

LA-UR-17-23386

Approved for public release; distribution is unlimited.

Title: Parameter Estimation for Thurstone Choice Models

Author(s): Vojnovic, Milan
Yun, Seyoung

Intended for: Report

Issued: 2017-04-24

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Parameter Estimation for Thurstone Choice Models *

Milan Vojnović

*Department of Statistics
London School of Economics (LSE)
London, United Kingdom*

M.VOJNOVIC@LSE.AC.UK

Se-Young Yun

*Los Alamos National Laboratory
New Mexico*

YUNSEYOUNG@GMAIL.COM

Editor: x

Abstract

We consider the estimation accuracy of individual strength parameters of a Thurstone choice model when each input observation consists of a choice of one item from a set of two or more items (so called top-1 lists). This model accommodates the well-known choice models such as the Luce choice model for comparison sets of two or more items and the Bradley-Terry model for pair comparisons.

We provide a tight characterization of the mean squared error of the maximum likelihood parameter estimator. We also provide similar characterizations for parameter estimators defined by a rank-breaking method, which amounts to deducing one or more pair comparisons from a comparison of two or more items, assuming independence of these pair comparisons, and maximizing a likelihood function derived under these assumptions. We also consider a related binary classification problem where each individual parameter takes value from a set of two possible values and the goal is to correctly classify all items within a prescribed classification error.

The results of this paper shed light on how the parameter estimation accuracy depends on given Thurstone choice model and the structure of comparison sets. In particular, we found that for unbiased input comparison sets of a given cardinality, when in expectation each comparison set of given cardinality occurs the same number of times, for a broad class of Thurstone choice models, the mean squared error decreases with the cardinality of comparison sets, but only marginally according to a diminishing returns relation. On the other hand, we found that there exist Thurstone choice models for which the mean squared error of the maximum likelihood parameter estimator can decrease much faster with the cardinality of comparison sets.

We report empirical evaluation of some claims and key parameters revealed by theory using both synthetic and real-world input data from some popular sport competitions and online labor platforms.

1. Introduction

We consider the statistical inference problem of estimating individual strength or skill parameters of items based on observed choices of items from sets of two or more items. This

*. A preliminary version of this work was published in ICML 2016.

accommodates the case of pair comparisons as a special case, where each comparison set consists of two items. In our more general case, each observation consists of a comparison set of two or more items and the identity of the chosen item from this set. In other words, each observation is a partial ranking, which is often referred to as a top-1 list. Many applications are accommodated by this framework, e.g., choices indicated by user clicks in various information retrieval systems, outcomes of single-winner contests in crowdsourcing services such as TopCoder or Taskcn, outcomes of hiring decisions where one applicant is hired among those who applied for a job, e.g., in online labour markets such as Fiverr and Upwork, as well as numerous sport competitions and online gaming platforms.

We consider the parameter estimation for the statistical choice model known as the *Thurstone choice* model; also referred to as the *random utility model*. According to the Thurstone choice model, items are associated with latent performance random variables that are independent across different items and different comparisons. For any given comparison set, the choice is the item from this set with the largest performance random variable. For any given item, the performance random variable is equal to the sum of a strength parameter, whose value can be specific to this item, and a noise random variable according to a given cumulative distribution function. The values of the strength parameters are unknown and have to be estimated from the observed comparisons, and the distribution of noise random variables is assumed to be known. The Thurstone choice model accommodates many known choice models by admitting different distribution for noise random variables, e.g., Luce choice model (Luce (1959)) for comparison sets of two or more items, and its special case for pair comparisons, often referred to as the Bradley-Terry model (Bradley and Terry (1954)).

In this paper, we study the accuracy of the maximum likelihood estimator of the parameter of the Thurstone choice model. Our goal is to characterize the accuracy of the maximum likelihood estimator and shed light on how it depends on the given Thurstone choice model and properties of the observed input data such as the number of observations and the structure of comparison sets. In particular, we consider the following statistical inference question. Suppose that the input observations are such that all comparison sets are of the same cardinality $k \geq 2$ and are unbiased, meaning that in expectation every comparison set of given cardinality occurs the same number of times in the input data. Then, we would like to understand how does the accuracy of the maximum likelihood estimator of the strength parameters depend on the cardinality of comparison sets. Notice that from any comparison set of cardinality k , we can deduce at most $k(k-1)/2$ pair comparisons. Intuitively, we would expect that the parameter estimation accuracy would increase with the cardinality of comparison sets. However, it is not a priori clear how fast the accuracy would improve and whether any significant gains can be achieved by increasing the sizes of comparison sets. Moreover, it is also not a priori clear whether or not there can be any significant difference between different Thurstone choice models, with respect to how the parameter estimation accuracy is related to the cardinality of comparison sets. We also consider these questions for parameter estimators that are derived by rank breaking methods, which amount to deducing one or more pair comparisons from each comparison of two or more items, assuming independence of these pair comparisons, and defining the estimator as the maximum likelihood estimator under these assumptions.

The main contributions of this paper can be summarized as follows.

We provide upper bounds on the mean squared error of the maximum likelihood estimator, and a lower bound that establishes their minimax optimality. We show that the effect of the structure of comparison sets on the mean squared error is captured by one key parameter: algebraic connectivity of a suitably defined weighted-adjacency matrix. The elements of this matrix correspond to distinct pairs of items and are equal to a weighted sum of the number of input comparisons of different cardinalities that contain the corresponding pair of items, where the weights are specific to given Thurstone choice model.

For the statistical inference question of how the estimation accuracy improves with the cardinality of comparison sets, we derive a tight characterization of the mean squared error in terms of the cardinality of comparison sets (Corollary 12). This characterization reveals that for a broad class of Thurstone choice models, which includes the well-known cases such as the Luce choice model, there is a diminishing returns decrease of the mean squared error with the cardinality of comparison sets. For this class of Thurstone choice models, the mean squared error tends to be largely insensitive to the cardinality of comparison sets. On the other hand, we show that there exist Thurstone choice models for which the mean squared error decreases much faster with the cardinality of comparison sets. Perhaps surprisingly, in these cases, the amount of information extracted from a comparison set of cardinality k is in the order of k^2 independent pair comparisons, which yields a $1/k^2$ reduction of the mean squared error of the maximum likelihood estimator. Section 6 provides more discussion.

We consider two natural rank-breaking methods, one that deduces $k-1$ pair comparisons and one that deduces 1 pair comparison from a comparison set of cardinality k (Section 4). We derive mean squared error upper bounds when choices are according to the Luce choice model for these two rank-breaking methods in Theorem 17 and Theorem 18, respectively. These results show that both estimators are consistent. Interestingly, both mean squared error upper bounds are equal to that of the maximum likelihood estimator up to a constant factor. Hence, they both inherit all the properties that we established to hold for the mean squared error upper bound for the maximum likelihood estimator.

We also consider a binary classification problem where all strength parameters associated with items take one of two possible values (separating items into two classes), and the goal is to correctly classify each item within a prescribed probability of classification error (Section 5). We identify sufficient conditions for correctness of a simple point score classification algorithm (Theorem 19) and establish their tightness (Theorem 20). These conditions are of the same form as those that we imposed for deriving upper bounds on the mean squared error of the maximum likelihood parameter estimator.

We present experimental results using both simulation and real-world data (Section 7). In particular, we validate the claim that the mean squared error can decrease with the cardinality of comparison sets in a qualitatively different way depending on the given Thurstone choice model. We also evaluate algebraic connectivity of weighted-adjacency matrices for several real-world input data, demonstrating that it can cover a wide range of values depending on specific application scenario.

1.1 Related Work

A model of comparative judgement for pair comparisons was introduced by Thurstone (1927), which is a special case of a model that we refer to as a Thurstone choice model,

for the case of pair comparisons and Gaussian random noise variables. A statistical model of pair comparisons that postulates that an item is chosen from a set of two items with probability proportional to the strength parameter of this item was introduced by Zermelo (1929), and was then popularized by the work of Bradley and Terry (1952, 1954) and others, and is often referred to as the Bradley-Terry model. The statistical model of choice where for any set of two or more items an item is chosen with probability proportional to its strength parameter was shown to be a unique model satisfying a set of axioms introduced by Luce (1959), and is referred as the Luce choice model. The Bradley-Terry model is the special case of the Luce choice model for pair comparisons. The choice probabilities of the Luce choice model correspond to those of a Thurstone choice model with noise random variables according to a double-exponential distribution. Relationships between the Luce choice model and Thurstone choice model were studied by Yellott (1977). A statistical model for full ranking outcomes (the outcome of a comparison is an ordered list of the compared items) where the ranking is in the order of sampling of items without replacement from the set of compared items and the sampling probabilities are proportional to the strengths of items is referred as the Plackett-Luce model (Luce (1959) and Plackett (1975)).

The Thurstone choice models have been used in the design of several popular skill rating systems, e.g., Elo rating system by Elo (1978) used for skill rating of chess players and in other sports, and TrueSkill by Graepel et al. (2006) used for skill rating of gamers of a popular online gaming platform. All these models are instances of Thurstone models, and are special instances of generalized linear models, see, e.g., Nelder and Wedderburn (1972), McCullagh and Nelder (1989), and Chapter 9 in Murphy (2012). An exposition to the principles of skill rating systems is available in Chapter 9 in Vojnović (2016).

The parameter estimation problem for the Bradley-Terry model of pair comparisons has been studied by many. The iterative methods for computing a maximum likelihood parameter estimate have been studied in the early work by Hunter (2004) and the recent work by Maystre and Grossglauser (2015). Simons and Yao (1999) shown that the maximum likelihood parameter is consistent and asymptotically normal as the number of items n grows large, under assumption that each pair is compared the same number of times and that the true parameter vector is such that the maximum distance between any of its coordinates is $o(\log(n))$. Maydeau-Olivares (1999) studied Thurstonian model parameter estimation with noise random variables according to a Gaussian distribution.

The accuracy of the parameter estimation for various instances of Thurstone models has been studied in recent work. Negahban et al. (2012) found a sufficient number of input pair comparisons to achieve a given mean squared error of a parameter estimator for the Bradley-Terry model and the input comparisons such that in expectation each distinct pair is compared the same number of times. In particular, they shown that under given assumptions, it suffices to observe $O(n \log(n))$ input pair comparisons. Rajkumar and Agarwal (2014) studied a statistical convergence property of ranking aggregation algorithms for pair comparisons not only for the Bradley-Terry model but also under some more general conditions referred to as low-noise and generalized low-noise. Hajek et al. (2014) provided a characterization of the mean squared error of the maximum likelihood parameter estimate for the Plackett-Luce model of full ranking outcomes. This work found that the algebraic connectivity of a weighted-adjacency matrix captures the effect of the structure of input comparison sets on the mean squared error of the maximum likelihood parameter estimator.

Shah et al. (2016) established similar characterization results for the case of pair comparisons according to Thurstone choice models.

Our work differs from previous work in that we consider characterization of the estimation accuracy for a general class of Thurstone choice models for arbitrary sizes of comparisons. Specifically, our work provides a first characterization of the estimation accuracy with respect to the cardinality of comparison sets for unbiased input comparisons, which reveals an insight into the fundamental limits of statistical inference for given cardinality of comparison sets. It is a folklore that different models of pair comparisons yield similar performance with respect to the prediction error, e.g. Stern (1992), which suggests that the precise choice of a Thurstone choice model does not matter much in applications. Our results show that there can be significant difference between two Thurstone choice models with respect to the statistical inference of model parameters.

Parameter estimators derived by various rank-breaking methods have been studied by various authors. For instance, Soufiani et al. (2013) and Soufiani et al. (2014) studied rank-breaking methods for full ranking data and Khetan and Oh (2016b) studied rank-breaking methods for partial rankings. Recently, Khetan and Oh (2016a) characterized a trade-off between the amount of information used per comparison and the mean squared error of a parameter estimate based on rank breaking. Our work is different in that we are interested in top-1 list observations and the effect of the structure of comparison sets for rank-breaking methods. For the top-1 list observations and any given comparison structure, our work provides upper bounds for the mean squared error of two natural ranking breaking methods and shows the optimality of them.

1.2 Organization of the Paper

Section 2 introduces problem formulation, some basic concepts and key technical results used to establish our main results. Section 3 provides a characterization of the mean squared error for the maximum likelihood parameter estimation, including both upper and lower bounds. Section 4 shows the same type of characterizations for two rank-breaking based parameter estimators. Section 5 establishes tight conditions for correct classification of items, when the strength parameters of items are of two possible types. Section 6 discusses how the estimation accuracy depends on the cardinality of comparison sets. Section 7 contains results of our experiments. Finally, Section 8 concludes the paper. Appendix contains some background facts and proofs of our theorems.

2. Problem Formulation and Notation

Let $N = \{1, 2, \dots, n\}$ be a set of two or more items. The input data consists of a sequence of one or more observations $(S_1, y_1), (S_2, y_2), \dots, (S_m, y_m)$, where each observation t consists of (a) a *comparison set* $S_t \subseteq N$ and (b) a *choice* of an item y_t from S_t . The case of *pair comparisons* is accommodated as a special case when each comparison set consists of two items. Let $w_{i,S}$ denote the number of observations for which the comparison set is S and the chosen item is $i \in S$. With a slight abuse of notation, for pair comparisons, let $w_{i,j}$ be the number of observations for which the comparison set is $\{i, j\}$ and the chosen item is i .

A Thurstone choice model, denoted as \mathcal{T}_F , is defined by a cumulative distribution function F of a zero mean random variable and parameter vector $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ that takes

value in $\Theta \subseteq \mathbf{R}^n$. We can interpret θ_i as the strength of item $i \in N$. The cumulative distribution function F is assumed to have a density function, denoted by f .

According to Thurstone choice model \mathcal{T}_F , for any given sequence of comparison sets, choices are independent random variables according to the following distribution: conditional on that the comparison set is S , with k denoting the cardinality of S , the distribution of choice is

$$p_{i,S}(\theta) = p_k(\mathbf{x}_{i,S}(\theta)), \text{ for } i \in S \quad (1)$$

where $\mathbf{x}_{i,S}(\theta)$ is a vector in \mathbf{R}^{k-1} with elements $\theta_i - \theta_j$, for $j \in S \setminus \{i\}$, and

$$p_k(\mathbf{x}) = \int_{\mathbf{R}} \left(\prod_{v=1}^{k-1} F(x_v + z) \right) f(z) dz, \text{ for } \mathbf{x} \in \mathbf{R}^{k-1}. \quad (2)$$

With a slight abuse of notation, for the case of pair comparisons, we let $p_{i,j}(\theta)$ denote the probability that item i is chosen from the comparison set $\{i, j\}$. In this case, we have

$$p_{i,j}(\theta) = p_2(\theta_i - \theta_j)$$

where

$$p_2(x) = \int_{\mathbf{R}} F(x + z) f(z) dz, \text{ for } x \in \mathbf{R}.$$

A Thurstone choice model \mathcal{T}_F corresponds to the following *probabilistic generative model* of choice, also referred to as a *random utility model*. The items of each comparison set are associated with latent *performance random variables*, which are independent across different items and different comparison sets. For any comparison set S and item $i \in S$, the performance random variable X_i is equal to the sum of a strength parameter θ_i and a noise random variable with distribution F . The given probabilistic generative model assumes that for any given comparison set S , the chosen item is the one with the largest performance. Hence, the distribution of choice is $p_{i,S}(\theta) = \mathbf{P}[X_i \geq \max_{j \in S} X_j]$ for $i \in S$, which can be expressed as asserted in (1).

It is noteworthy that under a Thurstone choice model, the probability distribution of choice depends only on pairwise differences of the strength parameters. This implies that the probability distribution of choice is shift-invariant with respect to the parameter vector. In order to allow for identifiability of the parameter vector, we assume that θ is such that $\sum_{i=1}^n \theta_i = 0$.

We refer to several examples of Thurstone choice models as follows: (i) *Gaussian distribution*: $f(x) = 1/(\sqrt{2\pi}\sigma) \exp(-x^2/(2\sigma^2))$ with variance σ^2 ; (ii) *Double-exponential distribution*: $F(x) = \exp(-\exp(-(x + \beta\gamma)/\beta))$ with parameter $\beta > 0$ and γ denoting the Euler-Mascheroni constant, which has variance $\sigma^2 = \pi^2\beta^2/6$; (iii) *Laplace distribution*: $F(x) = 1/2 \exp(x/\beta)$, for $x < 0$, and $F(x) = 1 - 1/2 \exp(-x/\beta)$, for $x \geq 0$, with parameter β , which has variance $\sigma^2 = 2\beta^2$; and (iv) *Uniform distribution*: $f(x) = 1/(2a)$ for $x \in [-a, a]$, which has variance $\sigma^2 = a^2/3$.

