiments) are sub-optimal for our setting. [Jamieson and Nowak(2011)] work in a setting where the probabilities come from some unknown low dimensional feature embedding of the items. However they require the pairs to be queried actively while we work in the passive setting. There is a rich ranking literature on noisy sorting [Braverman and Mossel(2008)], approximation algorithms [Ailon et al.(2008)] etc., which are fundamentally different from the passive setting under the BTL model considered in this work.

1.2 Summary of Contributions

The main contributions of this work are as follows:

• We introduce a new probabilistic model, f-BTL, for ranking from pairwise comparisons which explicitly uses features associated with items (Section 2).
• We give a novel analysis for the sample complexity of the proposed model by using ideas from graph matching theory that captures the dependencies among features explicitly than previous approaches (Section 3).

• We propose a novel least squares based algorithm, fBTL-LS, and provide its sample complexity guarantees for recovering a ‘good estimate’ of the score vector (and ranking) under the f-BTL model (Section 4).

• We further prove that the sample complexity guarantee of our proposed fBTL-LS algorithm is tight proving a matching lower bound for the problem (Section 5).

• We finally give supporting experimental results to show the efficacy of our algorithm on both synthetic data (which follows the f-BTL model) and on real datasets which not necessarily follow the f-BTL model (Section 6).

1.3 Organization

We give the necessary preliminaries in Section 2.1 and define the problem formally. In Section 3, we analyze the case when the probability values for the sampled pairs are known exactly and derive a graph matching theory based sample complexity bound. In Section 4, we propose our least squares based algorithm and show theoretical guarantees of it’s performance. Section 5 proves a matching lower bound guarantee for the problem. In section 6, we experimentally evaluate our algorithm on various synthetic and real world data sets. We conclude in Section 7 with directions for future work. All proofs are presented in the appendix.

Notation: We use lower case boldface letters for vectors, upper case boldface letters for matrices, lower case letters for scalars and upper case letters for constants. \( \| \cdot \|_2 \) denotes the \( \ell_2 \) norm for vectors and spectral norm for matrices. \( \| \cdot \|_F \) denotes the Frobenius norm for matrices. We denote the set \{1, \ldots, n\} by \([n]\). For a square matrix \( A \in \mathbb{R}^{n \times n} \), we denote by \( \lambda_{\text{max}}(A) \) the magnitude of the largest eigenvalue of \( A \).

2 Preliminaries

BTL Model: ([Bradley and Terry(1952)], [Luce(1959)]) The standard probabilistic model for pairwise comparisons is the Bradley–Terry–Luce model. In this model, the probability of preferring item \( i \) over item \( j \) is given by

\[
P_{ij} = \frac{w_i}{w_i + w_j}
\]

where \( w \in \mathbb{R}_+ \) is a vector of scores associated with the items.
Hall’s Marriage Theorem: In our sample complexity analysis, we will use this classical result in bipartite graph matching.

**Theorem 2.1 ([Hall(1935)])**. Let $C = (A \cup B, E)$ be a finite bipartite graph. For $S \subseteq A$, let $N_C(S)$ denote the neighbours of $S$ in $B$. Then $C$ admits a matching that entirely covers $A$ if and only if

$$|N_C(S)| \geq |S| \quad \forall S \subseteq A$$

2.1 Problem Setting

Let $[n] = \{1, \ldots, n\}$ be the set of items to be ranked. We assume that the items are related to each other using a relation graph $G([n], E)$ where a pair of items are related iff there is a corresponding edge in $G$. The items have associated feature vectors $U = \{u_1, \ldots, u_n\}$ where each $u_i \in \mathbb{R}^d$. The following natural assumption relates $G$ and $U$:

**Assumption**: The subset of vectors in $U$ corresponding to the items in the independent set $\mathcal{I}(G)$ of $G$ are linearly independent and form a basis for $\text{span}(U)$.

Let $\alpha = \alpha(G) = |\mathcal{I}(G)|$ be the independence number of $G$ and $d \geq \alpha$. Let $B \in \mathbb{R}^{n \times \alpha}$ be the coefficient matrix that expresses $U$ in terms of the basis vectors as follows:

$$u_i = \sum_{j \in \mathcal{I}(G)} B_{ji} u_j \quad \forall i \in [n] \quad (2)$$

We will assume (without loss of generality) the set $[\alpha]$ to be $\mathcal{I}(G)$. When there are multiple independent sets of $G$, we arbitrarily fix one.

**Model**: We introduce the feature Bradley–Terry–Luce model (f-BTL) where the probability of preferring item $i$ over $j$ is given by:

$$P_{ij} = \frac{\exp(w^T u_i)}{\exp(w^T u_i) + \exp(w^T u_j)}$$

where $w \in \mathbb{R}^d$. The f-BTL model reduces to the standard BTL model when $\alpha = n$ and $U$ is the standard basis. Let $\theta \in \mathbb{R}^n$ be the score vector where $\theta_i = w^T u_i$.

**Sampling**: We assume that a set $M$ of $m \in \binom{[n]}{2}$ pairs is generated where each pair is chosen with probability $p$. Each pair in $M$ is compared $K$ times independently according to f-BTL model.

**Problem**: Given $G$ and $U$, for what values of $m$ and $K$ under the above sampling model does one have an algorithm whose estimated score vector $\hat{\theta}$ satisfies

$$P(\|\theta - \hat{\theta}\|_2 \leq \epsilon) > 1 - \delta$$

5
Previous results show that under the standard BTL model, the rank centrality [Negahban et al. (2012) Negahban, Oh, and Shah] [Rajkumar and Agarwal (2016)], maximum likelihood under the BTL model [Shah and Wainwright (2015)] and the least squares [Borkar et al. (2016) Borkar, Karamchandani, and Mirani] algorithms need $O(n \log(n))$ comparisons to achieve a small error with probability at least $1 - O(poly(1/n))$. However, these algorithms do not consider the features explicitly. The feature low rank model of [Niranjan and Rajkumar (2017)] uses features but requires $O(d^2 \log(n))$ pairs to be compared. As we will see, the fBTL-LS algorithm that we propose will require much lesser samples than $O(n \log(n))$.

3 Analysis: Case When Probabilities are Known

We begin by analyzing the problem for the noiseless case where for every pair $(i, j)$ that is compared, we have access to the exact value for $P_{ij}$. This analysis will shed light into the structure of the problem which will be useful to analyse the case when the probabilities for each $(i, j)$ have to be estimated using the $K$ comparisons. Under this setting, the goal is to bound the number of samples $m$ needed to exactly recover the score vector $\theta$ where $\theta_i = w^T u_i \forall i \in [n]$. From Equation 2, we have that

$$w^T u_i = \sum_{j \in I(G)} B_{ji} w^T u_j,$$

or equivalently,

$$\theta_i = \sum_{j \in I(G)} B_{ji} \theta_j \forall i \in [n]. \quad (3)$$

As we have access to $U$ and $B$, we only need to recover the scores of $\theta_j = w^T u_j \forall j \in [\alpha]$ so that the remaining scores can be computed using Equation 3. For a pair $(i, j)$, under the f-BTL model, the following holds:

$$\sum_{k=1}^{\alpha} \gamma_{k}^{ij} \theta_k = \sum_{k=1}^{\alpha} \gamma_{k}^{ij} (w^T u_k) = 0 \quad (4)$$

where $\gamma_{k}^{ij} = B_{ik} - \frac{P_{ij}}{P_{ji}} B_{jk}$. Equation 4 shows that knowing $P_{ij}$ for any pair $(i, j)$ gives rise to a linear equation involving the score vectors corresponding to the items in $I(G)$. Since the f-BTL model is invariant to scaling the score vector $\theta$ by a positive constant, we can w.l.o.g assume that one of the scores is normalized to 1. Thus to recover the scores for all the items, it seems like we only need $I(G) - 1$ equations like Equation 4 that can be used to solve for the scores of the items in $I(G)$. However, if the coefficients $\gamma_{k}^{ij}$ are 0, then the corresponding equation does not involve the $k$-th independent set item. Thus, the equations (i.e., the pairs selected) should be such that each item in $I(G)$ appears in at least one of the selected equations so that it can be solved for.

Thus our problem now is to compute the number of pairs needed (under the sampling
model) to ensure that with high probability each item in $\mathcal{I}(G)$ appears in at least one equation of the form of Equation 4. To compute this number, we need to explicitly model the dependencies among features. We do this below and prove the necessary result using the Hall’s marriage theorem, a classical result from graph matching theory (refer Section 2).

The bipartite graph $C = (A \cup B, \Delta)$ that will be of interest to us is as follows: The set $A$ is just the set of items in the independent set i.e., $A = [\alpha]$. The set $B$ consists of $\binom{n}{2}$ nodes, one corresponding to each edge $(i,j)$. For an edge $(i,j)$, define

$$F_{ij} = \{ k \in [\alpha] : \gamma_{k}^{ij} \neq 0 \}$$

(5)

$F_{ij}$ is the set of independent set nodes whose coefficients are non-zero in the equation (refer Eq: 4) induced by the pair $(i,j)$. Thus, by observing the pair $(i,j)$, we have an equation involving the items in $F_{ij}$. We define the edge set $\Delta$ such that an edge from node $k \in [\alpha]$ to an edge $(i,j)$ is present in the bipartite graph $C$ iff $k \in (i,j)$. For any set of edges $M \subseteq \binom{n}{2}$, we define the reduced bipartite graph $C_M = (A \cup M, \Delta_M)$ by restricting the $B$ to $M$ and defining $\Delta_M$ correspondingly.

**Theorem 3.1.** Given a set of edges $M \subseteq \binom{n}{2}$, the bipartite graph $C_M = (A \cup M, \Delta_M)$ as defined above admits a matching that covers $A$ if and only if the system of linear equations induced by the edges can be solved for.