For the general case of comparison sets of two or more items, the distribution of choice admits an explicit form only for some special cases. For example, when noise random variables are according to a double-exponential distribution, we have

$$p_k(\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{k-1} e^{-x_i/\beta}}, \text{ for } \mathbf{x} \in \mathbf{R}^{k-1}.$$

This amounts to the choice probabilities $p_{i,S}(\theta) = \exp(\theta_i/\beta) / \sum_{v \in S} \exp(\theta_v/\beta)$, for $i \in S$ and $S \subseteq N$, which under suitable re-parametrisation corresponds to the well-known Luce choice model. In particular, for pair comparisons, we have $p_2(x) = 1/(1 + \exp(-x/\beta))$, which under suitable re-parametrisation corresponds to the well-known Bradley-Terry model. For pair comparisons, the choice probabilities admit an explicit form also for some other Thurstone choice models; for example, when noise has a Gaussian distribution, we have $p_2(x) = \Phi(x/(2\sigma))$ where Φ is the cumulative distribution function of a standard normal random variable.

Maximum Pseudo Likelihood Estimation We consider parameter estimators that are defined as maximizers of a pseudo log-likelihood function $\ell : \Theta \rightarrow \mathbf{R}$. We refer to the parameter estimator $\hat{\theta} \in \arg \max_{\theta \in \Theta} \ell(\theta)$ as a maximum pseudo likelihood estimator.

We devote a special attention to maximum likelihood estimator, defined as a maximizer of the log-likelihood function, which for a Thurstone choice model is given by

$$\ell(\theta) = \sum_{t=1}^m \log(p_{y_t, S_t}(\theta)). \quad (3)$$

The log-likelihood function can be written as follows

$$\ell(\theta) = \sum_{S \subseteq N} \sum_{i \in S} w_{i,S} \log(p_{|S|}(\mathbf{x}_{i,S}(\theta))) + \text{const} \quad (4)$$

where recall $w_{i,S}$ is the number of observations for which the comparison set is S and i is the choice from S . In particular, for pair comparisons, we have

$$\ell(\theta) = \sum_{i=1}^n \sum_{j=1}^n w_{i,j} \log(p_2(\theta_i - \theta_j)) + \text{const}. \quad (5)$$

Evaluating the value of the log-likelihood function in (4) for given parameter vector requires evaluating a sum that in the worst-case consists of exponentially many elements in n (all possible combinations of two or more elements from the ground set of n elements). On the other hand, for pair comparisons, the log-likelihood function in (5) is a sum of at most n^2 elements; thus, polynomially many elements in n . A common approach to reduce computational complexity is to use the so-called *rank breaking*, which amounts to deducing pair comparisons from any given comparison set of two or more items, and assuming that these pair comparisons are independent (if this is not the case). Using these pair comparisons, one then defines a pseudo log-likelihood function as the log-likelihood function under the assumption that the deduced pair comparisons are independent.

We shall consider two natural rank-breaking methods. The first rank-breaking method deduces $k-1$ pair comparisons from each comparison set of k items, by taking all pairs that consist of the chosen item and each other item in the given comparison set. The pseudo log-likelihood function in this case is given by

$$\ell_{k-1}(\theta) = \sum_{t=1}^m \sum_{v \in S_t \setminus \{y_t\}} \log(p_{y_t, v}(\theta)). \quad (6)$$

The second rank-breaking method that we consider deduces 1 pair comparison from each comparison set of k items, by taking a pair that consists of the chosen item and a randomly picked item from the remaining set of items in the given comparison set. The pseudo log-likelihood function in this case is given by

$$\ell_1(\theta) = \sum_{t=1}^m \log(p_{y_t, z_t}(\theta)). \quad (7)$$

The first rank-breaking method uses maximum amount of information that is contained in a comparison; by observing choice of one item from a comparison set of k items, we can indeed deduce at most $k - 1$ pair comparisons (between the chosen item and each other item in the given comparison set). In general, these pair comparisons are not mutually independent. The second rank-breaking method is conservative in deducing only one pair comparison from each comparison set of two or more items.

Parameter Estimation Accuracy We study the accuracy of a maximum pseudo log-likelihood estimator $\hat{\theta}$ of the true parameter vector θ^* by using the *mean squared error* defined as follows:

$$\text{MSE}(\hat{\theta}, \theta^*) = \frac{1}{n} \|\hat{\theta} - \theta^*\|_2^2. \quad (8)$$

We also consider the probability of classification error for the case when the strength parameters belong to one of two classes and the goal is to correctly classify each item.

2.1 Eigenvalues, Adjacency, and Laplacian Matrices

Here we review some basic definitions that are used throughout the paper. We denote eigenvalues of a matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$ as $\lambda_1(\mathbf{A}), \lambda_2(\mathbf{A}), \dots, \lambda_n(\mathbf{A})$. By convention, we assume that $\lambda_1(\mathbf{A}) \leq \lambda_2(\mathbf{A}) \leq \dots \leq \lambda_n(\mathbf{A})$. The *spectral norm* $\|\mathbf{A}\|_2$ of matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$ is defined by $\|\mathbf{A}\|_2 = \sqrt{\lambda_n(\mathbf{A}^\top \mathbf{A})}$. The spectral norm of \mathbf{A} is induced by the Euclidean vector norm as follows $\|\mathbf{A}\|_2 = \max\{\|\mathbf{A}\mathbf{x}\|_2 : \mathbf{x} \in \mathbf{R}^n, \|\mathbf{x}\| = 1\}$. If $\mathbf{A} \in \mathbf{R}^{n \times n}$ is a real symmetric matrix, then $\|\mathbf{A}\|_2 = \lambda_n(\mathbf{A})$.

For any *weighted-adjacency* matrix $\mathbf{A} \in \mathbf{R}_+^{n \times n}$, we consider a *Laplacian* matrix $L_{\mathbf{A}}$ defined by

$$L_{\mathbf{A}} = \text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$$

where for any given vector \mathbf{a} , $\text{diag}(\mathbf{a})$ denotes the diagonal matrix with diagonal \mathbf{a} .

For any weighted-adjacency matrix $\mathbf{A} \in \mathbf{R}_+^{n \times n}$, we refer to $\lambda_2(L_{\mathbf{A}})$ as the *Fiedler value* of \mathbf{A} (Fiedler (1973, 1989)). The Fiedler eigenvalue of a weighted-adjacency matrix quantifies its algebraic connectivity. A weighted-adjacency matrix \mathbf{A} corresponds to a connected graph if and only if it has strictly positive Fiedler value, i.e., $\lambda_2(L_{\mathbf{A}}) > 0$.

For any given observations and given *weight function* $w : \{1, 2, \dots, n\} \rightarrow \mathbf{R}_+$, we define the weighted-adjacency matrix \mathbf{M}_w as follows. Let $m_{i,j}(k)$ be the number of comparisons of cardinality k that contain the pair of items $\{i, j\}$. Then, we define \mathbf{M}_w to be the matrix in $\mathbf{R}_+^{n \times n}$ with zero diagonal elements and other elements given by

$$m_{i,j} = \frac{n}{m} \sum_{k \geq 2} w(k) m_{i,j}(k). \quad (9)$$

With a slight abuse of notation, let \mathbf{M}_a be the weighted-adjacency matrix defined for the weight function that takes constant value $a > 0$, and let \mathbf{M} be written in lieu of \mathbf{M}_1 .

If all comparison sets have identical cardinalities, then each element of the weighted-adjacency matrix is equal to the number of comparisons that contain the corresponding pair of items up to a multiplicative factor. The factor n/m can be interpreted as a normalization with the mean number of comparison sets per item. This normalization is admitted so that for the canonical case of pair comparisons when each pair is compared the same number of times, $\lambda_2(L_{\mathbf{M}_a})$ is a constant independent of the number of observations m and the number of items asymptotically for large n . Indeed, in this case, $\lambda_2(L_{\mathbf{M}_a}) = \dots = \lambda_n(L_{\mathbf{M}_a}) = (1 - 1/n)a$, which is equal to constant a , asymptotically for large n .

We say that comparison sets are *unbiased* if for any given cardinality, each set of the given cardinality occurs the same number of times. In particular, for pair comparisons, this means that each distinct pair is compared the same number of times. For any unbiased comparison sets, the weighted-adjacency matrix can be expressed as follows. Let $\mu(k)$ be the fraction of comparison sets of cardinality k . Then, for every integer $k \geq 2$ and pair of items $\{i, j\}$, $m_{i,j}(k) = \left(\binom{n-2}{k-2} / \binom{n}{k} \right) \mu(k)m = (k(k-1)/[n(n-1)]) \mu(k)m$. Hence, for every pair of items, we have

$$m_{i,j} = \frac{1}{n-1} \sum_{k \geq 2} w(k)k(k-1)\mu(k). \quad (10)$$

It follows that for any unbiased comparison sets, we have

$$\lambda_2(L_{\mathbf{M}_w}) = \left(1 - \frac{1}{n}\right) \sum_{k \geq 2} w(k)k(k-1)\mu(k) \quad (11)$$

which is a constant independent of n , asymptotically for large n .

We shall also consider comparison sets that are assumed to be an independent random sequence according to a given distribution. Specifically, we shall consider the case where all comparison sets are of the same cardinality, and are independent samples according to uniform random sampling without replacement from the set of all items. We denote with $\overline{\mathbf{M}}_w$ the *expected weighted-adjacency* matrix, where the expectation is with respect to the distribution of the sequence of comparison sets. We say that comparison sets are *a priori unbiased* if all non-diagonal elements of $\overline{\mathbf{M}}_w$ are equal.

2.2 A Key Lemma and Probability Tail Bounds

All upper bounds for the mean squared error of a maximum pseudo log-likelihood estimator in this paper are established by using the following key lemma.

Lemma 1 Suppose that $g : \mathbf{R}^n \rightarrow \mathbf{R}$ satisfies (i) $\nabla^2 g(\theta)\mathbf{1} = \mathbf{0}$ and (ii) $\lambda_2(\nabla^2 g(\theta)) > 0$ for all $\theta \in \Theta$, where $\Theta = \{\theta \in [-b, b]^n : \theta^\top \mathbf{1} = 0\}$ for $b > 0$. Let θ^* be an arbitrary vector in Θ and $\hat{\theta} \in \arg \min_{\theta \in \Theta} g(\theta)$. Then, we have

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{2 \|\nabla g(\theta^*)\|_2}{\min_{\theta \in \Theta} \lambda_2(\nabla^2 g(\theta))}.$$

We shall apply this lemma to the case where g is a negative pseudo log-likelihood function, $\hat{\theta}$ is a maximizer of the pseudo log-likelihood function, and θ^* is the true parameter vector. We upper bound the mean squared error by the following two steps:

S1 find $\alpha > 0$ such that $\|\nabla g(\theta^*)\|_2 \leq \alpha$, and

S2 find $\beta > 0$ such that $\min_{\theta \in \Theta} \lambda_2(\nabla^2 g(\theta)) \geq \beta$

which imply the upper bound $\|\hat{\theta} - \theta^*\|_2 \leq 2\alpha/\beta$.

All our proofs of the mean squared estimation error upper bounds follow the above two-step procedure, including the proof of Theorem 4 in Section 3.1 and other proofs provided in Appendix.

In step **S1**, $\nabla g(\theta^*)$ is a sum of random vectors. We will make use of the following vector version of Azuma-Hoeffding bound (Theorem 1.8 in Hayes (2003)) for a sum of random vectors.

Lemma 2 (vector Azuma-Hoeffding bound) *Suppose that $S_m = \sum_{t=1}^m X_t$ is a martingale where X_1, X_2, \dots, X_m are random variables that take values in \mathbf{R}^n and are such that $\mathbf{E}[X_t] = \mathbf{0}$ and $\|X_t\|_2 \leq \sigma$ for all $t \in \{1, 2, \dots, m\}$, for $\sigma > 0$. Then, for every $x > 0$,*

$$\mathbf{P} [\|S_m\|_2 \geq x] \leq 2e^2 e^{-\frac{x^2}{2m\sigma^2}}.$$

In step **S2**, we need to find a lower bound for the second-smallest eigenvalue of the Hessian matrix $\nabla^2 g(\theta)$ for all $\theta \in \Theta$. For pair comparisons according to a Thurstone choice model or comparisons of two or more items according to the Luce choice model, $\nabla^2 g(\theta)$ is determined by comparison sets and does not depend on the choices. We can find β from a Laplacian matrix when the comparison sets are given. In other cases, $\nabla^2 g(\theta)$ is a sum of random matrices. We will make use the following matrix version of Chernoff's bound along with properties of eigenvalues of a Laplacian matrix (which are given in Appendix A).

Lemma 3 (matrix Chernoff bound) *Let $S_m = \sum_{t=1}^m X_t$ where X_1, X_2, \dots, X_m are random independent real symmetric matrices in $\mathbf{R}^{n \times n}$ such that $\lambda_1(X_t) \geq 0$ and, $\|X_t\|_2 \leq \sigma$ for $t \in \{1, 2, \dots, m\}$, for $\sigma > 0$. Then, for $\epsilon \in [0, 1)$,*

$$\mathbf{P} [\lambda_1(S_m) \leq (1 - \epsilon)\lambda_1(\mathbf{E}[S_m])] \leq ne^{-\frac{\epsilon^2 \lambda_1(\mathbf{E}[S_m])}{2\sigma}}.$$

3. Maximum Likelihood Estimation

In this section, we present upper and lower bounds for the mean squared error of the maximum likelihood parameter estimator for the Thurstone choice model. We will first consider the case of pair comparisons. We then consider the more general case when each comparison set consists of two or more items. For this more general case, we first give an upper bound for the Luce choice model, and then present similar characterization for a class of Thurstone choice models. We end this section with a lower bound on the mean squared error of the maximum likelihood parameter estimator, which establishes minimax optimality of our upper bounds.

3.1 Pair Comparisons

We consider pair comparisons according to a Thurstone choice model \mathcal{T}_F with parameter vector θ^* that takes value in $\Theta = [-b, b]^n$ and that satisfies the following conditions:

P1 There exists $A > 0$ such that

$$\frac{d^2 \log(p_2(x))}{dx^2} \leq -A \text{ for all } x \in [-2b, 2b]. \quad (12)$$

P2 There exists $B > 0$ such that

$$\frac{d \log(p_2(x))}{dx} \leq B \text{ for all } x \in [-2b, 2b]. \quad (13)$$

P3 The weighted-adjacency matrix \mathbf{M} is irreducible, i.e. $\lambda_2(L_{\mathbf{M}}) > 0$.

Condition **P1** means that p_2 is a strictly log-concave function on $[-2b, 2b]$. Condition **P2** means that $\log(p_2(x))$ has a bounded derivative on $[-2b, 2b]$. Notice that this condition is equivalent to $dp_2(x)/dx \leq Bp_2(x)$ for all $x \in [-2b, 2b]$. Constants A and B are specific for given Thurstone choice model and the value of the parameter b . In particular, for the Bradley-Terry model, it can be easily checked that **P1** and **P2** hold with $A = e^{-2b/\beta}/[\beta^2(1 + e^{-2b/\beta})^2]$ and $B = 1/[\beta(1 + e^{-2b/\beta})]$. Condition **P3** means that the observations are such that the graph defined by the weighted-adjacency matrix is connected. Equivalently, there exists no partition of the set of items N into two non-empty sets S and $N \setminus S$ such that some pair of items $i \in S$ and $j \in N \setminus S$ is not compared in the input observations.

Theorem 4 *Under conditions **P1**, **P2** and **P3**, with probability at least $1 - 2/n$,*

$$\text{MSE}(\hat{\theta}, \theta^*) \leq D^2 \frac{n(\log(n) + 2)}{\lambda_2(L_{\mathbf{M}_{1/4}})^2} \frac{1}{m} \quad (14)$$

where $D = B/A$.

Before we show a proof of the theorem, we note the following remarks.

First, notice that D is a constant whose value is specific to given Thurstone choice model and the value of parameter b . In particular, for the Bradley-Terry model, we have $D = \beta(e^{2b/\beta} + 1)$.

Second, from (14), for the mean squared error to be smaller than or equal to ϵ^2 , for given $\epsilon > 0$, it suffices that the number of observations is such that

$$m \geq \frac{1}{\epsilon^2} D^2 \frac{1}{\lambda_2(L_{\mathbf{M}_{1/4}})^2} n(\log(n) + 2). \quad (15)$$

Third, and last, if each pair of items is compared the same number of times, then, from (10), we have $m_{i,j} = 1/(2(n-1))$ for all $i \neq j$. Hence, in this case $\lambda_2(L_{\mathbf{M}_{1/4}}) = n/(2(n-1))$, and, from (15), it suffices that the number of observations m is such that

$$m \geq \frac{4}{\epsilon^2} D^2 n(\log(n) + 2).$$

Proof of Theorem 4 We now go on to provide a proof of Theorem 4. For pair comparisons, the log-likelihood function (3) can be written as

$$\ell(\theta) = \sum_{t=1}^m \log(p_2(\theta_{y_t} - \theta_{z_t}))$$

where y_t denotes the choice from the comparison pair $S_t = \{y_t, z_t\}$ for each observation t . The negative log-likelihood function satisfies the relation in Lemma 1, which combined with the following two lemmas, establishes the statement of the theorem.