**Proof.** If there is a matching that covers $A$, then each node $i$ in $\mathcal{I}(G)$ has a distinct representative edge in $M$ which induces an equation containing $i$. Thus there are at least $\mathcal{I}(G)$ equations with each node appearing in at least one of them and hence the system can be solved for. On the other hand, if there is no matching that covers $A$, then by Hall’s theorem 2.1, there must exist some subset $S \subseteq A$ such that it’s neighbours $\left| N_{C_M}(S) \right| < |S|$. As the total number of equations that involve nodes in $S$ are less than the number of nodes, this set of equations cannot be solved for.

Theorem 3.1 gives us a novel way to analyse the number of pairs needed to obtain enough equations to solve for the score vector $\theta$. In particular, we only need to bound the probability that the Hall’s marriage condition is not met to get an upper bound on the number of pairs needed. Before we prove the result, we need the following definitions for a given set $M$. Let $M_k$ denote the neighbours of node $k$ in $C_M$. Let $c_I = | \bigcup_{k \in I} M_k |$, $d_I = | \bigcap_{k \in I} M_k |$, $I \subseteq \mathcal{I}(G)$. We now prove the main result of this section

**Theorem 3.2 (Bound On Error Probability).** Given a relation graph $G$, feature matrix $U$, a set of pairs $M$ where $|M| = m$ generated according to the sampling model above (where each pair is chosen with probability $p$), and the exact preference probabilities $P_{ij} \forall (i,j) \in M$, the probability that the score vector $\theta$ is same as that estimated score vector $\hat{\theta}$ that is got by solving the equations obtained is bounded by
\[ P(\hat{\theta} \neq \theta) \leq \min\{\alpha(G), (d_{\text{max}}(G)+1)\} \sum_{q=1}^{\sum_{I \subseteq \mathcal{I}(G)} |I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{(c_I-(q-1))}, \]

\(d_{\text{max}}(G)\) being the maximum degree of \(G\). The above theorem gives us a way of choosing \(p\) such that the probability of not satisfying the Hall’s condition (and hence not having enough equations to solve) can be bounded by a suitable value. As can be seen in the Theorem, the quantities of interest are \(c_I\) and \(d_I\) which capture the dependencies among the feature vectors of the nodes in the graph. For several common types of graphs, these quantities are easily computable and readily yield sample complexity bounds. We prove these results for some special cases below:

**Theorem 3.3 (Sample Complexity for Common Graphs).** Under the settings of Theorem 3.2, the following sample complexity bounds hold with probability at least \(1 - \delta\):

- If \(G\) is a disconnected graph, star graph, or a cycle, \(m = O(n \log(\frac{n}{\delta}))\)
- If \(G\) is a clique, \(m = O(\log(\frac{1}{\delta}))\)
- If \(G\) is a \(r\)-disconnected clique (i.e. union of \(r\) cliques), \(m = O(r \log(\frac{r}{\delta}))\)

**Remarks:** The above theorem captures the relation between the structure of the graph (and the induced dependencies among the features) and the sample complexity needed to recover the score vector under the f-BTL model. For instance, if the graph is a clique, then there is only one independent vector and hence one recovers the score using \(O(1)\) pairs. However, when the graph is disconnected, there are \(n\) independent vectors and we recover the result for the BTL model. More importantly, there are graphs (such as \(r\)-disconnected cliques) where the number of pairs needed scale as \(O(\alpha \log(\alpha))\) and are independent of the total number of nodes \(n\). This is an example where we get significant improvements in the sample complexity by exploiting the structure of the features which [Niranjan and Rajkumar(2017)] fails to achieve. In the appendix, we also discuss the sample complexity for other graphs such as regular graphs and trees.

## 4 Algorithm For General Case

In this section, we consider the original problem where we don’t have access to the exact \(P_{ij}\) values but only estimates of it available from the \(K\) independent comparisons made. In this setting, we cannot expect to solve the linear equations exactly. We propose f-BTL, a least squares based algorithm, shown in Algorithm 1 to solve for the score vector. Let the graph induced by the edge set \(M\) on the \(n\) nodes be called the comparison graph. The node-edge Incidence matrix \(Q \in \mathbb{R}^{n \times m}\) used in the algorithm is such that \(QQ^T\) is the standard unnormalized Laplacian associated with the comparison graph i.e., \(L = QQ^T = D - A\) where \(D\) is the diagonal matrix of degrees and \(A\) is the adjacency matrix. Algorithm 1 is
motivated using the fact that when the true probabilities are known exactly, the following holds:

\[ Q^T B v = y \]  

where \( \forall (i, j) \in M, y_{ij} = \log \left( \frac{P_{ij}}{P_{ji}} \right) \) and where \( v \in \mathbb{R}^\alpha \) such that \( v_i = \theta_i, \forall i \in [\alpha] \). As we only have estimates \( \hat{y} \) instead of \( y \), we take a least squares approach to estimate the scores.

**Algorithm 1 Algorithm: fBTL-LS**

**Require:** \( G, U \), a set \( M \) of \( m \) pairs each compared \( K \) times.

- Compute \( B \) from \( U \) such that Equation 2 is satisfied for all \( u_i, i \in [n] \).
- Compute the node-edge incidence matrix \( Q \in \mathbb{R}^{n \times m} \) from \( M \). Let \( \tilde{Q} = B^T Q \)
- Compute \( \hat{P}_{ij} = \begin{cases} \text{fraction of times } i \text{ beats } j \forall (i, j) \in M \\ 0 \forall (i, j) \notin M \end{cases} \)
- Compute \( \hat{y} \in \mathbb{R}^m \) where \( \forall (i, j) \in M, \hat{y}_{ij} = \log \left( \frac{\hat{P}_{ij}}{\hat{P}_{ji}} \right) \)
- Solve \( \hat{v} = \arg \min_{x \in \mathbb{R}^\alpha} \| \tilde{Q}^T x - \hat{y} \| \)
- Set \( \hat{\theta}_i = \begin{cases} \hat{v}_i \forall i \in [\alpha] \\ \text{compute using Equation (3)} \forall i \notin [\alpha] \end{cases} \)

**return** score vector \( \hat{\theta} \)

### 4.1 Connectivity

The results of [Borkar et al.(2016)](Borkar, Karamchandani, and Mirani) show that the sample complexity for the least squares algorithm for the standard BTL model depends on how well connected the comparison graph is. In particular this is measured w.r.t the second Eigenvalue of the Laplacian \( L \) which is 0 if and only if the comparison graph is disconnected. Thus when the comparison graph is disconnected, there is no way to recover the score vector in the standard BTL case. However, as we will see below, our analysis will depend on the least eigenvalue of the matrix \( \tilde{Q} \tilde{Q}^T \) and not the Laplacian matrix. The important point to note here is that even if the comparison graph is disconnected, the fBTL-LS algorithm may still recover the score vector. This is because of the fact that the algorithm makes use of the matrix \( B \) of coefficients to relate scores across possibly disconnected components in the comparison graph.

An example of this is shown in Figure ??.

Here \( n = 3 \) and \( M = \{(1, 2), (1, 3), (4, 5)\} \) and \( m = |M| = 3 \). The comparison graph as can be seen in the figure is disconnected. The nodes circled in red are assumed to be the independent set nodes. The exact relation between the feature vectors of the independent set i.e., \( \{u_1, u_2\} \) and those not in the independent set i.e., \( \{u_3, u_4, u_5\} \) are given by the matrix \( B \) shown in the figure. It can be verified for this example
that the matrix $B^T L B$ (also shown in the figure) has non zero eigenvalues though the Laplacian is block diagonal (which happens if and only if the comparison graph is disconnected as in this case).

We now prove the main result of this section:

**Theorem 4.1 (Recovery Guarantee for fBTL-LS Algorithm).** Let $M$ be a set of $m$ edges generated as per the sampling model and let each pair in $M$ be compared $K$ times independently according to the f-BTL model. Then for any positive scalar $K \geq 6(1 + e^{2b})^2 \log n$, with probability at least $1 - \frac{2}{m} \alpha$, the normalized $\ell_2$-error of Algorithm 1 satisfies

$$\frac{\|\hat{\theta} - \theta\|}{\|	heta\|} \leq \frac{2}{a} \sqrt{\frac{\lambda_{\text{max}}(B^T B)}{\lambda_{\text{min}}(B^T B)}} \cdot \sqrt{\frac{m}{\alpha}} \cdot \sqrt{\frac{\lambda_n}{\lambda_1}},$$

where $\lambda_1 = \min\{\lambda > 0 \mid \lambda \text{ is an eigen value of } B^T L B\}$, $\lambda_n = \lambda_{\text{max}}(B^T L B)$. Similarly $\lambda_{\text{min}}(B^T B)$ and $\lambda_{\text{max}}(B^T B)$ respectively denotes the minimum and maximum non-zero eigenvalues of the positive semi-definite matrix $B^T B$. $a, b > 0$ denote the range of the f-BTL parameter such that $|\theta_i| \geq a$, $\forall i \in [\alpha]$ and $|\theta_i| \leq b$, $\forall i \in [n]$.

**Remarks:** Some explanation is in order regarding the bound in the above theorem. As can be seen, the normalized error is bounded by a product of 4 terms. The first term $\frac{2}{a}$ can be treated as a constant that depends on the minimum score vector corresponding to the f-BTL model. The second term is the condition number of the feature coefficient matrix $B$ and captures how the features interact with each other. The third term depends on the number of pairs seen in $M$. When $|M| = m = \alpha \log(\alpha)$, this term becomes $\sqrt{\log(\alpha)}$. The fourth term grows depending on how many samples one sees as it depends on $L$ which is the Laplacian of the comparison graph. If both $\lambda_n$ and $\lambda_1$ are $O(\log(\alpha))$, then the normalized error is a constant with probability at least $1 - \text{poly}(\frac{1}{n})$. Thus, the result essentially says that if one sees $O(\alpha \log(\alpha))$ samples and $B$ is such that both $\lambda_1$ and $\lambda_n$ are $O(\log(\alpha))$, then the normalized error is bounded by a small constant.

## 5 Lower Bound

In this section, we show that how the achievable $\ell_2$-error rate of the fBTL-LS algorithm (as derived in Theorem 4.1), compares to the minimax $\ell_2$-error rate possible, over the class of feature Bradley-Terry-Luce (f-BTL) model. More specifically, our result in Theorem 5.1 proves an information-theoretic lower bound for the $\ell_2$-error rate achievable by any learning algorithm for estimating the score parameters of the f-BTL model. Our proof technique uses a constructive argument to generate the score vectors $\theta$ from a uniform distribution that respects the f-BTL model in the dynamic range $|\theta_i| \in [a, b], \forall i \in [n]$, and solves the stochastic
inference problem into a multi-way hypothesis testing problem. Our derived lower bound guarantee is given below:

**Theorem 5.1 (Lower Bound for estimating the parameters of f-BTL model).** Let us consider the following set of score vectors \( \Theta_B(a, b) \) of a f-BTL model defined with respect to the coefficient matrix \( B \) and range parameters \( a, b > 0 \) such that:

\[
\Theta_B(a, b) = \{ \theta \in \mathbb{R}^n | \theta \text{ satisfies (3)}, |\theta_i| \leq a \forall i \in [\alpha], |\theta_i| \geq b \forall i \in [n] \}.
\]

Now suppose the learner (i.e. an algorithm which estimates scores of a f-BTL model) is given access to noisy pairwise preferences sampled according to a \( G(n, p) \) Erdős-Rényi random graph with \( p = \frac{\zeta}{n} \) for some \( \zeta > 0 \), such that \( K \) independent noisy pairwise preferences are available for each sampled pair, generated according to some unknown f-BTL model in \( \Theta_B(a, b) \). Then if \( \hat{\theta} \in \mathbb{R}^n \) be the learner’s estimated f-BTL score vector based on the sampled pairwise preferences, upon which environment chooses a worst case true score vector \( \theta \in \Theta_B(a, b) \), then for any such learning algorithm one can show that

\[
\sup_{\theta \in \Theta_B(a, b)} \frac{E[\|\hat{\theta} - \theta\|]}{\|\theta\|} \geq \frac{\sqrt{\lambda_{\min}(B^TB)}}{16b\lambda_{\max}(B^TB)\sqrt{448\zeta Ke^{2b+1}}},
\]

where the expectation is taken over the randomness of the algorithm.

### 6 Experiments

We now describe the experiments we ran with the fBTL-LS algorithm on various synthetic and real world datasets. We compared our results with three algorithms, (i) ordinary least squares (OLS) [Borkar et al.(2016)](Borkar, Karamchandani, and Mirani], (ii) rank centrality (RC) [Negahban et al.(2012)](Negahban, Oh, and Shah] and (iii) inductive pairwise ranking based on inductive matrix completion (IMC) [Niranjan and Rajkumar(2017)]. The first two algorithms do not use any feature information while the third algorithm does. Both the OLS and RC algorithms are guaranteed to work well for the standard BTL model with \( O(n \log(n)) \) pairs compared while the IPR algorithm requires \( O(d^2 \log(n)) \) pairs to be compared for its guarantees to hold good (for a generalized BTL model that they define), \( d \) being the feature dimension.

**Performance Measures:** We measure the performance of algorithms using the following metrics

1. Normalized \( \ell_2 \)-error \( \left( \frac{\|\hat{\theta} - \theta\|}{\|\theta\|} \right) \): For experiments where there is a true score vector that generates the comparisons, we use the normalized \( \ell_2 \) error between the estimated score vector and the true score vector.

2. Pairwise disagreement (pd) error: Suppose \( P^* \in [0, 1]^{n \times n} \) denotes the underlying pairwise preference matrix corresponding to the true (and unknown) score \( \theta \), given by
$P^*_ij = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}$ and $\hat{P} \in [0, 1]^{n \times n}$ be the estimated preference matrix return by the algorithm (note if the algorithm returns a score vector estimate $\hat{\theta}$, we compute $\hat{P}_{ij} = \frac{e^{\hat{\theta}_i}}{e^{\hat{\theta}_i} + e^{\hat{\theta}_j}} \forall i, j \in [n]$), then pd-error essentially counts the fraction of pairs on which $P^*$ and $\hat{P}$ disagree, defined as:

$$pd(\hat{P}, P^*) = \frac{1}{n^2} \sum_{i<j} (\mathbb{I}(\hat{P}_{ij} \geq 0.5 \land P^*_{ij} < 0.5) + \mathbb{I}(\hat{P}_{ij} < 0.5 \land P^*_{ij} > 0.5))$$

3. Sample complexity (sc(ε)): Minimum number of pairwise comparisons required to be observed to obtain normalized $\ell_2$-error ($\frac{\|\hat{\theta} - \theta\|}{\|\theta\|} < \epsilon$).

6.1 Experiments on Synthetic Datasets

We evaluate the four algorithms on three different type of graphs with respect to the above performance measures. We consider three different settings for this purpose — Type-I plots: with increasing node size (n), Type-II plots: with increasing sampling rate (p) but fixed node size (n) and independence number (α) Type-III plots: with increasing independence number (α) and fixed node size (n). The details of the experimental setups and results are provided below.

**Type of graphs.** We use three different type of graphs for synthetic experiments: (1) r-disconnected cliques: Union of r-cliques (2) d-regular graphs: Graphs with each node having degree d and (3) k-ary trees: Trees with every node having k-children (except the leaf nodes).

**Data generation** For each of the above type of graphs $G$, we first fix a maximum independent set $I(G)$ of $G$, and embed the $i^{th}$ node of $I(G)$ with $i^{th}$ canonical basis vector of $\mathbb{R}^{\alpha}$, denoted by $e_i \in \{0, 1\}^{\alpha}$ with $e_i(j) = 1$, if $i = j$ and 0 otherwise $\forall i, j \in [\alpha]$, $\alpha = |I(G)|$.

We next generate a random coefficient matrix $B \in \mathbb{R}^{n \times \alpha}$ and embed rest of the nodes in $[n] \setminus I(G)$ according to (2) as defined in the problem setting. We now choose a random vector $w \in \mathbb{R}^{\alpha}$ and assign a BTL score $\theta_i = w^Tu_i$ to every node $i \in [n]$ as defined in (3). Finally $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is normalized to $\ell_2$-norm 1, i.e. $\|\theta\|_2 = 1$, setting $\theta_i = \frac{\theta_i}{\|\theta\|_2}, \forall i \in [n]$.

**Parameter setting.** As clear from the above data generation procedure, the feature dimension $d$ is equal to the independence number $\alpha = |I(G)|$ of $G$ in all the experiments. We also fix $K = 1000$ throughout for all the experiments (unless performance is reported against $K$), and report the average performances over 50 runs.
In this setup, we compare the four algorithms, with respect to the above three performance measures with varying node size \( n \), on three different graphs: (1) Union of 10 disconnected cliques on \( n \) nodes, (2) \( d \)-Regular graph of \( n \) nodes with fixed degree \( d = 10 \) and (3) Full binary tree of \( n \) nodes. The results are reported in Figure 1. They clearly reflect the superior performance of \( fBTL-LS \) for each of the three performance measures.

Note that for graph type (1) of 10-disconnected cliques, the independence number \( \alpha = 10 \) is fixed for all \( n \), unlike graph (2) and (3) where \( \alpha \) scales with \( n \). Now the interesting observation is that the sample complexity \( \text{sc}(0.5) \) of \( fBTL-LS \) for achieving a target error \( \epsilon = 0.5 \) for 10-disconnected clique is almost constant even with \( n \) varying from 20 to 500, unlike the rest of the three algorithms where it scales with \( n \). This indeed justifies our claim of the required sample complexity to be \( O(\alpha \log \alpha) \), as remarked in Theorem 4.1. This also justifies why for 10-regular graph and full binary tree, sample complexity of \( fBTL-LS \) monotonically increases with \( n \), as the independence number \( \alpha \) itself scales with \( n \) for these two graphs.

**Type-II plots: with increasing sampling rate \( p \) but fixed node size \( n \) and independence number \( \alpha \)**

In this case we compare the four algorithms with varying sampling rate \( p \) with respect to the two estimation error metrics, normalized \( \ell_2 \)-error and pairwise disagreement \( \text{pd}(\hat{P}, P^*) \),
on the following three different graphs: (1) Union of 100 disconnected cliques on 500 nodes, i.e. with each clique having 5 nodes, (2) 50-Regular graphs on 500 nodes with each node of degree $d = 50$ and (3) Full binary tree of height 8 (511 nodes). Thus in each case, $n$ and $\alpha$ are kept fixed, with $p$ to be set as $p = \frac{C\alpha \log \alpha}{(n^2)}$, $C$ varying from 0.5 to 32. From the results in Figure 2, it clearly reflects that, as expected, the performance of all the algorithms get improved with higher sampling rate $p$. However, the rate of performance improvement is far more drastic for $fBTL$ compared to the rest due to its inherent ability to exploit the feature correlation, and thus hence reaching to more accurate score estimates faster.

Type-III plots: with increasing independence number ($\alpha$) and fixed node size ($n$)

Finally in the third setup, we compare the four algorithms with varying independence set size (or independence number) $\alpha$ for a fixed set of $n = 500$ nodes on the following two graphs: (1) Union of $r$-disconnected cliques over 500 nodes with varying $r$ and (2) $d$-Regular graph of 500 nodes with varying degree $d$ (Figure 3). The results again show that varying $p$ as $p = \frac{10(\alpha \log \alpha)}{(n^2)}$, the performance metrics, normalized $\ell_2$-error and $pd(\hat{P}, P^*)$, remains almost constant validating the claim of the required sample complexity of $fBTL$ to be $O(\alpha \log \alpha)$, as follows from Theorem 4.1. The sample complexity curves on the other hand validate the dependency of $sc(0.5)$ on $\alpha$, which clearly increases with higher values of $\alpha$, as expected.