Lemma 5 *The following relation holds:*

$$\min_{\theta \in \Theta} \lambda_2(\nabla^2(-\ell(\theta))) \geq 4A \frac{m}{n} \lambda_2(L_{\mathbf{M}_{1/4}}).$$

Proof It is easy to verify that for all $\theta \in \mathbf{R}^n$ and $i, j \in N$, we have the following identities

$$\begin{aligned} \frac{d^2 \log(p_2(\theta_i - \theta_j))}{dx^2} &= \frac{\partial^2 \log(p_2(\theta_i - \theta_j))}{\partial \theta_i^2} \\ &= \frac{\partial^2 \log(p_2(\theta_i - \theta_j))}{\partial \theta_j^2} \\ &= \frac{\partial^2 (-\log(p_2(\theta_i - \theta_j)))}{\partial \theta_i \partial \theta_j}. \end{aligned}$$

For all $i, j \in N$ such that $i \neq j$, we have

$$\frac{\partial^2 (-\log(p_2(\theta_{y_t} - \theta_{z_t})))}{\partial \theta_i \partial \theta_j} = \begin{cases} \frac{d^2}{dx^2} \log(p_2(\theta_{y_t} - \theta_{z_t})), & \text{if } \{i, j\} = \{y_t, z_t\} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\frac{\partial^2 (-\log(p_2(\theta_{y_t} - \theta_{z_t})))}{\partial \theta_i^2} = - \sum_{j \neq i} \frac{\partial^2 (-\log(p_2(\theta_{y_t} - \theta_{z_t})))}{\partial \theta_i \partial \theta_j}.$$

Combining with condition **P1**, we have

$$\frac{\partial^2 (-\log(p_2(\theta_{y_t} - \theta_{z_t})))}{\partial \theta_i \partial \theta_j} \leq -A 1_{\{y_t, z_t\} = \{i, j\}}.$$

It follows that

$$\nabla^2(-\ell(\theta)) \succeq 4A \frac{m}{n} L_{\mathbf{M}_{1/4}}, \text{ for all } \theta \in [-b, b]^n \quad (16)$$

where for two matrices \mathbf{A} and \mathbf{B} , $\mathbf{A} \succeq \mathbf{B}$ is equivalent to saying that $\mathbf{A} - \mathbf{B}$ is positive definite; see Appendix A. In (16), both $\nabla^2(-\ell(\theta))$ and $4A \frac{m}{n} L_{\mathbf{M}_{1/4}}$ are positive-definite matrices. Hence, by the elementary fact stated in Lemma 23 (Appendix), we obtain the assertion of the lemma. \blacksquare

Lemma 6 *With probability at least $1 - 2/n$,*

$$\|\nabla(-\ell(\theta^*))\|_2 \leq 2B\sqrt{m(\log(n) + 2)}.$$

Proof $\nabla(-\ell(\theta))$ is a sum of independent random vectors in \mathbf{R}^n given by

$$\nabla(-\ell(\theta)) = \sum_{t=1}^m \nabla(-\log(p_2(\theta_{y_t} - \theta_{z_t}))).$$

The elements of $\nabla(-\log(p_2(\theta_{y_t} - \theta_{z_t})))$ can be expressed as follows

$$\frac{\partial(-\log(p_2(\theta_{y_t} - \theta_{z_t})))}{\partial\theta_i} = \begin{cases} -\frac{d\log(p_2(\theta_{y_t} - \theta_{z_t}))}{dx}, & \text{if } i = y_t \\ +\frac{d\log(p_2(\theta_{y_t} - \theta_{z_t}))}{dx}, & \text{if } i = z_t \\ 0, & \text{otherwise.} \end{cases}$$

If $i \notin \{y_t, z_t\}$, then clearly

$$\mathbf{E} \left[\frac{\partial(-\log(p_2(\theta_{y_t}^* - \theta_{z_t}^*)))}{\partial\theta_i} \right] = 0.$$

Otherwise, we have

$$\begin{aligned} \mathbf{E} \left[\frac{\partial(-\log(p_2(\theta_{y_t}^* - \theta_{z_t}^*)))}{\partial\theta_i} \right] &= -p_2(\theta_{y_t}^* - \theta_{z_t}^*) \frac{d\log(p_2(\theta_{y_t}^* - \theta_{z_t}^*))}{dx} \\ &\quad + p_2(\theta_{z_t}^* - \theta_{y_t}^*) \frac{d\log(p_2(\theta_{z_t}^* - \theta_{y_t}^*))}{dx} \\ &= -\frac{dp_2(\theta_{y_t}^* - \theta_{z_t}^*)}{dx} + \frac{dp_2(\theta_{z_t}^* - \theta_{y_t}^*)}{dx} \\ &= 0 \end{aligned}$$

where the last equation is by the fact that $dp_2(x)/dx$ is an even function.

By condition **P2**, we have

$$\begin{aligned} \|\nabla(-\log(p_2(\theta_{y_t} - \theta_{z_t})))\|_2^2 &= \left(\frac{d\log(p_2(\theta_{y_t} - \theta_{z_t}))}{dx} \right)^2 + \left(-\frac{d\log(p_2(\theta_{y_t} - \theta_{z_t}))}{dx} \right)^2 \\ &\leq 2B^2. \end{aligned}$$

Hence, by the vector Azuma-Hoeffding bound in Lemma 2, we have

$$\mathbf{P}[\|\nabla(-\ell(\theta^*))\|_2 \geq 2B\sqrt{m(\log(n) + 2)}] \leq \frac{2}{n}.$$

■

3.2 Comparison Sets of Two or More Items

We now consider a more general case where each comparison set consists of two or more items. We first show an upper bound for the mean squared error of the maximum likelihood parameter estimator when the choices are according to the Luce choice model and comparison sets are of identical sizes. We then present a similar characterization under more general assumption that allow for a broader set of Thurstone choice models and non-identical sizes of comparison sets.

Theorem 7 *Suppose that choices are according to the Luce choice model, all comparison sets are of cardinality $k \geq 2$, and $\lambda_2(L_{\mathbf{M}}) > 0$. Then, with probability at least $1 - 2/n$,*

$$\text{MSE}(\hat{\theta}, \theta^*) \leq D^2 \frac{n(\log(n) + 2)}{\lambda_2(L_{\mathbf{M}})^2} \frac{1}{m}$$

where $D = 4k^2 e^{4b}$.

The proof of Theorem 7 is provided in Appendix C. The mean squared error upper bound in Theorem 7 corresponds that in Theorem 4 up to a constant factor. If the comparisons sets are unbiased, from (11), we have that $\lambda_2(L_{\mathbf{M}}) = (1 - 1/n)k(k - 1)$. Hence, the mean squared error upper bound in Theorem 7 depends on k only through the factor $1/(1 - 1/k)^2$, which decreases to value 1 with k in a diminishing returns fashion. This suggests that there is a limited dependence of the mean squared error on the size of comparison sets.

We now go on to establish a mean-squared upper bound for a class of Thurstone choice models. We will allow for comparison sets of different cardinalities taking values in a set K . We will admit the following conditions:

A1 There exists $A > 0$ such that for all $S \subseteq N$ with $|S| \in K$, all $y \in S$, all $i, j \in S$ with $i \neq j$, and all $\theta \in [-b, b]^n$,

$$\frac{\partial^2(-\log(p_{y,S}(\theta)))}{\partial \theta_i \partial \theta_j} \leq A \frac{\partial^2(-\log(p_{y,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} \leq 0$$

and, moreover, the following holds

$$\mathbf{E} \left[\frac{\partial^2(-\log(p_{y,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} \right] < 0.$$

A2 There exists $B > 0$ such that for all $S \subseteq N$ with $|S| \in K$, all $y \in S$, and all $\theta \in [-b, b]^n$,

$$\|\nabla p_{y,S}(\theta)\|_2 \leq B \|\nabla p_{y,S}(\mathbf{0})\|_2.$$

A3 There exists $C > 0$ such that for all $S \subseteq N$ with $|S| \in K$, all $y \in S$, and all $\theta \in [-b, b]^n$,

$$p_{y,S}(\theta) \geq C p_{y,S}(\mathbf{0}).$$

Condition **A1** ensures that $\nabla^2(-\log(p_{y,S}(\mathbf{0})))$ is a Laplacian matrix with non-negative weights, and that $\nabla^2(-\log(p_{y,S}(\theta))) \succeq A\nabla^2(-\log(p_{y,S}(\mathbf{0})))$ for all $\theta \in [-b, b]^n$. Condition **A1** also ensures that the expected value of $\nabla^2(-\log(p_{y,S}(\theta)))$ is a positive definite matrix where $\nabla^2(-\log(p_{y,S}(\mathbf{0})))$ is a random matrix when $|S| > 2$. Condition **A2** requires that $\|\nabla p_{y,S}(\theta)\|_2$ is bounded for all $\theta \in [-b, b]^n$, while condition **A3** ensures that the choice probabilities are not too much imbalanced. Conditions **A1** and **A2** may be seen as generalizations of conditions **P1** and **P2** for the case of pair comparisons.

Conditions **A1**, **A2** and **A3** can be easily shown to hold for the Luce choice model. For the Luce choice model, we have

$$\frac{\partial^2(-\log(p_{y,S}(\theta)))}{\partial\theta_i\partial\theta_j} = -\frac{1}{\beta^2}p_{i,S}(\theta)p_{j,S}(\theta)$$

hence, **A1** holds with $A = e^{-4b/\beta}$ and $\partial^2(-\log(p_{y,S}(\mathbf{0})))/\partial\theta_i\partial\theta_j = -1/(|S|\beta)^2$. Conditions **A2** and **A3** hold with $B = 4$ and $C = e^{-2b/\beta}$.

Note that constants A , B and C that appear in **A1**, **A2** and $\mathbf{A3}$, respectively, may depend on F , the cardinalities of comparison sets, and the parameter b , but are independent of any other parameters. In particular, these constants are independent of the number of observations. For any Thurstone choice model, the constants A , B , and C can be taken to have values arbitrarily near to value 1 by taking b small enough.

We next show an upper bound for the mean squared error of the maximum likelihood parameter estimator for a class of Thurstone choice models that satisfy the above stated conditions. Before we do that, we need to introduce some new definitions.

Definition 8 (weight function) *Let w^* be a function defined on positive integers greater than or equal to 2, we refer to as a weight function, which is defined by*

$$w^*(k) = \left(k \frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2, \text{ for } k = 2, 3, \dots \quad (17)$$

where

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx. \quad (18)$$

Notice that for the Luce choice model, $\partial p_k(\mathbf{0})/\partial x_1 = 1/(\beta k)^2$. Hence, in this case, $w^*(k) = 1/(\beta k)^2$. We will see later in Section 6 that for a broad class of Thurstone choice models, which includes well-known cases with noise according to either Gaussian or double-exponential distribution, $\partial p_k(\mathbf{0})/\partial x_1 = \Theta(1/k^2)$ and, hence, $w^*(k) = \Theta(1/k^2)$.

Definition 9 ($\gamma_{F,k}$ parameter) *Let $\gamma_{F,k}$ be a parameter defined by*

$$\frac{1}{\gamma_{F,k}} = k^3(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2. \quad (19)$$

We note that for any comparison set S of cardinality k and all $y \in S$, we have

$$\frac{1}{\gamma_{F,k}} = \|\nabla \log(p_{y,S}(\mathbf{0}))\|_2^2 = k(k-1)w^*(k)$$

which is discussed in more detail in the proof of Lemma 30. In particular, for the Luce choice model, we have $\gamma_{F,k} = (1 - 1/k)/\beta^2$.

Theorem 10 Assume **A1**, **A2** and **A3**, let $\sigma_{F,K}$ be such that $1/\gamma_{F,k} \leq \sigma_{F,K}$ for all $k \in K$, and $\lambda_2(L_{\overline{\mathbf{M}}_{w^*}}) \geq 32(\sigma_{F,K}/C)n \log(n)/m$. Then, with probability at least $1 - 3/n$,

$$\text{MSE}(\hat{\theta}, \theta^*) \leq 32D^2 \sigma_{F,K} \frac{n(\log(n) + 2)}{\lambda_2(L_{\overline{\mathbf{M}}_{w^*}})^2} \frac{1}{m}$$

where $D = B/(AC)$.

The proof of the theorem is provided in Appendix D. The main technical difference of the proof with respect to that of Theorem 7 is that $\nabla^2(-\ell(\theta))$ is a sum of random matrices. Every $\nabla^2(-\log(p_{y_t, S_t}(\theta)))$ is a random matrix for the following two reasons: (a) $\nabla^2(-\log(p_{y_t, S_t}(\theta)))$ depends on the randomly chosen y_t and (b) S_t is allowed to be a random set of items. We use the matrix Chernoff bound in the proof of Theorem 10.

We next show two corollaries of Theorem 10, which cover two interesting special cases.

Corollary 11 Suppose that all comparison sets are of identical cardinality of value $k \geq 2$, **A1**, **A2**, **A3** hold, and $\lambda_2(L_{\overline{\mathbf{M}}_{1/k^2}}) \geq 32(k-1)/(Ck)$. Then, with probability at least $1 - 3/n$,

$$\text{MSE}(\hat{\theta}, \theta^*) \leq 32D^2 \left(1 - \frac{1}{k}\right)^2 \gamma_{F,k} \frac{n(\log(n) + 2)}{\lambda_2(L_{\overline{\mathbf{M}}_{1/k^2}})^2} \frac{1}{m}$$

where $\gamma_{F,k}$ is given by (19), and \mathbf{M}_{1/k^2} is the weighted-adjacency matrix with the weight function $w(k) = 1/k^2$.

Corollary 12 Suppose that comparison sets are independent with each comparison set being a sample without replacement from the set of all items, conditions **A1**, **A2**, **A3** hold, and $m \geq 32(1 - 1/n)/Cn \log(n)$. Then, with probability at least $1 - 3/n$,

$$\text{MSE}(\hat{\theta}, \theta^*) \leq 32D^2 \left(1 - \frac{1}{n}\right)^2 \gamma_{F,k} \frac{n(\log(n) + 2)}{m}.$$

3.3 Lower Bound

In this section, we present a lower bound for the mean squared error of the maximum likelihood parameter estimator, which establishes minimax optimality of the established upper bounds. We define the following conditions:

A1' There exists $\tilde{A} > 0$ such that for all $S \subseteq N$ with $|S| \in K$, all $y \in S$, all $i, j \in S$ such that $i \neq j$, and all $\theta \in [-b, b]^n$, it holds

$$\frac{\partial^2(-\log(p_{y,S}(\theta)))}{\partial \theta_i \partial \theta_j} \geq \tilde{A} \frac{\partial^2(-\log(p_{y,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j}.$$

A3' There exists $\tilde{C} > 0$ such that for all $S \subseteq N$ with $|S| \in K$, all $y \in S$, and all $\theta \in [-b, b]^n$, it holds

$$p_{y,S}(\theta) \leq \tilde{C} p_{y,S}(\mathbf{0}).$$

Notice that, in particular, for the special case of F being a double-exponential distribution with parameter β , we have that **A1'** and **A3'** hold with $\tilde{A} = e^{4b/\beta}$ and $\tilde{C} = e^{2b/\beta}$.

Theorem 13 *Under conditions **A1'** and **A3'**, for any unbiased estimator $\hat{\theta}$, we have*

$$\mathbf{E}[\text{MSE}(\hat{\theta}, \theta^*)] \geq \frac{1}{\tilde{A}\tilde{C}} \left(\sum_{i=2}^n \frac{1}{\lambda_i(L_{\mathbf{M}_F})} \right) \frac{1}{m}.$$

The following two corollaries follow from the last theorem.

Corollary 14 *If all comparison sets are of cardinality $k \geq 2$, then any unbiased estimator $\hat{\theta}$ satisfies*

$$\mathbf{E}[\text{MSE}(\hat{\theta}, \theta^*)] \geq \frac{1}{\tilde{A}\tilde{C}} \left(1 - \frac{1}{k} \right) \gamma_{F,k} \left(\sum_{i=2}^n \frac{1}{\lambda_i(L_{\mathbf{M}})} \right) \frac{1}{m}.$$

Corollary 15 *If, in addition, each comparison set is drawn independently, uniformly at random from the set of all items, then any unbiased estimator $\hat{\theta}$ satisfies*

$$\mathbf{E}[\text{MSE}(\hat{\theta}, \theta^*)] \geq \frac{1}{\tilde{A}\tilde{C}} \left(1 - \frac{1}{n} \right)^2 \gamma_{F,k} \frac{n}{m}.$$

The last corollary implies that under the given assumptions, for the mean squared error to be smaller than a constant, it is necessary that the number of observed comparisons is $m = \Omega(\gamma_{F,k}n)$.

Proof of Theorem 13 We denote by $\text{cov}[Y]$ the covariance matrix of a multivariate random variable Y i.e.,

$$\text{cov}[Y] = \mathbf{E}[(Y - \mathbf{E}[Y])(Y - \mathbf{E}[Y])^\top].$$

The proof uses the Cramér-Rao inequality, which is stated as follows.

Lemma 16 (Cramér-Rao bound) *Let X be a multivariate random variable with distribution $p(x; \theta)$, for parameter $\theta \in \Theta$, and let $\psi : \Theta \rightarrow \mathbf{R}^r$ be a differentiable function. Then, for any unbiased estimator $\mathbf{T}(X) = (T_1(X), \dots, T_r(X))^\top$ of $\psi(\theta) = (\psi_1(\theta), \dots, \psi_r(\theta))^\top$, we have*

$$\text{cov}[\mathbf{T}(X)] \geq \frac{\partial \psi(\theta)}{\partial \theta} F^{-1}(\theta) \frac{\partial \psi(\theta)}{\partial \theta}^\top$$

where $\frac{\partial \psi(\theta)}{\partial \theta}$ is the Jacobian matrix of ψ and $F(\theta)$ is the Fisher information matrix given by

$$F(\theta) = \mathbf{E}[\nabla^2(-\log(p(X; \theta)))].$$

Let us define $\psi_i(\theta) = \theta_i - \frac{1}{n} \sum_{l=1}^n \theta_l$ for all $i = 1, 2, \dots, n$. Since $\sum_{i=1}^n \theta_i = 0$, we have $\sum_{i=1}^n \psi_i(\theta) = 0$. Note that we can write

$$\frac{\partial \psi(\theta)}{\partial \theta} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top. \quad (20)$$

Let $F(\theta)$ be the Fisher information matrix given by

$$F(\theta) = \sum_{t=1}^m \mathbf{E} [\nabla^2(-\log(p_{y_t, S_t}(\theta)))] . \quad (21)$$

By conditions **A1'** and **A3'**, and Lemma 36, we have

$$\begin{aligned} \mathbf{E} [\nabla^2(-\log(p_{y_t, S_t}(\theta)))] &= \sum_{y \in S_t} p_{y, S_t}(\theta) \nabla^2(-\log(p_{y, S_t}(\theta))) \\ &\preceq \sum_{y \in S_t} \frac{\tilde{C}}{|S_t|} \nabla^2(-\log(p_{y, S_t}(\theta))) \\ &\preceq \sum_{y \in S_t} \frac{\tilde{A}\tilde{C}}{|S_t|} \nabla^2(-\log(p_{y, S_t}(\mathbf{0}))) \\ &= \tilde{A}\tilde{C} \left(|S_t| \frac{\partial p_{|S_t|}(\mathbf{0})}{\partial x_1} \right)^2 L_{\mathbf{M}_{S_t}} \end{aligned} \quad (22)$$

where \mathbf{M}_S is a matrix in $\mathbf{R}^{n \times n}$ that has each element (i, j) such $\{i, j\} \subseteq S$ equal to 1 and all other elements equal to 0.

From (21) and (22),

$$F(\theta) \preceq \tilde{A}\tilde{C} \frac{m}{n} L_{\overline{\mathbf{M}}_F}. \quad (23)$$

Note that

$$\mathbf{E}[\|\hat{\theta} - \theta\|_2^2] = \text{trace}(\text{cov}[\mathbf{T}(X)]) = \sum_{i=1}^n \lambda_i(\text{cov}[\mathbf{T}(X)]).$$

By the Cramér-Rao bound and (23), we have

$$\begin{aligned} \frac{1}{n} \mathbf{E}[\|\hat{\theta} - \theta\|_2^2] &\geq \frac{1}{n} \sum_{i=1}^n \lambda_i \left(\frac{\partial \psi(\theta)}{\partial \theta} F^{-1}(\theta) \frac{\partial \psi(\theta)}{\partial \theta}^\top \right) \\ &= \frac{1}{n} \sum_{i=1}^{n-1} \lambda_i(\mathbf{U}^\top F^{-1}(\mathbf{0}) \mathbf{U}) \\ &= \frac{1}{n} \sum_{i=1}^{n-1} \frac{1}{\lambda_i(\mathbf{U}^\top F(\mathbf{0}) \mathbf{U})} \\ &\geq \frac{1}{\tilde{A}\tilde{C}} \sum_{i=1}^{n-1} \frac{1}{\lambda_i(\mathbf{U}^\top L_{\overline{\mathbf{M}}_F} \mathbf{U})} \\ &= \frac{1}{\tilde{A}\tilde{C}m} \sum_{i=2}^n \frac{1}{\lambda_i(L_{\overline{\mathbf{M}}_F})}. \end{aligned}$$

4. Rank-Breaking Parameter Estimation Methods

The maximum likelihood parameter estimation requires to find a maximizer of a log-likelihood function which for comparison sets of cardinality k has the form of a sum of $\binom{n}{k}$ elements in the worst case. For pair comparisons, this sum consists of at most n^2 elements. It is common for parameter estimators to be defined as maximizers of a pseudo log-likelihood function, which is defined as the log-likelihood function of pair comparisons deduced from the input observations under assumption that these pair comparisons are independent (which in general is false under a Thurstone choice model for comparison sets of three or more items). This is commonly referred as *rank breaking*. In what follows, we consider two different rank-breaking methods: (a) one that deduces $k - 1$ pair comparisons from a choice from a comparison set of cardinality k , we refer to as rank-breaking method ALL, and (b) one that deduces 1 pair comparison from a choice from a comparison set of cardinality k , we refer to as rank-breaking method ONE.

Rank-breaking method ALL This rank-breaking method deduces $k - 1$ pair comparisons from each comparison set of cardinality k . Specifically, for every comparison set S_t , the method uses pair comparisons between the chosen item y_t and each non-chosen item $v \in S_t \setminus \{y_t\}$. Notice that for any comparison set of three or more items, the pair comparisons selected from this set by the given rank-breaking method are not independent.

For the Luce choice model, the pseudo log-likelihood function is given by (6), which can be written in a more explicit form as follows

$$\ell_{k-1}(\theta) = \sum_{t=1}^m \sum_{v \in S_t \setminus \{y_t\}} \log \left(\frac{e^{\theta_{y_t}}}{e^{\theta_{y_t}} + e^{\theta_v}} \right).$$

We consider the maximum pseudo log-likelihood estimator $\hat{\theta}_{k-1} := \arg \max_{\theta \in \Theta} \ell_{k-1}(\theta)$.

Theorem 17 *If $\lambda_2(L_M) \geq 128(k-1)^2 e^{2b} n \log(n)/m$, then with probability at least $1 - 3/n$,*

$$\text{MSE}(\hat{\theta}_{k-1}, \theta^*) \leq D^2 \frac{n(\log(n) + 2)}{\lambda_2(L_M)^2} \frac{1}{m}.$$

where $D = 16\sqrt{2}\sqrt{k(k-1)^3}e^{2b}$.

The proof of Theorem 17 is given in Appendix E.

The mean squared error upper bound in Theorem 17 implies that the given parameter estimator is consistent. For any fixed size of a comparison set, the bound in Theorem 17 is equal to that in Theorem 7 up to a constant factor. Both bounds have the same scaling with parameter k .

Rank-breaking method ONE This rank-breaking method deduces 1 pair comparison from a comparison set of cardinality k . From each comparison set S_t , this rank-breaking methods selects a pair that consists of the chosen item y_t and an item z_t selected uniformly at random from the set of non-chosen items $S_t \setminus \{y_t\}$.

For the Luce choice model, the pseudo log-likelihood function is given by (7), which can be written in a more explicit form as follows

$$\ell_1(\theta) = \sum_{t=1}^m \log \left(\frac{e^{\theta_{y_t}}}{e^{\theta_{y_t}} + e^{\theta_{z_t}}} \right).$$

We consider the maximum pseudo log-likelihood estimator $\hat{\theta}_1 := \arg \max_{\theta \in \Theta} \ell_1(\theta)$.

Theorem 18 *If $\lambda_2(L_{\mathbf{M}}) \geq 8k(k-1)e^{2b}n \log(n)/m$, then with probability at least $1 - 3/n$,*

$$\text{MSE}(\hat{\theta}_1, \theta^*) \leq D^2 \frac{n(\log(n) + 2)}{\lambda_2(L_{\mathbf{M}})^2} \frac{1}{m}$$

where $D = 4k(k-1)e^{2b}$.

The proof of Theorem 18 is given in Appendix F.

It is noteworthy that the mean squared error upper bounds in Theorem 17 and Theorem 18 are equal up to a constant factor. Intuitively, one would expect that rank-breaking method **ALL** would yield a smaller mean squared error than rank-breaking method **ONE** because it uses more information from each observed choice. The reason why the two mean squared error upper bounds are equal up to a constant factor is as follows. When applying Lemma 1 we need to find an upper bound α for the norm of the gradient of the negative pseudo log-likelihood function and a lower bound β for the second-smallest eigenvalue of the Hessian matrix of the negative pseudo log-likelihood function. In our proofs, for the case of Theorem 18, we obtained α and β that scale with k as 1 and $1/k^2$, respectively. On the other hand, for the case of Theorem 17, we obtained α and β that scale with k as k and $1/k$, respectively. For both cases, it follows that the ratio α/β scales as k^2 .

5. Binary Classification

We now consider a Thurstone choice model \mathcal{T}_F where the strength parameter vector θ takes value in $\Theta = \{-b, b\}^n$, for a parameter $b > 0$. Here each individual strength parameter takes either a low or a high value. We say that each item is either of a low or a high class. We consider a binary classification problem, where the goal is to correctly classify all items with a prescribed probability of classification error. Let N_1 be the set of high class items and N_2 be the set of low class items. We shall consider the case when the total number of items is even and the number of high class items is equal to the number of low class items.

We consider a simple classification algorithm that uses *point scores* defined as follows: each item is associated with a point score equal to the number of comparison sets in which the given item is the chosen item. The algorithm outputs a classification of items with \hat{N}_1 and \hat{N}_2 denoting the sets of items classified to be high class or low class, respectively. The algorithm follows the following three steps: (a) for each item compute its point score, (b) sort the items in decreasing order of point scores, and (c) let \hat{N}_1 contain $n/2$ items with highest point scores and \hat{N}_2 contain remaining items. We refer to this algorithm as a *point score ranking method*.

Theorem 19 *Suppose that $b \leq 4/(k^2 \partial p_k(\mathbf{0})/\partial x_1)$ and*

$$b \max_{\mathbf{x} \in [-2b, 2b]^{k-1}} \|\nabla^2 p_k(\mathbf{x})\|_2 \leq \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (24)$$

Then, for every $\delta \in (0, 1]$, if

$$m \geq 64 \frac{1}{b^2} \left(1 - \frac{1}{k}\right) \gamma_{F,k} n(\log(n) + \log(1/\delta)) \quad (25)$$

then, the point score ranking method correctly classifies all items with probability at least $1 - \delta$.

The proof of Theorem 19 is provided in Appendix J.

The sufficient condition in (25) for the point score ranking method to correctly classify all the items with probability at least $1 - \delta$ is shown to be necessary up to a constant-factor for any classification algorithm, which is given in the following theorem.

Theorem 20 *Suppose that $b \leq 1/(6k^2 \partial p_k(\mathbf{0})/\partial x_1)$ and that condition (24) holds. Then, for every even $n \geq 16$ and $\delta \in (0, 1/4]$, for any algorithm to correctly classify all the items with probability at least $1 - \delta$, it is necessary that the following condition holds*

$$m \geq \frac{1}{62} \frac{1}{b^2} \left(1 - \frac{1}{k}\right) \gamma_{F,k} n(\log(n) + \log(1/\delta)).$$

The proof of Theorem 20 is given in Appendix K. In the proof, we use the statistical difference between the case when all the items are correctly classified and the case that an item is incorrectly classified. This proof strategy is motivated by that in Yun and Proutiere (2014) where it was used to analyze the classification error of the stochastic block model.

6. Discussion of Results

In this section, we discuss how the number of observations needed to attain a prescribed parameter estimation error depends on the cardinality of comparison sets.

In Section 3, we found that for a priori unbiased input comparisons, where each comparison set is of cardinality k and is drawn uniformly at random from the set of all items, the number of observations needed for the mean squared error to be within a prescribed tolerance is of the order $\gamma_{F,k}$, defined by (19). In Section 5, we found that this also so to ensure that the probability of classification error is within a prescribed tolerance.

In Table 1, we show the values of the parameter $\gamma_{F,k}$ for several special instances of the Thurstone choice model, along with the values of $\partial p_k(\mathbf{0})/\partial x_1$. From the expressions in Table 1, we observe that for all the cases but the case of uniform distribution of noise, $\gamma_{F,k}$ decreases with the cardinality of comparison sets, but in a slow manner according to a diminishing returns relation. In particular, observe that for both double-exponential and Laplace distribution of noise, $\gamma_{F,k} = \Theta(1)$ and for Gaussian distribution of noise $\gamma_{F,k} = O(1/k^\epsilon)$. On the other hand, for uniform distribution of noise, $\gamma_{F,k} = \Theta(1/k^2)$.

It is noteworthy that $\gamma_{F,k}$ satisfies the following bounds.

Table 1: The values of parameters for our examples of \mathcal{T}_F .

F	$\frac{\partial p_k(\mathbf{0})}{\partial x_1}$	$\gamma_{F,k}$
Gaussian	$O(\frac{1}{k^{2-\epsilon}})$	$\Omega(\frac{1}{k^{2\epsilon}})$
Double-exponential	$\frac{1}{\beta k^2}$	$\beta^2 \frac{k}{k-1}$
Laplace	$\frac{1-1/2^{k-1}}{\beta k(k-1)}$	$\beta^2 \frac{k-1}{k(1-1/2^{k-1})^2}$
Uniform	$\frac{1}{2a(k-1)}$	$4a^2 \frac{k-1}{k^3}$

Lemma 21 *For any cumulative distribution function F with density function f :*

1. *If f is even and continuously differentiable, then $\gamma_{F,k} = O(1)$.*
2. *If f is such that $f(x) \leq C$ for all $x \in \mathbf{R}$, for a constant $C > 0$, then $\gamma_{F,k} = \Omega(1/k^2)$.*

We observe that both double-exponential distribution and Laplace distribution of noise are extremal in the sense that they achieve the upper bound $\gamma_{F,k} = O(1)$. On the other hand, uniform distribution of noise is extremal in the sense that it achieves the lower bound $\gamma_{F,k} = \Omega(1/k^2)$.

7. Experimental Results

In this section, we present our experimental results using both simulations and real-world data. Our first goal is to provide experimental validation of the claim that the mean squared error can depend on the cardinality of comparison sets in different ways for different Thurstone choice models, which is suggested by our theory. Our second goal is to evaluate Fiedler value for different weighted-adjacency matrices observed in real-world data, which demonstrates that it can assume a wide range of values depending on the application scenario.

7.1 MSE versus Cardinality of Comparison Sets

We consider the following simulation experiment. We fix the number of items n and the number of observations m . We then run experiments for different values of the cardinality of comparison sets k . For each given value of parameter k , we generate comparison sets as independent uniform random sets of cardinality k from the set of all items. We then draw choices according to a Thurstone choice model \mathcal{T}_F for the value of parameter vector $\theta^* = \mathbf{0}$. For every fixed value of k , we run 100 repetitions to estimate the mean squared error. We do this for the distribution of noise according to a double-exponential distribution (Bradley-Terry model) and according to a uniform distribution, both with unit variance.