6.2 Experiments on Read-World Datasets

We also evaluate the four algorithms on two real-world preference learning datasets: car and sushi. The datasets and the experimental setup is described below:
Figure 3: Performance vs $\alpha$ on disconnected cliques and $d$-regular graph ($n = 500$ nodes in each)

Figure 4: Pairwise disagreement error $pd(\hat{P}, P^*)$ vs sampling rate $\left( p = \frac{C\alpha \log \alpha}{n^2} \right)$, and number of repeated samples ($K$) on Car and Sushi

**Car Dataset.** ([Abbasnejad et al.(2013)Abbasnejad, Sanner, Bonilla, Poupart, et al.]) This dataset contains pairwise preferences of 20 cars given by 60 users, where each car represented by a 6-dimensional feature vector.

**Sushi Dataset.** ([Kamishima and Akaho(2009)]) This dataset contains over 100 sushis rated according to their preferences, where each sushi is represented by a 7-dimensional feature vector.

**Setup.** Note that the real world datasets does not satisfy any preference modelling assumption, e.g. BTL assumption, and hence there is no true score vector $\theta$ associated to the item preferences. From the user preferences, we first compute the underlying pairwise preference matrix $P^*$, where $P^*_{ij}$ is computed by taking the empirical average of number of times an item $i$ is preferred over item $j$. Further to construct the feature matrix $U$, we use the provided feature information of the item set, that is provided in each dataset. Specifically, if each item is represented by $d$-dimensional feature vector (as described before, $d = 6$ for Car and $d = 7$ for Sushi), we find a set of $d$ items whose corresponding features are linearly independent that forms a basis of $\mathbb{R}^d$ and use these $d$ items as the independent set $\mathcal{I}$. The coefficient matrix $B$ is then constructed by representing the rest of the items as a linear combination of $\mathcal{I}$, such that it satisfies (2). (see Section 2.1 for details)
Performance Measure. As noted above, the real world datasets does not satisfy the BTL assumption, so there is no true score vector $\theta$ associated to the item preferences. We however measure the performances of the algorithms with respect to the true preference matrix $P^*$, using pairwise disagreement error $pd(P, P^*)$. In both the cases, our proposed algorithm outperforms the rest. We also evaluate the algorithms with increasing number of repeated samples per pair ($K$). As expected, it shows that higher $K$ indeed leads to improved performance. The results are reported in Figure 4.

7 Conclusion

We considered the problem of ranking from pairwise comparisons in presence of feature information. The existing results either fail to utilize this feature information, or make broad low rank assumptions which cannot capture the item dependencies through their corresponding feature representations. In this work, we introduce a feature based probabilistic preference model, $fBTL-LS$ and have shown that the feature information could be used to obtain tight sample complexity bounds for recovering ‘good estimates’ of the underlying scores of the preference model. We have proposed a least squares based algorithm and have shown theoretical recovery guarantees for the same. Furthermore, our information theoretic based lower bound analysis show the optimality of our proposed algorithm with a matching lower bound guarantee.

While least square based algorithms are a natural choice for this problem, it would be interesting to see how Markov chain based approaches, such as rank centrality can be extended to accommodate feature information. It would also be interesting to analyze the problem in contextual setting introducing feature dependencies of the users as well.
References

[Abbasnejad et al.(2013)Abbasnejad, Sanner, Bonilla, Poupart, et al.] Ehsan Abbasnejad, Scott Sanner, Edwin V Bonilla, Pascal Poupart, et al. Learning community-based preferences via dirichlet process mixtures of gaussian processes. In Proceedings of the 18th International Joint Conferences on Artificial Intelligence, 2013.

[Ailon et al.(2008)Ailon, Charikar, and Newman] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. Journal of the ACM, 55(5), 2008.

[Borkar et al.(2016)Borkar, Karamchandani, and Mirani] Vivek S Borkar, Nikhil Karamchandani, and Sharad Mirani. Randomized kaczmarz for rank aggregation from pairwise comparisons. In Information Theory Workshop (ITW), 2016 IEEE. IEEE, 2016.

[Bradley and Terry(1952)] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4), 1952.

[Braverman and Mossel(2008)] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2008.

[Busa-Fekete et al.(2014)Busa-Fekete, Hüllermeier, and Szörényi] Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of mallows. In Proceedings of the 31st International Conference on Machine Learning, volume 32, 2014.

[Chen and Joachims(2016)] Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In Proceedings of the 9th ACM International Conference on Web Search and Data Mining. ACM, 2016.

[Gleich and Lim(2011)] David F Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. ACM, 2011.

[Hall(1935)] Philip Hall. On representatives of subsets. Journal of the London Mathematical Society, 1(1), 1935.

[Jamieson and Nowak(2011)] Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. In Advances in Neural Information Processing Systems, 2011.

[Kamishima and Akaho(2009)] Toshihiro Kamishima and Shotaro Akaho. Efficient clustering for orders. In Mining Complex Data. Springer, 2009.

[Luce(1959)] R Duncan Luce. Individual Choice Behavior: A Theoretical Analysis. Wiley, 1959.
[Negahban et al.(2012)Negahban, Oh, and Shah] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, 2012.

[Niranjan and Rajkumar(2017)] UN Niranjan and Arun Rajkumar. Inductive pairwise ranking: Going beyond the n log (n) barrier. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

[Rajkumar and Agarwal(2014)] Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of 31st International Conference on Machine Learning*, 2014.

[Rajkumar and Agarwal(2016)] Arun Rajkumar and Shivani Agarwal. When can we rank well from comparisons of o (n \log (n)) non-actively chosen pairs? In *Proceedings of the 29th Conference on Learning Theory*, 2016.

[Saaty(2008)] Thomas L Saaty. Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 2008.

[Shah et al.(2016)Shah, Balakrishnan, Guntuboyina, and Wainwright] Nihar Shah, Sivaraman Balakrishnan, Aditya Guntuboyina, and Martin Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *Proceedings of 33rd International Conference on Machine Learning*, 2016.

[Shah and Wainwright(2015)] Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.

[Thurstone(1927)] Louis L. Thurstone. A law of comparative judgment. In *Psychological Review*, volume 34:4, 1927.

[Wauthier et al.(2013)Wauthier, Jordan, and Jojic] Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[Yue et al.(2012)Yue, Broder, Kleinberg, and Joachims] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5), 2012.
A Proof of Theorem 3.2

Theorem 3.2 (Bound On Error Probability). Given a relation graph \(G\), feature matrix \(U\), a set of pairs \(M\) where \(|M| = m\) generated according to the sampling model above (where each pair is chosen with probability \(p\)), and the exact preference probabilities \(P_{ij} \forall (i, j) \in M\), the probability that the score vector \(\theta\) is same as that estimated score vector \(\hat{\theta}\) that is got by solving the equations obtained is bounded by

\[
P(\hat{\theta} \neq \theta) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G)+1)\}} \sum_{I \subseteq \mathcal{I}(G)||I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{(c_I-(q-1))},
\]

Proof. Note from Theorem 3.1 we have that one only fails to recover the true \(\theta\) if and only if the edge set \(\Delta_M\) of the bipartite graph \(C_M\) fails to cover \(A\). Thus we have

\[
P(\theta \neq \hat{\theta}) = P(\{A \text{ is not covered by } C_M\})
= P(\exists S' \subseteq A \text{ s.t. } |N_{CM}(S')| < |S'|) \quad \text{(by Theorem 2.1)}
\]

We use \(N_G(i)\) to denote the set of neighbours of node \(i \in [n]\) in a graph \(G\) and \(N_G(i)\) to denote the set of neighbours of node \(i \in [n]\) in \(G\) including \(i\) itself, i.e. \(\bar{N}_G(i) = N_G(i) \cup \{i\}\).

Define \(N_G(S) = \cup_{i \in S} N_G(i)\), \(\forall S \subseteq V(G)\) and \(\bar{N}_G(ij) = (\bar{N}_G(i) \cap \bar{N}_G(j)) \cap \mathcal{I}(G)\). Thus we can associate every node \(k \in \mathcal{I}(G) = [\alpha(G)]\) in the independent set to a set of edges \(M_k\) such that \((i, j) \in M_k \iff k \in \bar{N}_G(ij)\). Let us also denote \(n_k = |M_k|\) and let \(n_{\min} = \min_{k \in [\alpha(G)]} n_k\). More generally we denote \(n_I = |\cap_{i \in I} M_i|, \forall I \subseteq [\alpha(G)]\).

We will also find it convenient to define \(c_I = |\cup_{i \in I} M_i|\) and \(d_I = |\cap_{i \in I} M_i|, \forall I \subseteq [\alpha(G)]\).

Clearly when \(|I| = 1\), say \(I = \{i\}, i \in [n]\), \(c_I = d_I = n_i\). In general, for \(|I| = q, 1 \leq q \leq \alpha(G)\) we have \(c_I = \sum_{k=1}^{q} \sum_{I \subseteq |I|=k} (-1)^{q-k} d_I\), where the size of the intersecting sets \(d_I\) depends on specific the structure of the graph \(G\) (see Theorem 3.3 for graph specific analysis).