Figure 1 shows the results for the case of $n = 10$ and $m = 100$. The results clearly demonstrate that the mean squared error exhibits qualitatively different decay with the cardinality of comparison sets for the two Thurstone choice models under consideration. Our theoretical results in Section 3.2 suggest that the mean squared error should decrease with the cardinality of comparison sets as $1/(1-1/k)$ for the double-exponential distribution,

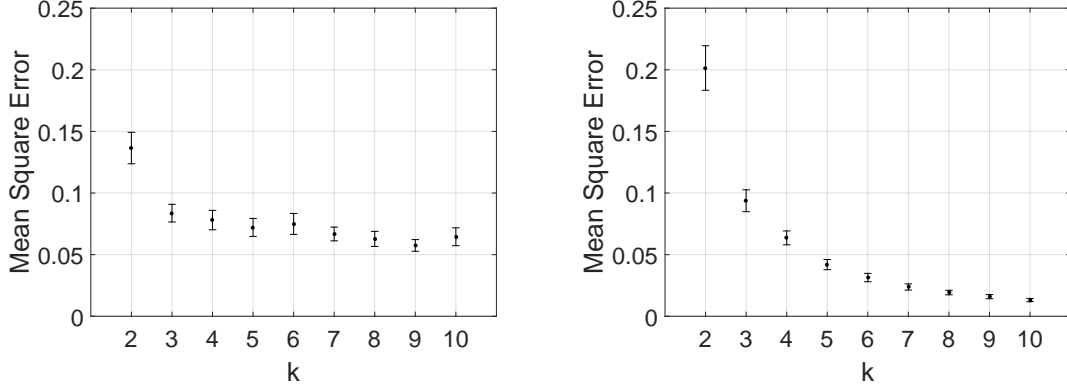


Figure 1: Mean squared error for two different Thurstone choice models \mathcal{T}_F : (left) double-exponential distribution of noise, and (right) uniform distribution of noise. The vertical bars denote 95% confidence intervals. The results demonstrate two qualitatively different relations between the mean squared error and the cardinality of comparison sets, which confirm the theory.

and as $1/k^2$ for the uniform distribution of noise. Observe that the latter two terms decrease with k to a strictly positive value and to zero value, respectively. The empirical results in Figure 1 confirm these claims.

7.2 Fiedler Values of Weighted-Adjacency Matrices

We found that Fiedler value of a weighted-adjacency matrix plays a key role in upper bounds on the mean squared error of parameter estimator in Section 3.1 and Section 3.2. Here we evaluate Fiedler value for different weighted-adjacency matrices of different schedules of comparisons. Throughout this section, we use the definition of a weighted-adjacency matrix in (9) with the weight function $w(k) = 1/k^2$. Our first two examples are representative of schedules in sport competitions, which are typically carefully designed by sport associations and exhibit a large degree of regularity. Our second two examples are representative of comparisons that are induced by user choices in the context of online services, which exhibit much more irregularity.

Sport competitions We consider the fixtures of games for the season 2014-2015 for (i) football Barclays premier league and (ii) basketball NBA league. In the Barclays premier league, there are 20 teams, each team plays a home and an away game with each other team; thus there are 380 games in total. In the NBA league, there are 30 teams, 1,230 regular

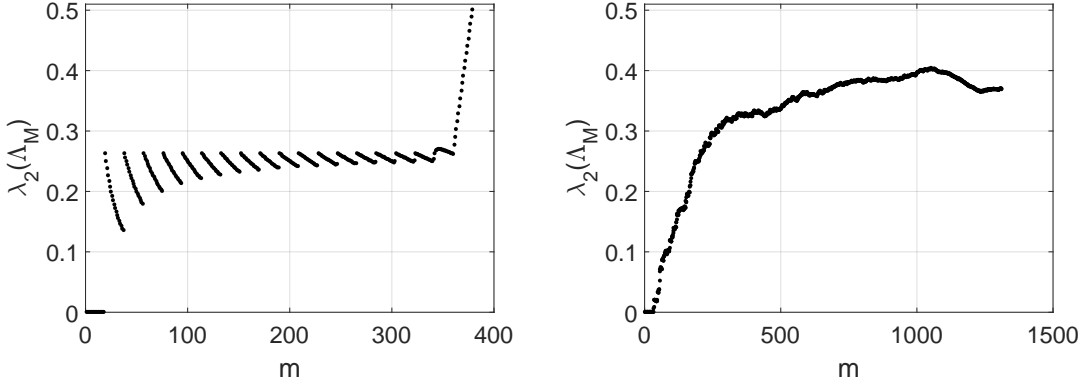


Figure 2: Fiedler value of the weighted-adjacency matrices for the game fixtures of two sports in the season 2014-2015: (left) football Barclays premier league, and (right) basketball NBA league.

games, and 81 playoff games.¹ We evaluate Fiedler value of weighted-adjacency matrices defined for first m matches of each season; see Figure 2.

For the Barclays premier league dataset, at the end of the season, the Fiedler value of the weighted-adjacency matrix is of value $n/[2(n-1)] \approx 1/2$. The schedule of matches is such that at the middle of the season, each team played against each other team exactly once, at which point the Fiedler value is $n/[4(n-1)] \approx 1/4$. The Fiedler value is of a strictly positive value after the first round of matches. For most part of the season, its value is near to $1/4$ and it grows to the highest value of approximately $1/2$ in the last round of the matches.

For the NBA league dataset, at the end of the season, the Fiedler value of the weighted-adjacency matrix is approximately 0.375. It grows more slowly with the number of games played than for the Barclays premier league; this is intuitive as the schedule of games is more irregular, with each team not playing against each other team the same number of times.

Crowdsourcing contests We consider participation of users in contests of two competition-based online labour platforms: (i) online platform for software development TopCoder and (ii) online platform for various kinds of tasks Tackn. We refer to coders in TopCoder and workers in Taskcn as users. We consider contests of different categories observed in year 2012. In both these systems, the participation in contests is according to choices made by users.

For each set of tasks of given category, we conduct the following analysis. We consider a conditioned dataset that consists only of a set of top- n users with respect to the number

1. The NBA league consists of two conferences, each with three divisions, and the fixture of games has to obey constraints on the number of games played between teams from different divisions.

of contests they participated in given year, and of all contests attended by at least two users from this set. We then evaluate Fiedler value of the weighted-adjacency matrix for parameter n ranging from 2 to the smaller of 100 or the total number of users. Our analysis reveals that the Fiedler value tends to decrease with n . This indicates that the larger the number of users included, the less connected the weighted-adjacency matrix is. See the top plots in Figure 3.

We also evaluated the smallest number of contests from the beginning of the year that is needed for the Fiedler value of the weighted-adjacency matrix to assume a strictly positive value. See the bottom plots in Figure 3. We observe that this threshold number of contests tends to increase with the number of top users considered. There are instances for which this threshold substantially increases for some number of the top users. This, again, indicates that the algebraic connectivity of the weighted-adjacency matrices tends to decrease with the number of top users considered.

8. Conclusion

The results of this paper elucidate how the estimation accuracy of a Thurstone choice model parameter depends on the given model and the structure of comparison sets. They show that a key factor is the algebraic connectivity of a weighted-adjacency matrix, which is specific to given model. It is shown that for a large class of Thurstone choice models, including well-known instances, there is a diminishing returns decrease of the estimation error with the cardinality of comparison sets at a slow rate for comparison sets of three or more items.

The results provide guidelines to the designers of competition schedules such as to ensure that a schedule has a well-connected weighted-adjacency matrix and to expect limited estimation accuracy gains by enlarging the size of comparison sets.

References

- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. method of paired comparisons. *Biometrika*, 39(3/4):324–345, Dec 1952.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: II. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4):502–537, Dec 1954.
- Arpad E. Elo. *The Rating of Chessplayers*. Ishi Press International, 1978.
- Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.
- Miroslav Fiedler. Laplacian of graphs and algebraic connectivity. *Combinatorics and Graph Theory*, 25:57–70, 1989.
- Thore Graepel, Tom Minka, and Ralf Herbrich. Trueskill(tm): A bayesian skill rating system. In *Proc. of NIPS 2006*, volume 19, pages 569–576, 2006.

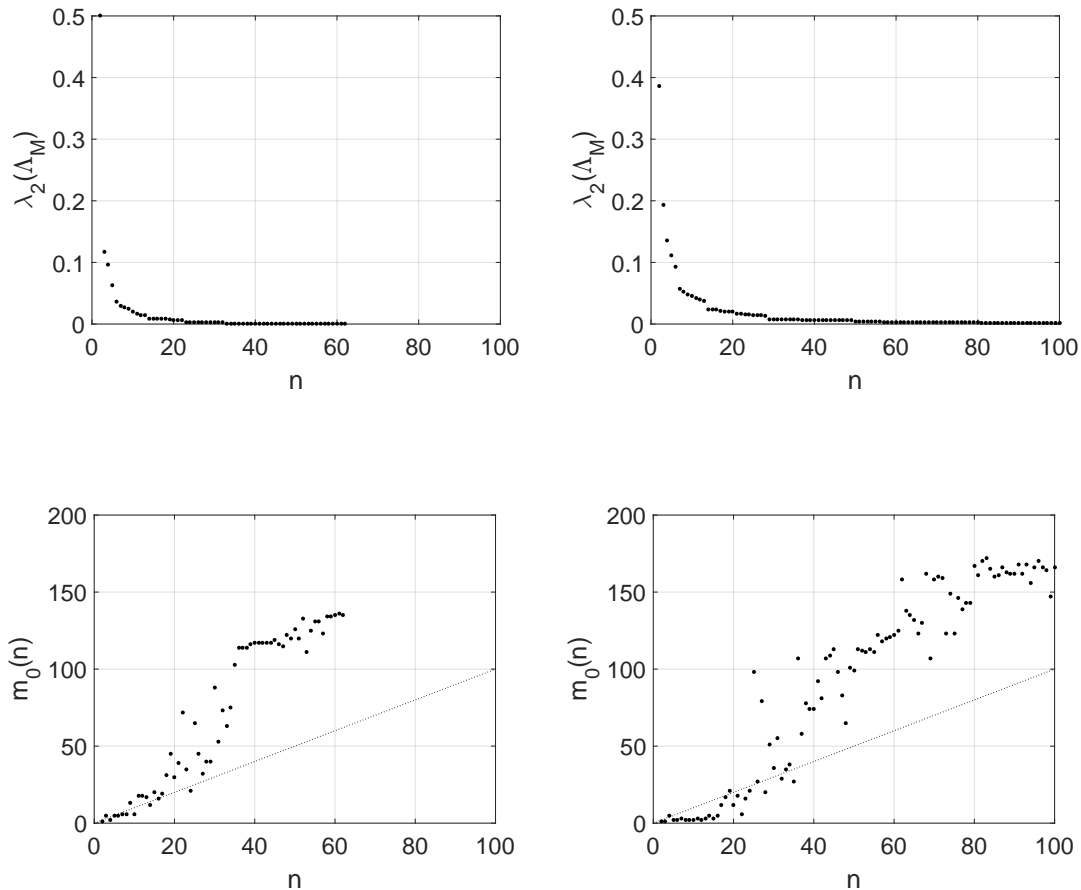


Figure 3: (Left) Topcoder data restricted to top- n coders and (Right) same as left but for Taskcn, for Design and Website task categories, respectively. The top plots show the Fiedler value and the bottom plots show the minimum number of contests to observe a strictly positive Fiedler value.

Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Proc. of NIPS 2014*, pages 1475–1483, 2014.

Thomas P. Hayes. A large-deviation inequality for vector-valued martingales. 2003. URL <http://www.cs.unm.edu/~hayes/papers/VectorAzuma/VectorAzuma20030207.pdf>.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

- David R Hunter. Mm algorithms for generalized bradley-terry models. *Annals of Statistics*, pages 384–406, 2004.
- Ashih Khetan and Sewoong Oh. Computational statistical tradeoffs in learning to rank. In *Proc. of NIPS 2016*, 2016a.
- Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016b.
- R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.
- Albert Maydeau-Olivares. Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, 64(3):325–340, 1999.
- Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. In *Advances in Neural Information Processing Systems*, pages 172–180, 2015.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, New York, 2 edition, 1989.
- Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Proc. of NIPS 2012*, pages 2483–2491, 2012.
- J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.
- Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.
- Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proc. of ICML 2014*, pages 118–126, 2014.
- Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *J. Mach. Learn. Res.*, 17(1):2049–2095, January 2016.
- Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.
- Azari Soufiani, David Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 360–368, 2014.
- Hossein Azari Soufiani, William Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Proc. of NIPS 2013*, 2013.

- Hal Stern. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1):103–117, 1992.
- L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(2):273–286, 1927.
- Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.
- Milan Vojnović. *Contest Theory: Incentive Mechanisms and Ranking Methods*. Cambridge University Press, 2016.
- John I. Yellott. The relationship between Luce’s choice axiom, Thurstone’s theory of comparative judgement and the double exponential distribution. *Journal of Mathematical Psychology*, 15:109–144, 1977.
- Se-Young Yun and Alexandre Proutiere. Community detection via random and adaptive sampling. In *COLT*, pages 138–175, 2014.
- E. Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Math. Z.*, 29:436–460, 1929.

Appendix A. Background Material

Location of Eigenvalues We make note of the well-known Geršgorin circles theorem, which we state as at the following lemma:

Lemma 22 *Let $\mathbf{A} \in \mathbf{R}^{n \times n}$, then all eigenvalues of \mathbf{A} are located in the union of n discs*

$$\cup_{i=1}^n \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{i,j}| \right\}.$$

Properties of Positive Definite and Laplacian Matrices A symmetric matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$ is said to be *positive semidefinite* if $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ for all nonzero $\mathbf{x} \in \mathbf{R}^n$. If the inequality is replaced with strict inequality, \mathbf{A} is said to be *positive definite*.

Each eigenvalue of a positive definite matrix is a positive real number. Each eigenvalue of a positive semidefinite matrix is a nonnegative real number.

For two matrices $\mathbf{A}, \mathbf{B} \in \mathbf{R}^{n \times n}$, we write $\mathbf{A} \succeq \mathbf{B}$ for the *positive semidefinite ordering*, which means that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. Similarly, we write $\mathbf{A} \succ \mathbf{B}$ for the *positive definite ordering*, which means that $\mathbf{A} - \mathbf{B}$ is a positive definite matrix. Note that $\mathbf{A} \succeq \mathbf{0}$ means that \mathbf{A} is a positive semidefinite matrix, and $\mathbf{A} \succ \mathbf{0}$ means that \mathbf{A} is a positive definite matrix.

We note the following ordering relations for eigenvalues of two positive definite matrices that satisfy the positive semidefinite ordering (e.g., see Corollary 7.7.4 Horn and Johnson (1985)).

Lemma 23 *For any two positive definite matrices $\mathbf{A}, \mathbf{B} \in \mathbf{R}^{n \times n}$ such that $\mathbf{A} \succeq \mathbf{B}$, $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for all $i = 1, 2, \dots, n$.*

For any $\mathbf{A} \in \mathbf{R}^{n \times n}$ with zero diagonal such that $\mathbf{A} \geq \mathbf{0}$, the Laplacian matrix $L_{\mathbf{A}}$ is positive semidefinite. This follows from the localization of eigenvalues by the Geršgorin circles theorem.

Lemma 24 *If $\mathbf{A} \in \mathbf{R}^n$ is a symmetric matrix with zero diagonal, then*

- (i) $\lambda_1(L_{\mathbf{A}}) = 0$, and
- (ii) $\lambda_i(L_{\mathbf{A}}) = \lambda_{i-1}(\mathbf{U}^\top L_{\mathbf{A}} \mathbf{U})$, for $i \in \{2, 3, \dots, n\}$

where $\mathbf{U} \in \mathbf{R}^{n \times n-1}$ is any matrix such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$.

We shall use the following property of real symmetric matrices:

Lemma 25 *If $\mathbf{A}, \mathbf{B} \in \mathbf{R}^{n \times n}$ are real symmetric matrices with zero diagonals such that $\mathbf{B} \geq \mathbf{A}$ where the inequality holds element-wise, then $L_{\mathbf{B}} \succeq L_{\mathbf{A}}$.*

Chernoff Tail Bounds The following bounds follow from the Chernoff bound:

Lemma 26 *Suppose that X is a sum of m independent Bernoulli random variables each with mean p , then if $q \leq p \leq 2q$,*

$$\mathbf{P}[X \leq qm] \leq \exp\left(-\frac{(q-p)^2}{4q}m\right) \quad (26)$$

and, if $p \leq q$,

$$\mathbf{P}[X \geq qm] \leq \exp\left(-\frac{(q-p)^2}{4q}m\right). \quad (27)$$

Proof We prove only prove (27) as (26) follows by similar arguments. By the Chernoff's bound, for every $s > 0$,

$$\begin{aligned} \mathbf{P}[X \geq qm] &\leq e^{-sqm} \mathbf{E}[e^{sX}] \\ &= e^{-sqm} (1 - p + pe^s)^m \\ &= e^{-mh(s)} \end{aligned}$$

where $h(s) = qs - \log(1 - p + pe^s)$. Since $\log(1 - x) \leq -x$ for all $x \in \mathbf{R}$, we have $h(s) \geq qs + p - pe^s$. Take $s = s^* := \log(q/p)$ to obtain $h(s^*) \geq q \log(q/p) + p - q$.

Now, let $\epsilon = q - p$, and note that $q \log(q/p) + p - q := g(\epsilon)$ where $g(\epsilon) = q \log(q/(q - \epsilon)) - \epsilon$. Since $g'(\epsilon) = q/(q - \epsilon) - 1 = \epsilon/(q - \epsilon) \geq \epsilon/(2q)$, we have $g(\epsilon) = \int_0^\epsilon g'(x) dx \geq \epsilon^2/(4q)$.

Hence, it follows that $h(s^*) \geq (p - q)^2/(4q)$, and, thus

$$\mathbf{P}[X \geq qm] \leq \exp\left(-\frac{1}{4q}(p - q)^2\right).$$

■

Appendix B. Proof of Lemma 1

Let $\Delta = \hat{\mathbf{x}} - \mathbf{x}^*$. By the Taylor expansion, we have

$$g(\hat{\mathbf{x}}) \geq g(\mathbf{x}^*) + \nabla g(\mathbf{x}^*)^\top \Delta + \frac{1}{2} \min_{\alpha \in [0,1]} \Delta^\top \nabla^2 g(\mathbf{x}^* + \alpha \Delta) \Delta. \quad (28)$$

Since $g(\hat{\mathbf{x}}) \leq g(\mathbf{x}^*)$, we have

$$\min_{\alpha \in [0,1]} \Delta^\top \nabla^2 g(\mathbf{x}^* + \alpha \Delta) \Delta \leq -\nabla g(\mathbf{x}^*)^\top \Delta.$$

Hence,

$$\min_{\mathbf{x} \in \mathcal{X}} \Delta^\top \nabla^2 g(\mathbf{x}) \Delta \leq 2 \|\nabla g(\mathbf{x}^*)\|_2 \|\Delta\|_2. \quad (29)$$

Fix an arbitrary $\mathbf{x} \in \mathcal{X}$. From condition (i) it follows that $\nabla^2 g(\mathbf{x})$ has eigenvalue 0 with eigenvector $\mathbf{1}$. Combining with condition (ii), we have

$$0 = \lambda_1(\nabla^2 g(\mathbf{x})) < \lambda_2(\nabla^2 g(\mathbf{x})) \leq \dots \leq \lambda_n(\nabla^2 g(\mathbf{x})). \quad (30)$$

Let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n] \in \mathbf{R}^{n \times n}$ where $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ are orthonormal eigenvectors of $\nabla^2 g(\mathbf{x})$, which correspond to eigenvalues $\lambda_1(\nabla^2 g(\mathbf{x})), \lambda_2(\nabla^2 g(\mathbf{x})), \dots, \lambda_n(\nabla^2 g(\mathbf{x}))$, respectively. Note that

$$\mathbf{u}_1^\top \Delta = 0. \quad (31)$$

Let $\Lambda = \text{diag}(\lambda_1(\nabla^2 g(\mathbf{x})), \lambda_2(\nabla^2 g(\mathbf{x})), \dots, \lambda_n(\nabla^2 g(\mathbf{x})))$.

We have the following relations:

$$\begin{aligned} \Delta^\top \nabla^2 g(\mathbf{x}) \Delta &= \Delta^\top \mathbf{U} \Lambda \mathbf{U}^\top \Delta \\ &= \sum_{i=1}^n \lambda_i(\nabla^2 g(\mathbf{x})) |(\mathbf{U}^\top \Delta)_i|^2 \\ &= \sum_{i=1}^n \lambda_i(\nabla^2 g(\mathbf{x})) |\mathbf{u}_i^\top \Delta|^2 \\ &= \sum_{i=2}^n \lambda_i(\nabla^2 g(\mathbf{x})) |\mathbf{u}_i^\top \Delta|^2 \\ &\geq \lambda_2(\nabla^2 g(\mathbf{x})) \sum_{i=2}^n |\mathbf{u}_i^\top \Delta|^2 \\ &= \lambda_2(\nabla^2 g(\mathbf{x})) \sum_{i=1}^n |\mathbf{u}_i^\top \Delta|^2 \\ &= \lambda_2(\nabla^2 g(\mathbf{x})) \|\Delta\|_2^2 \end{aligned}$$

where we use the properties in (30) and (31).

Hence, it follows that

$$\min_{\mathbf{x} \in \mathcal{X}} \Delta^\top \nabla^2 g(\mathbf{x}) \Delta \geq \|\Delta\|_2^2 \min_{\mathbf{x} \in \mathcal{X}} \lambda_2(\nabla^2 g(\mathbf{x})). \quad (32)$$

By combining (32) and (29), we conclude the proof of the lemma.

Appendix C. Proof of Theorem 7

The log-likelihood function in (3) can be written in the following more explicit form:

$$\ell(\theta) = \sum_{t=1}^m \log \left(\frac{e^{\theta_{y_t}}}{\sum_{v \in S_t} e^{\theta_v}} \right). \quad (33)$$

Since $\nabla^2(-\ell(\theta))\mathbf{1} = \mathbf{0}$, for all $\theta \in \mathbf{R}^n$, under assumption that $\min_{\theta \in \Theta} \lambda_2(-\nabla^2 \ell(\theta)) > 0$, the upper bound in Lemma 1 holds. This combined with the following two lemmas yields the statement of the theorem.

Lemma 27 *The following lower bound holds:*

$$\min_{\theta \in \Theta} \lambda_2(\nabla^2(-\ell(\theta))) \geq \frac{1}{k^2 e^{4b}} \frac{m}{n} \lambda_2(L_{\mathbf{M}}). \quad (34)$$

Proof From (33), we have for $i, j \in \{1, 2, \dots, n\}$ such that $j \neq i$,

$$\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} = \sum_{t: \{i, j\} \subseteq S_t} p_{i, S_t}(\theta) p_{j, S_t}(\theta) \text{ and } \frac{\partial^2 \ell(\theta)}{\partial \theta_i^2} = - \sum_{j \neq i} \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}.$$

Since for all $i \neq j$,

$$\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} \geq \frac{1}{k^2 e^{4b}} m_{i, j}$$

we have that $\nabla^2(-\ell(\theta)) - \frac{1}{k^2 e^{4b}} \frac{m}{n} L_{\mathbf{M}}$ is a Laplacian matrix of a matrix with nonnegative elements. Every such Laplacian matrix is positive semidefinite, hence

$$\nabla^2(-\ell(\theta)) \succeq \frac{1}{k^2 e^{4b}} \frac{m}{n} L_{\mathbf{M}}. \quad (35)$$

From (35), we conclude (34). ■

Lemma 28 *With probability at least $1 - 2/n$,*

$$\|\nabla(-\ell(\theta^*))\|_2 \leq 2\sqrt{m(\log(n) + 2)}. \quad (36)$$

Proof From (33), we have

$$\nabla(-\ell(\theta)) = \sum_{t=1}^m \nabla(-\log(p_{y_t, S_t}(\theta))) \quad (37)$$

where $p_{y_t, S_t}(\theta) = e^{\theta_{y_t}} / \sum_{v \in S_t} e^{\theta_v}$.

It is straightforward to derive that $\nabla(-\log(p_{y_t, S_t}(\theta)))$ has elements given by

$$\frac{\partial(-\log(p_{y_t, S_t}(\theta)))}{\partial \theta_i} = \begin{cases} -(1 - p_{i, S_t}(\theta)), & \text{if } i = y_t \\ p_{i, S_t}(\theta), & \text{if } i \in S_t \setminus \{y_t\} \\ 0, & \text{otherwise.} \end{cases} \quad (38)$$

From (37), $\nabla \ell(\theta)$ is a sum of independent random vectors in \mathbf{R}^n that satisfy for all $t \in \{1, 2, \dots, m\}$,

$$\mathbf{E}[\nabla(-\log(p_{y_t, S_t}(\theta^*)))] = \mathbf{0} \quad (39)$$

and

$$\|\nabla(-\log(p_{y_t, S_t}(\theta^*))\|_2 \leq \sqrt{2}. \quad (40)$$

The last two relations are easy to establish using (38) as follows. Equation (39) holds because for every $i \in N$,

$$\mathbf{E} \left[\frac{\partial(-\log(p_{y_t, S_t}(\theta^*)))}{\partial \theta_i} \right] = -p_{i, S_t}(\theta^*)(1 - p_{i, S_t}(\theta^*)) + (1 - p_{i, S_t}(\theta^*))p_{i, S_t}(\theta^*) = 0.$$

Equation (39) follows from

$$\|\nabla(-\log(p_{y_t, S_t}(\theta^*)))\|_2^2 = (1 - p_{y_t, S_t}(\theta^*))^2 + \sum_{j \in S_t \setminus \{y_t\}} p_{j, S_t}(\theta^*)^2 \leq 2.$$

By the vector Azuma-Hoeffding bound in Lemma 2, we have

$$\mathbf{P}[\|\nabla(-\ell(\theta^*))\|_2 \geq 2\sqrt{m(\log(n) + 2)}] \leq \frac{2}{n}.$$

which completes the proof of (36). ■

Appendix D. Proof of Theorem 10

Since $\nabla^2(-\ell(\theta))$ is a Laplacian matrix, by condition **A1**,

$$\nabla^2(-\ell(\theta)) \succeq A\nabla^2(-\ell(\mathbf{0})) \text{ for all } \theta \in [-b, b]^n.$$

Hence, in particular,

$$\min_{\theta \in [-b, b]^n} \lambda_2(-\ell(\theta)) \geq A\lambda_2(\nabla^2(-\ell(\mathbf{0}))). \quad (41)$$

We have the following two lemmas.

Lemma 29 *If $\lambda_2(L_{\overline{\mathbf{M}}_{w^*}}) \geq 32(\sigma_{F,K}/C)(n \log(n))/m$, then with probability at least $1 - 1/n$,*

$$\lambda_2(\nabla^2(-\ell(\mathbf{0}))) \geq \frac{1}{2}C\frac{m}{n}\lambda_2(L_{\overline{\mathbf{M}}_{w^*}}).$$

Proof $\nabla^2(-\ell(\mathbf{0}))$ is a sum of independent random matrices given by

$$\nabla^2(-\ell(\mathbf{0})) = \sum_{t=1}^m \nabla^2(-\log(p_{y_t, S_t}(\mathbf{0}))).$$

We have the following relations:

$$\begin{aligned} \mathbf{E} [\nabla^2(-\ell(\mathbf{0}))] &= \sum_{t=1}^m \mathbf{E} [\nabla^2(-\log(p_{y_t, S_t}(\mathbf{0})))] \\ &= \sum_{t=1}^m \sum_{y \in S_t} p_{y, S_t}(\theta^*) \nabla^2(-\log(p_{y, S_t}(\mathbf{0}))) \end{aligned}$$

$$\begin{aligned}
 &\succeq C \sum_{t=1}^m \sum_{y \in S_t} \frac{1}{|S_t|} \nabla^2(-\log(p_{y,S_t}(\mathbf{0}))) \\
 &= C \sum_{t=1}^m \sum_{y \in S_t} \frac{1}{|S_t|} L_{\mathbf{M}_{S_t}} \left(|S_t| \frac{\partial p_{|S_t|}(\mathbf{0})}{\partial x_1} \right)^2 \\
 &= C \frac{m}{n} L_{\overline{\mathbf{M}}_F}
 \end{aligned}$$

where we use Lemma 36.

Hence,

$$\lambda_2(\mathbf{E}[\nabla^2(-\ell(\mathbf{0}))]) \geq C \frac{m}{n} \lambda_2(L_{\overline{\mathbf{M}}_F}). \quad (42)$$

By Lemma 37, for all $t \in \{1, 2, \dots, m\}$,

$$\|\nabla^2 \log(p_{y_t, S_t}(\mathbf{0}))\|_2 \leq \frac{2}{\gamma_{F, |S_t|}} \leq 2\sigma_{F, K}. \quad (43)$$

Using the matrix Chernoff bound with $\epsilon = 1/2$, (42) and (43), we conclude the statement of the lemma. \blacksquare

Lemma 30 *With probability at least $1 - 2/n$,*

$$\|\nabla \ell(\theta^*)\|_2 \leq B \sqrt{\sigma_{F, K}} \sqrt{2m(\log(n) + 2)}.$$

Proof For every $S \subseteq N$ and $i, j \in S$ such that $j \neq i$,

$$\frac{\partial \log(p_{j,S}(\theta))}{\partial \theta_i} = \frac{1}{p_{j,S}(\theta)} \frac{\partial p_{j,S}(\theta)}{\partial \theta_i}. \quad (44)$$

and

$$\frac{\partial \log(p_{i,S}(\theta))}{\partial \theta_i} = -\frac{1}{p_{i,S}(\theta)} \sum_{v \in S \setminus \{i\}} \frac{\partial p_{v,S}(\theta)}{\partial \theta_i}. \quad (45)$$

From (44) and (45), for every $t \in \{1, 2, \dots, m\}$,

$$\mathbf{E}[\nabla \log p_{y_t, S_t}(\theta^*)] = \mathbf{0}. \quad (46)$$

From (44) and (45), for every $S \subseteq N$ such that $|S| = k \geq 2$ and $i, j \in S$ such that $i \neq j$,

$$\begin{aligned}
 \frac{\partial \log(p_{j,S}(\mathbf{0}))}{\partial \theta_i} &= k \frac{\partial p_k(\mathbf{0})}{\partial x_1} \\
 \frac{\partial \log(p_{i,S}(\mathbf{0}))}{\partial \theta_i} &= k(k-1) \frac{\partial p_k(\mathbf{0})}{\partial x_1}.
 \end{aligned}$$

Hence, for every $t \in \{1, 2, \dots, m\}$ such that $|S_t| = k$

$$\|\nabla \log(p_{y_t, S_t}(\mathbf{0}))\|_2^2 = k^3(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 = \frac{1}{\gamma_{F, k}}. \quad (47)$$

By condition **A3** and (47), for every $t \in \{1, 2, \dots, m\}$,

$$\|\nabla \log p_{y_t, S_t}(\theta^*)\|_2^2 \leq B^2 \|\nabla \log p_{y_t, S_t}(\mathbf{0})\|_2^2 \leq B^2 \sigma_{F,K}. \quad (48)$$

Using (46) and (48) and the vector Azuma-Hoeffding bound in Lemma 2, with probability at least $1 - 2/n$,

$$\|\nabla \ell(\theta^*)\|_2 \leq B \sqrt{\sigma_{F,K}} \sqrt{2m(\log(n) + 2)}.$$

■

The negative log-likelihood function satisfies the bound in Lemma 1. This combined with relation (41), Lemma 29 and Lemma 30 implies that if $\lambda_2(L_{\overline{\mathbf{M}}_{w^*}}) \geq 32(\sigma_{F,K}/C)n \log(n)/m$, then with probability at least $1 - 3/n$,

$$\frac{1}{n} \|\hat{\theta} - \theta^*\|_2^2 \leq 32 \left(\frac{B}{AC} \right)^2 \sigma_{F,K} \frac{n(\log(n) + 2)}{\lambda_2(L_{\overline{\mathbf{M}}_{w^*}})^2} \frac{1}{m}$$

which proves the theorem.

Appendix E. Proof of Theorem 17

Under condition that $\lambda_2(\nabla^2(-\ell_{k-1}(\theta))) > 0$, the negative pseudo log-likelihood function satisfies the bound in Lemma 1. This, combined with the following two lemmas implies the statement of the theorem.