Now if we denote the event \(F_i = \{\exists S' \subseteq A \text{ s.t. } |S'| = i \text{ and } S' \text{ is not covered by } C_M\}, \forall i \in [\alpha(G)]\), and recalling \(A = [\alpha(G)]\), we further get

\[
P(\theta \neq \hat{\theta}) = P(\exists S' \subseteq A \text{ s.t. } |N_{CM}(S')| < |S'|) = P(F_1 \cup F_2 \cup F_3 \ldots F_{\alpha(G)})
= P(F_1 \cup (F_2 \cap F_{\alpha(G)}) \cup (F_3 \cap F_{\alpha(G)}^c) \cup \ldots \cup (F_{\alpha(G)} \cap F_{\alpha(G)}^c))
= P(F_1) + P(F_2 \cap F_{\alpha(G)}^c) + \ldots + P(F_{\alpha(G)} \cap F_{\alpha(G)}^c) \quad (7)
\]

Assuming the pairwise node preferences are drawn according to the edges sampled from an Erdős-Rényi random graph \(G(n, p)\) and applying Theorem 3.1 on the event \(F_1\), it is easy to see that

\[
P(F_1) = P(\exists S' \subseteq A \text{ s.t. } |N_{CM}(S')| < |S'| = 1) = P\left(\exists S' = \{k\}, k \in [\alpha(G)] \text{ s.t. no edge from } M_k \text{ is sampled in } G(n, p)\right) \leq \sum_{i=1}^{\alpha(G)} (1-p)^{n_i},
\]

where

\[
\bar{N}_G(ij) = (\bar{N}_G(i) \cap \bar{N}_G(j)) \cap \mathcal{I}(G).
\]
where the last inequality follows taking union bound over all singletons in \( A = [\alpha(G)] \).

Note that one can further bound above as \( P(F_1) \leq \alpha(G) \exp(-pn_{\text{min}}) \). In general, for any \( 1 \leq q \leq \alpha(G) \), one can similarly derive

\[
P(F_q \cap F_{q-1}) = P(\{\exists S' \subseteq A, |S'| = q, S' \text{ is not covered by } C_M \text{ and } \forall S'_1 \subseteq A, |S'_1| < q, S'_1 \text{ is covered by } C_M\})
\leq \sum_{I \subseteq \mathcal{I}(G)||I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{c_I-q}, \tag{8}
\]

where the last inequality follows from the crucial observation that for any \( S' \subseteq A, |S'| = q \) if \( S' \) is not covered by \( C_M \) but all it subsets \( S'_1 \subset S' \) are, then \( G(n,p) \) must have sampled exactly \( q-1 \) edges from \( \cap_{i \in S'} M_i \) and none from \( \bigcup_{i \in I} M_i \setminus \cap_{i \in I} M_i \). Using (8) in (7) we finally get,

\[
P(\theta \neq \hat{\theta}) \leq P(F_1) + P(F_2 \cap F_1^c) + \ldots + P(F_{\alpha(G)} \cap F_{\alpha(G)-1}^c)
= \sum_{q=1}^{\alpha(G)} \sum_{I \subseteq \mathcal{I}(G)||I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{c_I-(q-1)},
\]

where we assume \( \binom{x}{y} = 0 \), if \( x < y \). Further note that if \( d_{\text{max}}(G) < \alpha(G) \), then for any \( I \subseteq [\alpha(G)] \) such that \( |I| > (d_{\text{max}} + 1) \), we have \( d_I = 0 \), using which we further get

\[
P(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G)+1)\}} \sum_{I \subseteq \mathcal{I}(G)||I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{c_I-(q-1)}
\]

Thus the claim follows. \( \Box \)

## B Proof of Theorem 3.3

**Theorem 3.3 (Sample Complexity for Common Graphs).** Under the settings of Theorem 3.2, the following sample complexity bounds hold with probability at least \( 1 - \delta \)

- If \( G \) is a disconnected graph, star graph, or a cycle, \( m = O(n \log(\frac{\alpha}{\delta})) \)
- If \( G \) is a clique, \( m = O(\log(\frac{1}{\delta})) \)
- If \( G \) is a r-disconnected clique (i.e. union of r cliques), \( m = O(r \log(\frac{1}{\delta})) \)

**Proof.** We will now analyse Theorem 3.2 for certain specific class of graphs. We will be using the same notations used in proof of Theorem 3.2 for the purpose.

1. **Fully Disconnected Graph:** Note that in this case \( \alpha(G) = n \). Also note that \( \forall k \in [n], M_k = \{(k,i) \mid i \in [n] \setminus \{k\}\} \). Thus \( n_k = n - 1 \). Moreover \( \forall \emptyset \subseteq I \subseteq [n], |I| = 2, c_I = 2n - 3, d_I = 1, \) and if \( |I| \geq 3, d_I = 0 \).
Now applying Theorem 3.2 and noting $d_{\text{max}}(G) = 0$, we further get that,

$$P(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G) + 1)\}} \sum_{I|\lvert I\rvert = q} \binom{d_I}{q-1} p^{q-1}(1-p)^{ci-q}$$

$$= \sum_{i=1}^{n} (1-p)^{n-1} + \sum_{i<j} p(1-p)^{2n-3-1}$$

$$= n(1-p)^{n-1} + \binom{n}{2} p(1-p)^{2n-4}$$

$$\leq n(e-p)^{n-1} + \frac{n(n-1)}{2} p(e-p)^{2n-4}$$

$$\leq n^2(e-p(n-1))$$

$$\leq \delta,$$

solving which we get $p \geq \frac{1}{(n-1)} \log \left( \frac{n^2}{\delta} \right)$. Thus the expected number of edges (pairwise preferences) in the random graph required is at least $p(\binom{n}{2}) \geq \frac{n}{2} \log \left( \frac{n}{\delta} \right)$, which recovers the result for the usual BTL model.

2. **Complete Graph:** In this case $\alpha(G) = 1$. Without loss of generality assuming $\mathcal{I}(G) = \{1\}$, thus we have $M_1 = \{(i, j) \mid i, j \in [n]\}$. Thus $n_1 = \binom{n}{2}$. Moreover $\forall I \subseteq [n], |I| \geq 2, d_I = 0$.

Applying Theorem 3.2 as before and noting $d_{\text{max}}(G) = n$, we further get,

$$P(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G) + 1)\}} \sum_{I|\lvert I\rvert = q} \binom{d_I}{q-1} p^{q-1}(1-p)^{ci-q}$$

$$= (1-p)^{\binom{n}{2}}$$

$$= (e-p)^{\binom{n}{2}}$$

$$\leq \delta,$$

solving which one gets $p \geq \frac{1}{(\binom{n}{2})} \log \left( \frac{1}{\delta} \right)$. Thus the expected number of edges (pairwise preferences) in the random graph required is at least $p(\binom{n}{2}) \geq \frac{1}{2} \log \left( \frac{1}{\delta} \right)$, which is intuitive as well since in a complete graph one needs the knowledge of only $\Omega(1)$ pairwise preferences to recover the exact ranking (i.e. $\theta$) with high probability $(1 - \delta)$.

3. **r-Disconnected Cliques:** Say $G$ has exactly $r \in [n]$ disconnected cliques, $G_1, G_2, \ldots G_r$, each with $d \in [n]$ edges (i.e. for each $k \in [r], |E(G_k)| = d$), assuming $n = rd$. Thus
in this case \( \alpha(G) = r \). Without loss of generality assume \( \mathcal{I}(G) = \{1, 2, \ldots, r\} \). Then \( \forall k \in [r] \), we have \( M_k = \{(i, j) \mid (i, j) \in E(G_k)\} \cup \{(k, j) \mid j \in [n] \setminus \{k\}\} \). Thus \( n_k = \binom{d}{2} + (r-1) \). Moreover \( \forall I \subseteq [n], |I| = 2, c_I = \binom{d}{2}((d-1) + (r-1)) - 1 = d(d-1) + (r-2), d_I = 1 \) and \(|I| \geq 3, d_I = 0\).

Then applying Theorem 3.2 as above and noting \( d_{\max}(G) \leq \left\lceil \frac{n}{r} \right\rceil \), we further get,

\[
\Pr(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{r} \left( \sum_{|I|=q} \left( \sum_{j=1}^{r} \binom{d}{2} + r-1 \right) \right) p^{q-1} (1-p)^{c_I - q}
\]

\[
= \sum_{i=1}^{r} (1-p)^{\binom{d}{2} + r-1} + \sum_{i<j,j \in [r]} p(1-p)^{d(d-1) + (r-2)-1}
\]

\[
= r(1-p)^{\binom{d}{2} + r-1} + \binom{r}{2} p(1-p)^{d(d-1) + (r-3)}
\]

\[
\leq r(e-p)^{\binom{d}{2} + r-1} + \frac{r(r-1)}{2} p(e-p)^{d(d-1) + (r-3)}
\]

\[
\leq r(r-1)(e-p)^{\binom{d}{2} + r-1})
\]

\[
\leq r^2(e-p)^{\binom{d}{2} + r-1}) \leq \delta,
\]

solving which one can derive \( p \geq \frac{1}{(d+1)(r-1)} \log \left( \frac{r^2}{\delta} \right) \). Thus the expected number of edges (pairwise preferences) in the random graph required is at least \( p(n) = \frac{n(n-1)/2}{d(d-1)/2 + r-1} \log \left( \frac{r^2}{\delta} \right) \geq \frac{n(n-1)^2}{n(n-r)+2r^2} \log \left( \frac{r^2}{\delta} \right) \geq r \log \left( \frac{r^2}{\delta} \right) \), where the last inequality follows assuming \( r < \frac{n}{\sqrt{2}} \).

Note that setting \( d = 1 \) and \( d = n \), one can recover the earlier bounds we derived for disconnected and complete graphs respectively.

4. **Star:** Note that in this case the size of the maximal independent set \( \alpha(G) = (n-1) \). Without loss of generality assume \( \mathcal{I}(G) = [n] \setminus \{1\} \). Thus we have that for any \( k \in \mathcal{I}(G) \), \( E_k = \{(k, j) \mid j \in [n] \setminus \{k\}\} \cup \{(1, j) \mid j \in [n] \setminus \{1\}\} \). Thus \( n_k = (n-1) + (n-2) = 2n-3 \).

Moreover \( \forall I \subseteq [n], |I| = 2, d_I = (n-2) + 1 = n-1 \) and \( c_I = 2(2n-3) - (n-1) = 3n-5 \).

For \( |I| \geq 3, d_I = n-2 \) and \( c_I = (2n-3)(|I| - (n-1)(|I|)) + (n-2)((|I|) - \ldots + (-1)^{|I|-1}(n-2), e.g. when |I| = 3, c_I = 4n-14 \text{ etc.}

Applying Theorem 3.2 as before and noting \( d_{\max}(G) = (n-1) \), we further get,
\[
P(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G)+1)\}} \sum_{I\mid |I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{c_I-q}
\]

\[
= (n-1) \left((1-p)^{2n-3} + \frac{n(n-2)}{2} p(1-p)^{2n-4}\right)
\]
\[
+ \left(\frac{n-1}{3}\right) \left(\frac{n-2}{2}\right) p^2(1-p)^{3n-12} + \ldots
\]
\[
\leq n^2 (e^{-p(n-1)})
\]
\[
\leq \delta.
\]

Similar to the case of fully disconnected graph, solving \(p\) from above one can get that the expected number of edges (pairwise preferences) in the random graph required is atleast \(p\left(\frac{n}{2}\right) = \left(\frac{n}{2} \log \left(\frac{n}{\delta}\right)\right)\).

5. Cycle: We will assume that \(n = 2n' \geq 4\) is even, similar analysis can be done for the odd number of nodes as well. Thus in this case \(\alpha(G) = n'\). Without loss of generality assume \(\mathcal{I}(G) = \{2i \in [n] \mid i \in [n]\}\). Thus we have that for any \(k \in \mathcal{I}(G), E_k = \{(k, j) \mid j \in [n] \setminus \{k\}\} \cup \{(k-1, j) \mid j \in [n] \setminus \{k-1\}\} \cup \{(k+1) \mod k, j \mid j \in [n] \setminus \{(k+1) \mod k\}\}\). Thus \(n_k = (n-1) + (n-2) + (n-3) = 3(n-2)\). Moreover \(\forall I \subseteq [n], |I| = 2, d_I = (n-2) + 1 = n-1\) and \(c_I = 2(3n-6) - (n-1) = 2n - 5\). For \(|I| \geq 3, d_I = 0\).

Further applying Theorem (3.2) and noting \(d_{\text{max}}(G) = 2\), we further get,

\[
P(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G)+1)\}} \sum_{I\mid |I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{c_I-q}
\]

\[
= \sum_{i=1}^{n'} (1-p)^{3n-6} + \sum_{I\subseteq \mathcal{I}(G), |I|=2} (n-1)p(1-p)^{2n-5-(n-1)}
\]
\[
= \frac{n}{2} (1-p)^{3n-6} + \frac{(n-1)(n-2)}{8} p(1-p)^{n-4}
\]
\[
\leq \frac{n}{2} (1-p)^{3n-6} + \frac{n-2}{4} (1-p)^{n-4} \quad \text{(as } p\left(\frac{n}{2}\right) \geq 1\}
\]
\[
\leq n(1-p)^{n-4}
\]
\[
\leq \delta,
\]

solving which one can derive \(p = f(\delta) \geq \frac{1}{n-4} \log \left(\frac{n}{\delta}\right)\). Thus the expected number of edges (pairwise preferences) in the random graph required is atleast \(p\left(\frac{n}{2}\right) = \frac{n(n-1)/2}{n-4} \log \left(\frac{n}{\delta}\right) \geq \frac{n}{2} \log \left(\frac{n}{\delta}\right)\).
6. **K-ary Tree**: Let $h$ be the height of the tree and 1 denotes the root node. For any node $i \in [n]$, $\text{par}(i)$ and $\text{ch}(i)$ respectively denotes the parent and child nodes $i$. We will consider only trees of even height for the purpose, it is easy to derive a similar analysis for trees of odd height. Note that $n = (1 + K + K^2 + \ldots + K^h) = \frac{K^{h+1} - 1}{K - 1}$. Clearly the maximum independent set contains all the nodes which which are at a even length distance from the root, including the root itself. Thus $\alpha(G) = (1 + K^2 + K^4 + \ldots + K^h) = \frac{K^{h+2} - 1}{K^2 - 1}$.

Note that for any $k \in \mathcal{I}(G)$, $N_G(k) \cap \mathcal{I}(G)^c = \{\text{par}(k) \cup \text{ch}(k)\}$. Also every node in $[n] \setminus \mathcal{I}(G)$ form $C = (K + K^3 + K^5 + \ldots + K^{h-1}) = \frac{K^{h+2} - 1}{K^2 - 1}$ clusters, we denote them by $H_1, H_2, \ldots, H_C$, such that for any $i \in [n] \setminus \mathcal{I}(G)$, $H_i = \{j \in \mathcal{I}(G) \mid j \in \text{par}(i) \cup \text{ch}(i)\}$. Thus $|H_i| = K + 1$. We will also abbreviate $N_G(\cdot)$ as $N(\cdot)$ for ease of notations.

Thus for any $k \in \mathcal{I}(G)$, $E_k = \{(k, j) \mid j \in [n] \setminus \{k\}\} \cup_{k' \in \text{par}(k) \cup \text{ch}(k)} \{(k', j) \mid j \in [n] \setminus \{k\}\}$. This gives that $n_k = (n - 1) + \sum_{i=2}^{K+2} (n - i) = (k + 2)\frac{2n - k - 3}{2}$. Moreover, for any $I \subseteq \mathcal{I}(G), |I| = 2$,

\[
d_I = \begin{cases} 1 + (n - 2), & \forall i, j \in I, |N(i) \cap N(j)| = 1, \\ 1, & \text{otherwise,} \end{cases}
\]

\[
c_I = \begin{cases} (K + 1)(2n - K - 2), & \forall i, j \in I, |N(i) \cap N(j)| = 1, \\ (K + 2)(2n - k - 3) - 1, & \text{otherwise,} \end{cases}
\]

for any $I \subseteq \mathcal{I}(G), 3 \geq |I| \leq k + 1$,

\[
d_I = \begin{cases} (n - 2), & \forall i, j \in I, |N(i) \cap N(j)| = 1, \\ 0, & \text{otherwise,} \end{cases}
\]

\[
c_I = \begin{cases} |I|(K + 2)\frac{2n - K - 3}{2} - \left(\frac{|I|}{2}\right)(n - 1) \ldots (-1)^{|I| - 1}(n - 2), & \forall i, j \in I, |N(i) \cap N(j)| = 1, \\ |I|(K + 2)\frac{2n - K - 3}{2} - \left(\frac{|I|}{2}\right)1, & \text{otherwise,} \end{cases}
\]

and for any $I \subseteq \mathcal{I}(G), |I| > K + 1, d_I = 0$. Now applying Theorem 3.2 as before and noting $d_{\text{max}}(G) = K$ we further get,
$$\mathbf{P}(\theta \neq \hat{\theta}) \leq \sum_{q=1}^{\min\{\alpha(G), (d_{\text{max}}(G)+1)\}} \sum_{I:|I|=q} \left( \frac{d_I}{q-1} \right) p^{q-1}(1-p)^{d_I-q}$$

$$= \alpha(G)(1-p)^{(K+2)/2(2n-K-3)} +$$

$$C \left( \binom{K+1}{2} (n-1)p(1-p) \right) (K+1)(2n-K-2)-(n-1)$$

$$\left( \binom{n}{2} - C \left( \binom{K+1}{2} \right) \right) p(1-p)^{(K+2)/2(2n-K-3)-1}$$

$$\sum_{K'=3}^{K+1} C \left( \binom{K+1}{3} (n-2)p^2(1-p)^{K+1} \right) -$$

$$\leq \delta.$$
On the other hand, assuming \( v \in \mathbb{R}^a \) to be such that \( v_i = \theta_i \), \( \forall i \in [a] \) and \( y \in \mathbb{R}^m \) be such that \( y_{ij} = \log \left( \frac{P_{ij}}{P_{ji}} \right) \), we have \( v = \arg \min_{x \in \mathbb{R}^m} \|Q^T x - y\|^2 \) which gives

\[
\tilde{Q} y = \tilde{L} v. \tag{10}
\]

The above optimality condition holds as for any \( i, j \in [n] \), \( y_{ij} = \theta_i - \theta_j \), and so \( y = L^T \theta = L^T B v = \tilde{Q}^T v \), where the second equality holds due to (3). Thus combining (9) and (10), we get

\[
\tilde{Q} (y - \hat{y}) = \tilde{L} (v - \hat{v})
\]

which further gives,

\[
(y - \hat{y})^T \tilde{Q}^T \tilde{Q}(y - \hat{y}) = \|\tilde{Q} (y - \hat{y})\|^2 = \|\tilde{L} (v - \hat{v})\|^2 = (v - \hat{v})^T \tilde{L} \tilde{L}^T (v - \hat{v}),
\]

from which we get

\[
\lambda_{\text{min}}(\tilde{L} \tilde{L}^T) \|v - \hat{v}\|^2 \leq \|\tilde{L} (v - \hat{v})\|^2 = \|\tilde{Q} (y - \hat{y})\|^2 \leq \lambda_{\text{max}}(\tilde{Q}^T \tilde{Q}) \|y - \hat{y}\|^2 \tag{11}
\]

where \( \lambda_{\text{min}}(\tilde{L} \tilde{L}^T) \) is the smallest non-zero eigenvalue of the positive semi-definite matrix \( (\tilde{L} \tilde{L}^T) \) and \( \lambda_{\text{max}}(\tilde{Q}^T \tilde{Q}) \) being the largest eigenvalue of \( (\tilde{Q}^T \tilde{Q}) \). Now from standard results on matrix eigenvalues, we know that the set of non-zero eigenvalues of \( \tilde{Q}^T \tilde{Q} \) and \( \tilde{Q} \tilde{Q}^T \) are exactly same, which implies \( \lambda_{\text{max}}(\tilde{Q}^T \tilde{Q}) = \lambda_{\text{max}}(\tilde{Q} \tilde{Q}^T) = \lambda_n \). Moreover, \( \lambda_{\text{min}}(\tilde{L} \tilde{L}^T) = (\lambda_{\text{min}}(\tilde{L}))^2 = (\lambda_{\text{min}}(\tilde{Q} \tilde{Q}^T))^2 = \lambda_n^2 \). Thus from Equation 11, we get

\[
\|v - \hat{v}\| \leq \frac{\|y - \hat{y}\| \sqrt{\lambda_n}}{\lambda_1}. \tag{12}
\]