Lemma 31 *If $\lambda_2(L_{\mathbf{M}}) \geq 8k(k-1)e^{2b}n \log(n)/m$, then with probability at least $1 - 1/n$,*

$$\min_{\theta \in \Theta} \lambda_2(\nabla^2(-\ell_{k-1}(\theta))) \geq \frac{1}{4ke^{2b}} \frac{m}{n} \lambda_2(L_{\mathbf{M}}).$$

Proof We will establish the lemma by using the matrix Chernoff bound in Lemma 3 as follows. Note that $\nabla^2(-\ell_{k-1}(\theta))$ is a sum independent random matrices given by:

$$\nabla^2(-\ell_{k-1}(\theta)) = \sum_{t=1}^m \nabla^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta)) \right).$$

The nonzero elements of $\nabla^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta)) \right)$ are, for every $i, j \in S_t$ such that $i \neq j$,

$$\frac{\partial^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta)) \right)}{\partial \theta_i \partial \theta_j} = \begin{cases} -p_{i,j}(\theta)(1 - p_{i,j}(\theta)), & \text{if } \{i, j\} \subseteq S_t, y_t \in \{i, j\} \\ 0, & \text{otherwise} \end{cases} \quad (49)$$

and

$$\frac{\partial^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right)}{\partial \theta_i^2} = - \sum_{j \in S_t \setminus \{i\}} \frac{\partial^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right)}{\partial \theta_i \partial \theta_j}.$$

From (49), we have

$$\begin{aligned} \mathbf{E} \left[\frac{\partial^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right)}{\partial \theta_i \partial \theta_j} \right] &= -p_{i,j}(\theta)(1 - p_{i,j}(\theta))(p_{i,S_t}(\theta) + p_{j,S_t}(\theta)) \\ &= -\frac{e^{\theta_i} + e^{\theta_j}}{\sum_{v \in S_t} e^{\theta_v}} \frac{e^{\theta_i} e^{\theta_j}}{(e^{\theta_i} + e^{\theta_j})^2} \\ &= -\frac{1}{\sum_{v \in S_t} e^{\theta_v}} \frac{1}{e^{-\theta_i} + e^{-\theta_j}} \\ &\leq -\frac{1}{2ke^{2b}}. \end{aligned}$$

Hence, we have

$$\mathbf{E} [\nabla^2(-\ell_{k-1}(\theta))] \succeq \frac{1}{2ke^{2b}} \frac{m}{n} L_{\mathbf{M}}.$$

and, in particular,

$$\lambda_2(\mathbf{E} [\nabla^2(-\ell_{k-1}(\theta))]) \geq \frac{1}{2ke^{2b}} \frac{m}{n} \lambda_2(L_{\mathbf{M}}). \quad (50)$$

To apply the matrix Chernoff bound in Lemma 3, we use the following identities that follow by Lemma 24,

$$\lambda_2(\nabla^2((-\ell_{k-1}(\theta)))) = \lambda_1(\mathbf{U}^\top \nabla^2((-\ell_{k-1}(\theta))) \mathbf{U}) \quad (51)$$

and

$$\lambda_2(\nabla^2(\mathbf{E}[(-\ell_{k-1}(\theta))])) = \lambda_1(\mathbf{U}^\top \nabla^2(\mathbf{E}[(-\ell_{k-1}(\theta))]) \mathbf{U}) \quad (52)$$

and the following fact:

$$\left\| \mathbf{U}^\top \nabla^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right) \mathbf{U} \right\|_2 \quad (53)$$

$$\begin{aligned} &\leq \left\| \mathbf{U}^\top \right\|_2 \left\| \nabla^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right) \right\|_2 \|\mathbf{U}\|_2 \\ &= \left\| \nabla^2 \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right) \right\|_2 \\ &\leq \max_i \left| \frac{\partial^2}{\partial \theta_i^2} \left(\sum_{v \in S_t \setminus \{y_t\}} \log(p_{y_t,v}(\theta)) \right) \right| + \sum_{j \neq i} \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left(\sum_{v \in S_t \setminus \{y_t\}} \log(p_{y_t,v}(\theta)) \right) \right| \\ &= 2 \max_i \sum_{j \neq i} p_{i,j}(\theta)(1 - p_{i,j}(\theta)) 1_{\{i,j\} \subseteq S_t, y_t \in \{i,j\}} \quad (54) \\ &\leq \frac{1}{2}(k-1). \quad (55) \end{aligned}$$

where the first equation follows by $\|\mathbf{U}^\top\|_2 = \|\mathbf{U}\|_2 = 1$, and the second inequality is by the Geršgorin circles theorem (Lemma 22).

$$\begin{aligned}
 & \mathbf{P}[\lambda_2(\nabla^2(-\ell_{k-1}(\theta))) \leq \frac{1}{4ke^{2b}} \frac{n}{m} \lambda_2(L_{\mathbf{M}})] \\
 & \leq \mathbf{P}[\lambda_2(\nabla^2(-\ell_{k-1}(\theta))) \leq \frac{1}{2} \lambda_2(\nabla^2 \mathbf{E}[(-\ell_{k-1}(\theta))])] \\
 & = \mathbf{P}[\lambda_1(\mathbf{U}^\top \nabla^2(-\ell_{k-1}(\theta)) \mathbf{U}) \leq \frac{1}{2} \lambda_1(\mathbf{U}^\top \nabla^2 \mathbf{E}[(-\ell_{k-1}(\theta))] \mathbf{U})] \\
 & \leq ne^{-\frac{\lambda_1(\mathbf{U}^\top \nabla^2 \mathbf{E}[(-\ell_{k-1}(\theta))] \mathbf{U})}{4^{\frac{k-1}{2}}}} \\
 & \leq ne^{-\frac{\lambda_2(L_{\mathbf{M}})m}{4k(k-1)e^{2b}n}} \\
 & \leq \frac{1}{n}
 \end{aligned}$$

where the first inequality is by (50), the equality is by (51) and (52), the second inequality is by the matrix Chernoff bound in Lemma 3 and (55), the third inequality is by (50), and the last inequality is by the condition $\lambda_2(L_{\mathbf{M}}) \geq 8k(k-1)e^{2b}n \log(n)/m$. \blacksquare

Lemma 32 *With probability at least $1 - 2/n$,*

$$\|\nabla(-\ell_{k-1}(\theta^*))\|_2 \leq 2\sqrt{k(k-1)m(\log(n) + 2)}. \quad (56)$$

Proof $\nabla(-\ell_{k-1}(\theta))$ is a sum of independent random vectors given by:

$$\nabla(-\ell_{k-1}(\theta)) = \sum_{t=1}^m \nabla \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta)) \right).$$

It is straightforward to show that $\nabla \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta)) \right)$ has the elements given by

$$\frac{\partial \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta)) \right)}{\partial \theta_i} = \begin{cases} -\sum_{v \in S_t \setminus \{y_t\}} p_{v, y_t}(\theta), & \text{if } i = y_t \\ p_{i, y_t}(\theta), & \text{if } i \in S_t \setminus \{y_t\} \\ 0, & \text{otherwise} \end{cases} \quad (57)$$

For every $t \in \{1, 2, \dots, m\}$ and $i \in \{1, 2, \dots, n\}$, we have

$$\mathbf{E} \left[\frac{\partial \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t, v}(\theta^*)) \right)}{\partial \theta_i} \right] = 0.$$

The last equation obviously holds for every $i \notin S_t$, and it holds for $i \in S_t$ by the following derivations

$$\begin{aligned} \mathbf{E} \left[\frac{\partial}{\partial \theta_i} \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta^*)) \right) \right] &= -p_{i,S_t}(\theta^*) \sum_{v \in S_t \setminus \{i\}} p_{v,i}(\theta^*) + \sum_{v \in S_t \setminus \{i\}} p_{v,S_t}(\theta^*) p_{i,v}(\theta^*) \\ &= \sum_{v \in S_t \setminus \{i\}} (-p_{i,S_t}(\theta^*) p_{v,i}(\theta^*) + p_{v,S_t}(\theta^*) p_{i,v}(\theta^*)) = 0. \end{aligned}$$

From (57), we have

$$\begin{aligned} \left\| \nabla \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right) \right\|_2^2 &= \left(\sum_{v \in S_t \setminus \{y_t\}} p_{v,y_t}(\theta) \right)^2 + \sum_{v \in S_t \setminus \{y_t\}} p_{v,y_t}(\theta)^2 \\ &\leq (k-1)^2 + k-1 = k(k-1). \end{aligned}$$

Hence,

$$\left\| \nabla \left(\sum_{v \in S_t \setminus \{y_t\}} -\log(p_{y_t,v}(\theta)) \right) \right\|_2 \leq \sqrt{k(k-1)}. \quad (58)$$

Therefore, by vector Azuma-Hoeffding bound in Lemma 2, (56) holds with probability at least $1 - 2/n$. \blacksquare

Appendix F. Proof of Theorem 18

The proof follows by the same steps as that of Theorem 17, and the following two lemmas.

Lemma 33 *If $\lambda_2(L_{\mathbf{M}}) \geq 8k(k-1)e^{2b}n \log(n)/m$, then with probability at least $1 - 1/n$,*

$$\lambda_2(\nabla^2(-\ell_1(\theta))) \geq \frac{1}{4k(k-1)e^{2b}} \frac{m}{n} \lambda_2(L_{\mathbf{M}}). \quad (59)$$

Proof $\nabla^2(-\ell_1(\theta))$ is a sum of random matrices given by:

$$\nabla^2(-\ell_1(\theta)) = \sum_{t=1}^m \nabla^2(-\log(p_{y_t,z_t}(\theta)))$$

where $p_{y_t,z_t}(\theta) = e^{\theta_{y_t}} / (e^{\theta_{y_t}} + e^{\theta_{z_t}})$.

It is easy to establish that for all $t \in \{1, 2, \dots, m\}$ and $i \neq j$,

$$\frac{\partial^2(-\log(p_{y_t,z_t}(\theta)))}{\partial \theta_i \partial \theta_j} = \begin{cases} -p_{i,j}(\theta)(1 - p_{i,j}(\theta)), & \text{if } \{i, j\} = \{y_t, z_t\} \\ 0, & \text{otherwise} \end{cases} \quad (60)$$

and

$$\frac{\partial^2(-\log(p_{y_t,z_t}(\theta)))}{\partial \theta_i^2} = - \sum_{v \in S_t \setminus \{i\}} \frac{\partial^2(-\log(p_{y_t,z_t}(\theta)))}{\partial \theta_i \partial \theta_v}. \quad (61)$$

Hence, for every $i, j \in \{1, 2, \dots, n\}$ such that $i \neq j$,

$$\begin{aligned} \mathbf{E} \left[\frac{\partial^2 (-\log(p_{y_t, z_t}(\theta)))}{\partial \theta_i \partial \theta_j} \right] &= -p_{i,j}(\theta)(1 - p_{i,j}(\theta)) \mathbf{P}[\{y_t, z_t\} = \{i, j\}] \\ &= -p_{i,j}(\theta)(1 - p_{i,j}(\theta)) \frac{1}{k-1} (p_{i, S_t}(\theta) + p_{j, S_t}(\theta)) \\ &= -\frac{1}{k-1} \frac{1}{\sum_{v \in S_t} e^{\theta_v}} \frac{1}{e^{-\theta_i} + e^{-\theta_j}} \\ &\leq -\frac{1}{2k(k-1)e^{2b}}. \end{aligned}$$

Hence,

$$\nabla^2 \mathbf{E}[(-\ell_1(\theta))] \succeq \frac{1}{2k(k-1)e^{2b}} \frac{m}{n} L_{\mathbf{M}}$$

and, in particular,

$$\lambda_2(\nabla^2 \mathbf{E}[(-\ell_1(\theta))]) \geq \frac{1}{2k(k-1)e^{2b}} \frac{m}{n} \lambda_2(L_{\mathbf{M}}). \quad (62)$$

By (60) and (61) and Grešgorin circle theorem,

$$\|\nabla^2 (-\log(p_{y_t, z_t}(\theta)))\|_2 \leq 2p_{y_t, z_t}(\theta)(1 - p_{y_t, z_t}(\theta)) \leq \frac{1}{2}. \quad (63)$$

The rest of the proof follows by the same arguments as in the proof of Lemma 31. \blacksquare

Lemma 34 *With probability at least $1 - 2/n$,*

$$\|\nabla(-\ell_1(\theta^*))\|_2 \leq 2\sqrt{m(\log(n) + 2)}. \quad (64)$$

Proof $\nabla(-\ell_1(\theta))$ is a sum of independent random vectors in \mathbf{R}^n with elements given by

$$\frac{\partial(-\log(p_{y_t, z_t}(\theta)))}{\partial \theta_i} = \begin{cases} -p_{z_t, i}(\theta), & \text{if } i = y_t \\ p_{i, y_t}(\theta), & \text{if } i = z_t \\ 0, & \text{otherwise.} \end{cases}$$

It follows that for all $t \in \{1, 2, \dots, m\}$ and $i \in \{1, 2, \dots, n\}$,

$$\begin{aligned} \mathbf{E} \left[\frac{\partial(-\log(p_{y_t, z_t}(\theta^*)))}{\partial \theta_i} \right] &= -p_{i, S_t}(\theta^*) \sum_{j \in S_t \setminus \{i\}} \frac{1}{k-1} p_{j, i}(\theta^*) + \sum_{j \in S_t \setminus \{i\}} p_{j, S_t}(\theta^*) \frac{1}{k-1} p_{i, j}(\theta^*) \\ &= \frac{1}{k-1} \sum_{j \in S_t \setminus \{i\}} (-p_{i, S_t}(\theta^*) p_{j, i}(\theta^*) + p_{j, S_t}(\theta^*) p_{i, j}(\theta^*)) = 0 \end{aligned}$$

and

$$\|\nabla(-\log(p_{y_t, z_t}(\theta)))\|_2^2 = p_{z_t, y_t}(\theta)^2 + p_{z_t, y_t}(\theta)^2 \leq 2. \quad (65)$$

The statement of the lemma then follows by vector Azuma-Hoeffding bound in Lemma 2. \blacksquare

Appendix G. Proof of Lemma 3

From Theorem 5.1.1 in Tropp (2015),

$$\mathbf{P}[\lambda_1(S_m) \leq (1 - \epsilon)\lambda_1(\mathbf{E}[S_m])] \leq n \left(\frac{e^{-\epsilon}}{(1 - \epsilon)^{1-\epsilon}} \right)^{\frac{\lambda_1(\mathbf{E}[S_m])}{\sigma}} \quad \text{for } \epsilon \in [0, 1]. \quad (66)$$

which combined with the fact

$$\frac{e^{-\epsilon}}{(1 - \epsilon)^{1-\epsilon}} \leq e^{-\frac{\epsilon^2}{2}}, \quad \text{for all } \epsilon \in (0, 1]$$

yields the lemma.

G.1 Proof of Lemma 24

Since $L_{\mathbf{A}} = \text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$, we have

$$L_{\mathbf{A}}\mathbf{1} = \text{diag}(\mathbf{A}\mathbf{1})\mathbf{1} - \mathbf{A}\mathbf{1} = \mathbf{A}\mathbf{1} - \mathbf{A}\mathbf{1} = \mathbf{0}.$$

Hence, 0 is an eigenvalue of $L_{\mathbf{A}}$ for eigenvector $\mathbf{1}$.

Let $\mathbf{y} \in \mathbf{R}^n$ be an eigenvector with corresponding eigenvalue λ of $L_{\mathbf{A}}$. Since the columns of \mathbf{U} are independent, \mathbf{U} is nonsingular and it has inverse \mathbf{U}^{-1} (e.g., Section 0.5 Horn and Johnson (1985)). Let $\mathbf{x} = \mathbf{U}^{-1}\mathbf{y}$.

Note that the following equations hold:

$$\begin{aligned} \mathbf{x}^T \mathbf{U}^T L_{\mathbf{A}} \mathbf{U} \mathbf{x} &= (\mathbf{U} \mathbf{x})^T L_{\mathbf{A}} (\mathbf{U} \mathbf{x}) \\ &= \mathbf{y}^T L_{\mathbf{A}} \mathbf{y} \\ &= \lambda \mathbf{y}^T \mathbf{y} \\ &= \lambda \mathbf{x}^T \mathbf{U}^T \mathbf{U} \mathbf{x} \\ &= \lambda \mathbf{x}^T \mathbf{x}. \end{aligned}$$

Hence, it follows that λ is an eigenvalue of $\mathbf{U}^T L_{\mathbf{A}} \mathbf{U}$ with corresponding eigenvector \mathbf{x} .

Appendix H. Proof of Lemma 25

Let $\mathbf{C} = \mathbf{B} - \mathbf{A} \geq \mathbf{0}$. Note that

$$\begin{aligned} L_{\mathbf{B}} - L_{\mathbf{A}} &= (\text{diag}(\mathbf{B}\mathbf{1}) - \mathbf{B}) - (\text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}) \\ &= \text{diag}((\mathbf{B} - \mathbf{A})\mathbf{1}) - (\mathbf{B} - \mathbf{A}) \\ &= L_{\mathbf{C}}. \end{aligned}$$

Since $L_{\mathbf{C}}$ is positive semidefinite, it follows that $L_{\mathbf{B}} - L_{\mathbf{A}}$ is positive semidefinite, i.e. $L_{\mathbf{B}} \succeq L_{\mathbf{A}}$.

Lemma 35 *For all $S \subseteq \{1, 2, \dots, n\}$ such that $|S| = k \geq 2$ we have: for all $i, j \in \{1, 2, \dots, n\}$ such that $i \neq j$, for $v \in \{1, 2, \dots, n\} \setminus \{i, j\}$,*

$$\frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial \theta_i \partial \theta_j} = -k \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} + k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \quad (67)$$

and

$$\frac{\partial^2(-\log(p_k(\mathbf{x}_j(\mathbf{0}))))}{\partial\theta_i\partial\theta_j} = \frac{k(k-2)}{2} \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} - k^2(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \quad (68)$$

Moreover,

$$\frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial\theta_i^2} = - \sum_{j \in S \setminus \{i\}} \frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial\theta_i \partial\theta_j}. \quad (69)$$

Proof Let $S \subseteq N$ be such that $|S| = k$, for an integer $2 \leq k \leq n$. Without loss of generality, let $S = \{1, 2, \dots, k\}$. Let $\mathbf{x}_v(\theta) = (\theta_v - \theta_u, u \in S \setminus \{v\})$, for $v \in S$. We first consider $\frac{\partial^2}{\partial\theta_i \partial\theta_j}(-\log(p_k(\mathbf{x}_v(\theta))))$ for $i \neq j$. It is easy to note that

$$\begin{aligned} \frac{\partial^2(-\log(p_k(\mathbf{x}_v(\theta))))}{\partial\theta_i \partial\theta_j} &= -\frac{1}{p_k(\mathbf{x}_v(\theta))} \frac{\partial^2 p_k(\mathbf{x}_v(\theta))}{\partial\theta_i \partial\theta_j} \\ &\quad + \frac{1}{p_k(\mathbf{x}_v(\theta))^2} \frac{\partial p_k(\mathbf{x}_v(\theta))}{\partial\theta_i} \frac{\partial p_k(\mathbf{x}_v(\theta))}{\partial\theta_j}. \end{aligned}$$

We separately consider two different cases.