Now in order to bound \( \|y - \hat{y}\| = \sqrt{\sum_{(i,j) \in E} (y_{ij} - \hat{y}_{ij})^2} \), first recall from the definition of \( y_{ij} \) that \( y_{ij} = \log \left( \frac{P_{ij}}{P_{ji}} \right) = \log P_{ij} - \log P_{ji} \), for any edge \( (i, j) \in M \). Similarly we have \( \hat{y}_{ij} = \log \hat{P}_{ij} - \log \hat{P}_{ji} \). Thus we have,

\[
|y_{ij} - \hat{y}_{ij}| = |(\log P_{ij} - \log \hat{P}_{ij}) - (\log P_{ji} - \log \hat{P}_{ji})| \\
\leq |(\log P_{ij} - \log \hat{P}_{ij})| + |(\log P_{ji} - \log \hat{P}_{ji})| \tag{13}
\]

Let us denote \( \nu_{ij} = |P_{ij} - \hat{P}_{ij}| \). Clearly \( |P_{ij} - \hat{P}_{ij}| = \nu_{ij} \) since \( P_{ij} + P_{ji} = \hat{P}_{ij} + \hat{P}_{ji} = 1 \). Note that the random variable \( \hat{P}_{ij} \) is the average of \( K \) samples from \( \text{Bernoulli}(P_{ij}) \), applying \text{Hoeffding’s Inequality} we get

\[
\mathbb{P} \left( \nu_{ij} \geq \eta \right) = \mathbb{P} \left( |P_{ij} - \hat{P}_{ij}| \geq \eta \right) \leq 2e^{-2\eta^2 K} \tag{14}
\]

Now since \( |\theta_i| \leq b, \forall i \in [n] \), we have \( \frac{1}{1+e^{2b}} \leq P_{ij} \leq \frac{e^{2b}}{1+e^{2b}}, \forall i, j \in [n] \). Also as \( K \geq 6(1 + e^{2b})^2 \log n \), using (14) we further have

\[
\mathbb{P} \left( \nu_{ij} \geq \frac{P_{ij}}{2} \right) \leq \mathbb{P} \left( \nu_{ij} \geq \frac{1}{2(1+e^{2b})} \right) \leq \frac{2}{n^3}, \forall i, j \in [n] \tag{15}
\]
Above thus implies that \( \nu_{ij} = |P_{ij} - \hat{P}_{ij}| < \frac{P_{ij}}{2} \) with high probability of at least \( 1 - \frac{2}{n^3} \), for \( K = 6 \log n(1 + e^{2b})^2 \). Further since \( \nu_{ij} = \nu_{ji} \), using union bound over all pairs in \( M \), we get that (15) holds true for all pairs \( (i,j) \in [n] \) with probability at least \( 1 - \frac{2m}{n^3} \), i.e.

\[
P\left( \forall i,j \in [n], \nu_{ij} < \frac{P_{ij}}{2} \right) > 1 - \frac{2m}{n^3}.
\]

Define \( g : [0,1] \mapsto \mathbb{R} \), such that \( g(p) = \log(p) \), \( \forall p \in [0,1] \). Using Taylor’s theorem, one can obtain a \( p^* \in [P_{ij} - \nu_{ij}, P_{ij} + \nu_{ij}] \) such that

\[
\log(\hat{P}_{ij}) = \log P_{ij} + \frac{1}{p^*} (\hat{P}_{ij} - P_{ij}),
\]

or equivalently,

\[
\log(\hat{P}_{ij}) - \log P_{ij} = \frac{1}{p^*} \leq \frac{2}{P_{ij}},
\]

where the last inequality follows from (15) with probability at least \( 1 - \frac{2m}{n^3} \).

Furthermore, in the high probability event, as \( |\hat{P}_{ij} - P_{ij}| < \frac{P_{ij}}{2} \), Thus we have

\[
|\log(\hat{P}_{ij}) - \log P_{ij}| \leq 1, \forall i,j \in [n].
\]

Combining above with (13) we get

\[
|y_{ij} - \hat{y}_{ij}| \leq 2,
\]

which implies \( \|y - \hat{y}\| \leq 2\sqrt{m} \). Applying above to (12) we thus get

\[
\|v - \hat{v}\| \leq \frac{\|y - \hat{y}\|\sqrt{\lambda_1}}{\lambda_1} \leq \frac{2\sqrt{m\lambda}}{\lambda_1} \tag{16}
\]

with probability at least \( 1 - \frac{1}{n} \). Finally note that since \( |\theta_i| \geq a, \forall i \in [\alpha] \), we have \( \|v\| \geq a\sqrt{\alpha} \).

Moreover, as \( \theta = Bv \), \( \|\theta\| = \|Bv\| \geq \sqrt{\lambda_{\min}(B^TB)}\|v\| \geq a\sqrt{\alpha\lambda_{\min}(B^TB)} \). On the other hand, we have set \( \hat{\theta} = B\hat{v} \) thus,

\[
\|\theta - \hat{\theta}\| = \|B(v - \hat{v})\| \leq \sqrt{\lambda_{\max}(B^TB)}\|v - \hat{v}\|.
\]

Combining above with (16), we finally have

\[
\frac{\|\theta - \hat{\theta}\|}{\|\theta\|} \leq \frac{2\sqrt{m\lambda}\lambda_{\max}(B^TB)}{a\lambda_1\sqrt{\alpha\lambda_{\min}(B^TB)}},
\]

with probability at least \( 1 - \frac{2m}{n^3} \) and the claim follows.

\[\square\]

D Proof of Theorem 5.1

Theorem 5.1 (Lower Bound for estimating the parameters of f-BTL model). Let us consider the following set of score vectors \( \Theta_B(a,b) \) of a f-BTL model defined with respect to the coefficient matrix \( B \) and range parameters \( a,b > 0 \) such that:

\[
\Theta_B(a,b) = \{ \theta \in \mathbb{R}^n \mid \theta \text{ satisfies } (3), |\theta_i| \leq a \forall i \in [\alpha], |\theta_i| \geq b \forall i \in [n] \}.
\]

27
Now suppose the learner (i.e. an algorithm which estimates scores of a f-BTL model) is given access to noisy pairwise preferences sampled according to a $\mathcal{G}(n, p)$ Erdős-Rényi random graph with $p = \frac{\zeta}{n}$ for some $\zeta > 0$, such that $K$ independent noisy pairwise preferences are available for each sampled pair, generated according to some unknown f-BTL model in $\Theta_B(a, b)$. Then if $\hat{\theta} \in \mathbb{R}^n$ be the learner’s estimated f-BTL score vector based on the sampled pairwise preferences, upon which environment chooses a worst case true score vector $\theta \in \Theta_B(a, b)$, then for any such learning algorithm one can show that

$$\sup_{\theta \in \Theta_B(a, b)} \frac{E[\|\hat{\theta} - \theta\|]}{\|\theta\|} \geq \frac{\sqrt{\lambda_{\min}(B^TB)}}{16b\lambda_{\max}(B^TB)\sqrt{448\zeta Ke^{2(b+1)}}},$$

where the expectation is taken over the randomness of the algorithm.

Proof. We solve the above problem reducing it to a multi-class hypothesis testing problem as follows: Consider we are given a set of $N$ score vectors $\{\theta^1, \theta^2, \ldots \theta^N\} \subset \Theta_B(a, b)$ such that $\|\theta^{k_1} - \theta^{k_2}\| \geq \delta$, for any two score vectors $\theta^{k_1}, \theta^{k_2}$ such that $k_1, k_2 \in [N]$. Then given the set of pairwise preferences generated by an unknown score vector $\theta = \theta^L$, where $L$ is a random index selected uniformly from the set $[N]$, the hypothesis testing task is to identify the index of the true score vector $L$.

Now given any algorithm that predicts a score vector $\hat{\theta}$ based on the given set of pairwise preferences from the f-BTL model $\theta^L$, sampled according to a $\mathcal{G}(n, p)$ Erdős-Rényi random graph with $p = \frac{\zeta}{n}$ for some $\zeta > 0$, such that $K$ independent noisy pairwise preferences are available for each sampled pair, one natural way to estimate $L$ is by $\hat{L} = \arg \min_{k \in [N]} \|\hat{\theta} - \theta^k\|$. Note that for $\hat{L}$ to be different that $L$, it has to be the case that $\|\hat{\theta} - \theta\| \geq \frac{\delta}{2}$. Thus one can write

$$E[\|\hat{\theta} - \theta\|] \geq \frac{\delta}{2} P(\hat{L} \neq L)$$

Further applying a similar information theoretic analysis as [Negahban et al.(2012)Negahban, Oh, and Shah], one gets

$$E[\|\hat{\theta} - \theta\|] \geq \frac{\delta}{2} \left[1 - \frac{K\zeta}{2N^2} \sum_{k_1 \in [N]} \sum_{k_2 \in [N]} \|e^{\theta^{k_1}} - e^{\theta^{k_2}}\|^2 + \log 2\right] \log N$$

(17)

Thus the remaining task is to construct a set of $N$ score vectors $\{\theta^1, \theta^2, \ldots \theta^N\} \subset \Theta_B(a, b)$ which are well separated, so to get suitable bounds on the terms $\|e^{\theta^{k_1}} - e^{\theta^{k_2}}\|^2$, $\forall k_1, k_2 \in [N]$ in (17). We use the following construction for the purpose:

**Constructing the set of score vectors.** For any $k \in [N]$, we construct the $k^{th}$ score vector $\theta^k$ set of the set of $N$ random score vectors as follows:

- Draw $\alpha$ many random variables $X_{1}^k, X_2^k, \ldots X_{\alpha}^k \sim Unif\left[\left(\frac{1}{2} - \beta\delta\right), \left(\frac{1}{2} + \beta\delta\right)\right]$, where $\beta$ is a constant to be adjusted later.
- Set $\theta_i^k = a + (b - a)X_i^k$, $\forall i \in [\alpha], 0 < a < b < 1$. 

28
• Consider the coefficient matrix $B \in \mathbb{R}^{n \times \alpha}$ such that $\sum_{j=1}^{\alpha} B_{ij} = 1$, $\forall i \in [n]$.

• Set the remaining score vectors $\theta_i^k$ according to (3) for all $i \in [n] \setminus [\alpha]$.

We denote the restriction of the score vector $\theta_i$ to the independent set $\mathcal{I}(G)$ by $\theta_i^k \in \mathbb{R}^\alpha$, where wlog we assume $\mathcal{I}(G) = [\alpha]$ as before. Furthermore, from (3) for any two $k_1, k_2 \in [N]$, we have

$$\lambda_{\min}(B^T B)\|\theta_{[\alpha]}^{k_1} - \theta_{[\alpha]}^{k_2}\|^2 \leq \|\theta^{k_1} - \theta^{k_2}\|^2 \leq \lambda_{\max}(B^T B)\|\theta_{[\alpha]}^{k_1} - \theta_{[\alpha]}^{k_2}\|^2$$

(18)

where $\lambda_{\min}(B^T B)$ and $\lambda_{\max}(B^T B)$ respectively denotes the minimum and maximum non-zero eigenvalues of the positive semi-definite matrix $B^T B$.

**Lemma D.1.** $\frac{1}{6}(b-a)^2\alpha^2\beta^2\delta^2 \leq \|\theta^{k_1}_{[\alpha]} - \theta^{k_2}_{[\alpha]}\|^2 \leq \frac{7}{6}(b-a)^2\alpha^2\beta^2\delta^2$, for all $k_1, k_2 \in [N] \times [N]$, with probability at least $(1 - N^2 e^{-\frac{\delta^2}{\alpha\beta^2}})$.

**Proof.** Firstly we note that $\|\theta^{k_1}_{[\alpha]} - \theta^{k_2}_{[\alpha]}\|^2 = \sum_{i=1}^{\alpha} (\theta^{k_1}_i - \theta^{k_2}_i)^2$ and for any $i \in [\alpha]$, $(\theta^{k_1}_i - \theta^{k_2}_i)^2 = (b-a)^2(X^{k_1}_i - X^{k_2}_i)^2$ and $E[(X^{k_1}_i - X^{k_2}_i)^2] = \frac{2}{3}\beta^2\delta^2$. Now applying Hoeffding’s inequality we have that

$$P \left( \left| \sum_{i=1}^{\alpha} (X^{k_1}_i - X^{k_2}_i)^2 - \frac{2}{3}\alpha^2\beta^2\delta^2 \right| \geq \frac{1}{2}\alpha^2\beta^2\delta^2 \right) \leq 2e^{-\frac{\delta^2}{2\alpha^2\beta^2}},$$

for any fixed $k_1, k_2 \in [N] \times [N]$, and applying union bounding above holds true for all $\binom{N}{2}$ $(k_1, k_2)$ pairs with probability $N(N-1)e^{-\frac{\delta^2}{2\alpha^2\beta^2}} \leq N^2e^{-\frac{\delta^2}{2\alpha^2\beta^2}}$. Now for any $N < e^{\frac{\delta^2}{2\alpha^2\beta^2}}$, we have $N^2e^{-\frac{\delta^2}{2\alpha^2\beta^2}} < 1$ for all $\alpha > 0$, and hence with some non-zero probability of at least $(1 - N^2 e^{-\frac{\delta^2}{2\alpha^2\beta^2}}) > 0$, we have

$$\frac{1}{6}\alpha^2\beta^2\delta^2 \leq \sum_{i=1}^{\alpha} (X^{k_1}_i - X^{k_2}_i)^2 \leq \frac{7}{6}\alpha^2\beta^2\delta^2, \ \forall k_1, k_2 \in [N] \times [N].$$

Combining above we get,

$$\frac{1}{6}(b-a)^2\alpha^2\beta^2\delta^2 \leq \|\theta^{k_1}_{[\alpha]} - \theta^{k_2}_{[\alpha]}\|^2 \leq \frac{7}{6}(b-a)^2\alpha^2\beta^2\delta^2,$$

for all $k_1, k_2 \in [N] \times [N]$, with probability at least $(1 - N^2 e^{-\frac{\delta^2}{2\alpha^2\beta^2}})$. □

For convenience let us fix $N = e^{\frac{\alpha}{2\beta^2}}$. Thus using Lemma D.1 on (18), we get

$$\frac{\lambda_{\min}(B^T B)}{6}(b-a)^2\alpha^2\beta^2\delta^2 \leq \|\theta^{k_1} - \theta^{k_2}\|^2 \leq \frac{7\lambda_{\max}(B^T B)}{6}(b-a)^2\alpha^2\beta^2\delta^2,$$

with probability at least $(1 - e^{-\frac{\alpha}{2\beta^2}})$. Now setting $\beta = \frac{\sqrt{6}}{(b-a)\sqrt[3]{\alpha\lambda_{\min}(B^T B)}}$ in above, we get

$$\delta^2 \leq \|\theta^{k_1} - \theta^{k_2}\|^2 \leq \frac{7\lambda_{\max}(B^T B)}{\lambda_{\min}(B^T B)}\delta^2,$$

with probability at least $(1 - e^{-\frac{\alpha}{2\beta^2}})$ (19)
Lemma D.2. Given any two $\theta, \theta' \in [a, b]^n$, such that $0 < a < b < 1$, we have
\[
\|e^\theta - e^{\theta'}\|^2 \leq e^{2(b+1)}\|\theta - \theta'\|^2
\]

Proof. The proof follows from the following straightforward deduction:
\[
\|e^\theta - e^{\theta'}\|^2 = \sum_{i=1}^{n}(e^{\theta_i} - e^{\theta'_i})^2 = \sum_{i=1}^{n}(e^{\theta_i})^2(e^{\theta_i} - 1)^2
\leq \sum_{i=1}^{n}e^{2b}(e^{\theta_i} - 1)^2 \leq e^{2b}\sum_{i=1}^{n}((\theta_i - \theta'_i)(e - 1))^2
\leq e^{2(b+1)}\|\theta - \theta'\|^2,
\]
where the second last inequality follows from the fact that $-1 < \theta_i - \theta'_i < 1$, for all $i \in [n]$.  

We will now assume our constructed score vectors, $\theta^k$, indeed satisfy $0 < a < \theta^k_i < b < 1$, $\forall i \in [n], \forall k \in [N]$. We will shortly show this is indeed true by our construction of $\theta^k$. Then applying Lemma D.2 and subsequently D.1 to (17) we further get,
\[
E[\|\hat{\theta} - \theta\|] \geq \frac{\delta}{2} \left[ 1 - \frac{448e^{2(b+1)}K\zeta\lambda\delta^2 + 128\log 2}{\alpha} \right], \tag{20}
\]
where $\Lambda = \frac{\lambda_{\text{max}}(B^T B)}{\lambda_{\text{min}}(B^T B)}$ and $N = \frac{\alpha}{128}$.

Thus setting $\delta = \frac{\sqrt{\alpha}}{4\sqrt{448\zeta K\lambda e^{2(b+1)}}}$, we have that
\[
448e^{2(b+1)}K\zeta\lambda\delta^2 + 128\log 2 \leq \frac{\alpha}{2}, \text{ for any } \alpha \geq 512\log 2,
\]
using which in (20) further gives
\[
E[\|\hat{\theta} - \theta\|] \geq \frac{\delta}{4} = \frac{\sqrt{\alpha}}{16\sqrt{448\zeta K\lambda e^{2(b+1)}}} = \frac{\sqrt{\alpha\lambda_{\text{min}}(B^T B)}}{16\sqrt{448\zeta K\lambda_{\text{max}}(B^T B)e^{2(b+1)}}}.
\]

Finally, the only thing left to show is that indeed in the above construction of the score vectors $\theta^k$ lies in the set $\Theta_B(a, b)$, $\forall k \in [N]$. Note that if we can show $X^k_i \in [0, 1], \forall i \in [\alpha]$, then that immediately implies $\theta^k_i \in [a, b], \forall i \in [n]$ by our construction of $\theta^k$ and the assumption on the coefficient matrix $B \in \mathbb{R}^{n \times \alpha}$ such that $\sum_{j=1}^{\alpha}B_{ij} = 1, \forall i \in [n]$.

Now we have $\left(\frac{1}{2} - \beta\delta\right) \leq X^k_i \leq \left(\frac{1}{2} + \beta\delta\right)$, $\forall i \in [n]$ and $k \in [N]$. And with $\beta = \frac{\sqrt{\delta}}{(b-a)\sqrt{\alpha\lambda_{\text{min}}(B^T B)}}$ and $\delta = \frac{\sqrt{\alpha\lambda_{\text{min}}(B^T B)}}{4\sqrt{448\zeta K\lambda_{\text{max}}(B^T B)e^{2(b+1)}}}$, we have
\[
\beta\delta = \frac{6}{4(b-a)\sqrt{448\zeta K\lambda_{\text{max}}(B^T B)e^{2(b+1)}}} < \frac{1}{2}.
\]
Hence $0 \leq X^k_i \leq 1$, $\forall i \in [n]$ and indeed we have $\theta^k \in \Theta_B(a, b)$, $\forall k \in [N]$. The desired lower bound now follows as:

$$\frac{E[\|\hat{\theta} - \theta\|]}{\|\theta\|} \geq \frac{\sqrt{\lambda_{\min}(B^T B)}}{16b\lambda_{\max}(B^T B)\sqrt{448\zeta K\epsilon^{2(b+1)}}},$$

since $\|\theta\| \leq \sqrt{\lambda_{\max}(B^T B)}\|\theta_{[a]}\| \leq b\sqrt{\lambda_{\max}(B^T B)}\alpha.$

$\square$