Consider first the case when $\{i\} \cap \{j\} \cap \{v\} = \emptyset$. By differentiation, we have

$$\frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial\theta_i \partial\theta_j} = -k \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} + k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2$$

which establishes (67).

Consider now the case when $i \neq v$ and $j = v$. First, note

$$\begin{aligned} \frac{\partial^2}{\partial\theta_i \partial\theta_j}(-\log(p_k(\mathbf{x}_j(\theta)))) &= -\frac{1}{p_k(\mathbf{x}_j(\theta))} \frac{\partial^2 p_k(\mathbf{x}_j(\theta))}{\partial\theta_i \partial\theta_j} \\ &\quad + \frac{1}{p_k(\mathbf{x}_j(\theta))^2} \frac{\partial p_k(\mathbf{x}_j(\theta))}{\partial\theta_i} \frac{\partial p_k(\mathbf{x}_j(\theta))}{\partial\theta_j}. \end{aligned} \quad (70)$$

For every $u \in S$, $p_k(\mathbf{x}_u(\theta))$ does not change its value by changing θ with $\theta + c\mathbf{1}$, for every constant $c \in \mathbf{R}$. Hence, by full differentiation, we have

$$\frac{\partial p_k(\mathbf{x}_j(\theta))}{\partial\theta_j} = - \sum_{v \in S \setminus \{j\}} \frac{\partial p_k(\mathbf{x}_u(\theta))}{\partial\theta_v}. \quad (71)$$

From (71),

$$\frac{\partial^2 p_k(\mathbf{x}_j(\theta))}{\partial\theta_i \partial\theta_j} = -\frac{\partial^2 p_k(\mathbf{x}_j(\theta))}{\partial\theta_i^2} - \sum_{v \in S \setminus \{i, j\}} \frac{\partial^2 p_k(\mathbf{x}_j(\theta))}{\partial\theta_i \partial\theta_v}. \quad (72)$$

Now, note that

$$\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i^2} = \int_{\mathbf{R}} f(z) f'(x_i + z) \prod_{v \in S \setminus \{i, j\}} F(x_v + z) dz.$$

Hence,

$$\begin{aligned}
\frac{\partial^2 p_k(\mathbf{x}_j(\mathbf{0}))}{\partial \theta_i^2} &= \int_{\mathbf{R}} f(z) f'(z) F(z)^{k-2} dz \\
&= f(z)^2 F(z)^{k-2} \Big|_{-\infty}^{\infty} - \int_{\mathbf{R}} f(z) (f(z) F(z)^{k-2})' dz \\
&= - \int_{\mathbf{R}} f(z) f'(z) F(z)^{k-2} dz - (k-2) \int_{\mathbf{R}} f(z)^2 F(z)^{k-3} dz \\
&= - \frac{\partial^2 p_k(\mathbf{x}_j(\mathbf{0}))}{\partial \theta_i^2} - (k-2) \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2}.
\end{aligned}$$

From this it follows

$$\frac{\partial^2 p_k(\mathbf{x}_j(\mathbf{0}))}{\partial \theta_i^2} = - \frac{k-2}{2} \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2}. \quad (73)$$

From (72) and (73),

$$\frac{\partial^2 p_k(\mathbf{x}_j(\mathbf{0}))}{\partial \theta_i \partial \theta_j} = - \frac{k-2}{2} \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2}. \quad (74)$$

From (71),

$$\frac{\partial p_k(\mathbf{x}_j(\mathbf{0}))}{\partial \theta_j} = (k-1) \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (75)$$

Combining (70), (74) and (75), we obtain (68).

Since for every $v \in S$, $\log(p_k(x_v(\theta)))$ does not change its value by changing θ to $\theta + c\mathbf{1}$ for all $c \in \mathbf{R}$, by full differentiation

$$\frac{\partial(-\log(p_k(x_v(\theta))))}{\partial \theta_i} = - \sum_{j \in S \setminus \{i\}} \frac{\partial(-\log(p_k(x_v(\theta))))}{\partial \theta_j}.$$

Taking partial derivative with respect to θ_i on both sides implies (69). ■

Lemma 36 *Let $S \subseteq N$ be such that $|S| = k \geq 2$ and Y be a random variable according to distribution $p_{Y,S}(\theta)$, for $\theta \in \mathbf{R}^n$. Then,*

$$\mathbf{E} [\nabla^2(-\log(p_{Y,S}(\mathbf{0})))] = \left(k \frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 L_{\mathbf{M}_S}$$

where $\mathbf{M}_S = [m_{i,j}^S] \in \mathbf{R}^{n \times n}$ is such that $m_{i,j}^S = 1$ if $i, j \in S$ and $i \neq j$, and $m_{i,j}^S = 0$, otherwise.

Proof By (67) and (68) in Lemma 35, we have for all $i, j \in \{1, 2, \dots, n\}$ such that $i \neq j$,

$$\begin{aligned}
 \mathbf{E} \left[\frac{\partial^2(-\log(p_{Y,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} \right] &= \sum_{v \in S \setminus \{i, j\}} p_k(\mathbf{x}_v(\mathbf{0})) \frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial \theta_i \partial \theta_j} \\
 &\quad + p_k(\mathbf{x}_i(\mathbf{0})) \frac{\partial^2(-\log(p_k(\mathbf{x}_i(\mathbf{0}))))}{\partial \theta_i \partial \theta_j} \\
 &\quad + p_k(\mathbf{x}_j(\mathbf{0})) \frac{\partial^2(-\log(p_k(\mathbf{x}_j(\mathbf{0}))))}{\partial \theta_i \partial \theta_j} \\
 &= \frac{k-2}{k} \frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial \theta_i \partial \theta_j} \\
 &\quad + \frac{2}{k} \frac{\partial^2(-\log(p_k(\mathbf{x}_i(\mathbf{0}))))}{\partial \theta_i \partial \theta_j}, \text{ for any } v \in S \setminus \{i, j\} \\
 &= -k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \tag{76}
 \end{aligned}$$

By (69) in Lemma 35, we have

$$\begin{aligned}
 \mathbf{E} \left[\frac{\partial^2(-\log(p_{Y,S}(\mathbf{0})))}{\partial \theta_i^2} \right] &= - \sum_{u \in S \setminus \{i\}} \sum_{v \in S} p_k(\mathbf{x}_v(\mathbf{0})) \frac{\partial^2(-\log(p_k(\mathbf{x}_v(\mathbf{0}))))}{\partial \theta_i \partial \theta_u} \\
 &= k^2(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2.
 \end{aligned}$$

■

Lemma 37 *If for $S \subseteq N$ such that $|S| = k \geq 2$,*

$$\frac{\partial^2(-\log(p_{v,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} \leq 0 \text{ for all } i, j, v \in \{1, 2, \dots, n\} \text{ such that } i \neq j$$

then

$$\|\nabla^2(-\log(p_{y,S}(\mathbf{0})))\|_2 \leq \frac{1}{\gamma_{F,k}}$$

where $1/\gamma_{F,k} = (k^2 \partial p_k(\mathbf{0}) / \partial x_1)^2$.

Proof Without loss of generality, let $y = 1$ and $S = \{1, 2, \dots, k\}$. By Lemma 35, we have

$$\frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial \theta_1 \partial \theta_2} = -k^2(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 + \frac{k(k-2)}{2} \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} \tag{77}$$

and for $i \neq 1, j \neq 1$ and $j \neq i$,

$$\frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} = -k \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} + k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \tag{78}$$

Since by condition of the lemma the left-hand side in (77) is non positive, we have

$$(k-2) \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} \leq 2k(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \quad (79)$$

and, since by condition of the lemma the left-hand side in (78) is non positive, we have

$$\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} \geq 0. \quad (80)$$

From (77) and (80),

$$\frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial \theta_1 \partial \theta_2} \geq -k^2(k-1) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2.$$

Consider now the case when $i \neq 1$, $j \neq 1$ and $i \neq j$. If $k = 2$, then obviously

$$\frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} = 0.$$

Otherwise, if $k > 2$, from (78) and (79), we have

$$\begin{aligned} \frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial \theta_i \partial \theta_j} &\geq k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 - \frac{2k^2(k-1)}{k-2} \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \\ &= -k^2 \left(-1 + \frac{2(k-1)}{k-2} \right) \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \\ &= -k^2 \frac{k}{k-2} \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \\ &\geq -k^3 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \end{aligned}$$

Hence, it follows

$$k^3 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 L_{\mathbf{M}_S} \succeq \nabla^2(-\log(p_{y,S}(\mathbf{0}))).$$

Therefore, we conclude

$$\begin{aligned} \|\nabla^2(-\log(p_{y,S}(\mathbf{0})))\|_2 &\leq k^3 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \|L_{\mathbf{M}_S}\|_2 \\ &= k^4 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \end{aligned}$$

■

Appendix I. Remark for Theorem 10

For the special case of noise according to the double-exponential distribution with parameter β , we have

$$p_k(\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{k-1} e^{-x_i/\beta}}.$$

For every $\theta \in \theta_n$ and every $S \subseteq N$ of cardinality k and $i, j, y \in S$, we can easily check that

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} (-\log(p_{y,S}(\theta))) = -\frac{1}{\beta^2} p_{i,S}(\theta) p_{j,S}(\theta).$$

Furthermore, the following two relations hold

$$\frac{k}{\beta(k-1)} (1 - p_{y,S}(\theta))^2 \leq \|\nabla p_{y,S}(\theta)\|_2 \leq \frac{2}{\beta} (1 - p_{y,S}(\theta))^2.$$

Since

$$\min_{y \in S, \theta \in [-b, b]^n} p_{y,S}(\theta) = \frac{1}{1 + (k-1)e^{2b/\beta}} \geq p_{y,S}(\mathbf{0})e^{-2b/\beta}$$

and

$$\max_{y \in S, \theta \in [-b, b]^n} p_{y,S}(\theta) = \frac{1}{1 + (k-1)e^{-2b/\beta}} \leq p_{y,S}(\mathbf{0})e^{2b/\beta}$$

we have that

$$\sigma_{F,K} \leq \frac{1}{\beta^2}$$

and

$$e^{-4b/\beta} \leq A \leq \tilde{A} \leq e^{4b/\beta}, \quad (81)$$

$$e^{-4b/\beta} \leq \tilde{B} \leq B \leq 4, \quad (82)$$

$$e^{-2b/\beta} \leq C \leq \tilde{C} \leq e^{2b/\beta}. \quad (83)$$

Appendix J. Proof of Theorem 19

Let p^e denote the probability that the point score ranking method incorrectly classifies at least one item:

$$p^e = \mathbf{P} \left[\bigcup_{v \in N_1} \{v \in \hat{N}_1\} \cup \bigcup_{v \in N_2} \{v \in \hat{N}_2\} \right]$$

Let R_i denote the point score of item $i \in N$. If the point scores are such that $R_v > m/n$ for every $v \in N_1$ and $R_v < m/n$ for every $v \in N_2$, then this implies a correct classification. Hence, it must be that in the event of a misclassification of an item, $R_v \leq m/n$ for some $v \in N_1$ or $R_v \geq m/n$ for some $v \in N_2$. Combining this with the union bound, we have

$$p^e \leq \sum_{v \in N_1} \mathbf{P} \left[R_v \leq \frac{m}{n} \right] + \sum_{v \in N_2} \mathbf{P} \left[R_v \geq \frac{m}{n} \right]. \quad (84)$$

Let i and j be arbitrarily fixed items such that $i \in N_1$ and $j \in N_2$. We will show that for $t \in \{1, 2, \dots, m\}$,

$$\mathbf{P}[y_t = i] \geq \frac{1}{n} + \frac{bk^2}{4n} \frac{\partial p_k(\mathbf{0})}{\partial x_1} \quad (85)$$

and

$$\mathbf{P}[y_t = j] \leq \frac{1}{n} - \frac{bk^2}{4n} \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (86)$$

By the Chernoff bound (26) we have

$$\begin{aligned} \mathbf{P}\left[R_i \leq \frac{m}{n}\right] &\leq \exp\left(-\frac{1}{4}n\left(\frac{1}{n} - \mathbf{E}[y_1 = i]\right)^2 m\right) \\ &\leq \exp\left(-\frac{1}{4}\left(\frac{bk^2}{4n} \frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 m\right) \\ &\leq \exp(-\log(n/\delta)) \\ &= \frac{\delta}{n}. \end{aligned}$$

Similarly, by the Chernoff bound (27), we have

$$\mathbf{P}\left[R_j \geq \frac{m}{n}\right] \leq \frac{\delta}{n}.$$

Combining with (84), we have $p^e \leq \delta$.

In the remainder of the proof we show that (85) and (86) hold.

Let \mathcal{A} contain all $A \subseteq N$ such that $|A| = k-1$ and $A \cap \{i, j\} = \emptyset$ and \mathcal{B} contain all $B \subseteq N$ such that $|B| = k-2$ and $B \cap \{i, j\} = \emptyset$. Then, we have

$$\mathbf{P}[y_t = i] - \mathbf{P}[y_t = j] = \sum_{A \in \mathcal{A}} \mathbf{P}[S_t = A \cup \{i\}] D_{i,j}(A) + \sum_{B \in \mathcal{B}} \mathbf{P}[S_t = B \cup \{i, j\}] D_{i,j}(B) \quad (87)$$

where

$$D_{i,j}(A) = \mathbf{P}[y_t = i | S_t = A \cup \{i\}] - \mathbf{P}[y_t = j | S_t = A \cup \{j\}]$$

and

$$D_{i,j}(B) = \mathbf{P}[y_t = i | S_t = B \cup \{i, j\}] - \mathbf{P}[y_t = j | S_t = B \cup \{i, j\}].$$

Let \mathbf{b} be a $k-1$ -dimensional vector with all elements equal to b . Then, note that

$$D_{i,j}(A) = p_k(\mathbf{b} - \theta_A) - p_k(-\mathbf{b} - \theta_A).$$

By limited Taylor series development, we have

$$p_k(\mathbf{x}) \geq p_k(\mathbf{0}) + \nabla p_k(\mathbf{0})^\top \mathbf{x} - \frac{1}{2} \beta \|\mathbf{x}\|_2^2 \quad (88)$$

$$p_k(\mathbf{x}) \leq p_k(\mathbf{0}) + \nabla p_k(\mathbf{0})^\top \mathbf{x} + \frac{1}{2} \beta \|\mathbf{x}\|_2^2 \quad (89)$$

where

$$\beta = \max_{\mathbf{x} \in [-2b, 2b]^{k-1}} \|\nabla^2 p_k(\mathbf{x})\|_2. \quad (90)$$

Hence, it follows that for every $\theta_A \in \{-b, b\}^{k-1}$,

$$D_{i,j}(A) \geq 2(k-1)b \frac{\partial p_k(\mathbf{0})}{\partial x_1} - 4(k-1)b^2\beta. \quad (91)$$

Under the condition of the theorem, we have

$$\beta \leq \frac{1}{4b} \frac{\partial p_k(\mathbf{0})}{\partial x_1}.$$

Hence, combining with (91), for every $\theta_A \in \{-b, b\}^{k-1}$,

$$D_{i,j}(A) \geq (k-1)b \frac{\partial p_k(\mathbf{0})}{\partial x_1} \geq \frac{kb}{2} \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (92)$$

By the same arguments, we can show that

$$D_{i,j}(B) = p_k(\mathbf{b} - \theta_B^{(-b)}) - p_k(-\mathbf{b} - \theta_B^{(b)}) \geq \frac{kb}{2} \frac{\partial p_k(\mathbf{0})}{\partial x_1} \quad (93)$$

where $\theta_B^{(b)} \in \{-b, b\}^{k-1}$ and $\theta_B^{(-b)} \in \{-b, b\}^{k-1}$ are $(k-1)$ -dimensional vectors with the first two elements equal to b and $-b$, respectively, and other elements equal to the parameters of items B .

Since comparison sets are sampled uniformly at random without replacement,

$$\mathbf{P}[S_t = A \cup \{i\}] = \frac{\binom{n-1}{k-1}}{\binom{n}{k}}, \text{ for all } A \in \mathcal{A} \quad (94)$$

and

$$\mathbf{P}[S_t = B \cup \{i, j\}] = \frac{\binom{n-2}{k-2}}{\binom{n}{k}}, \text{ for all } B \in \mathcal{B}. \quad (95)$$

From (87), (92), (93), (94) and (95), we have

$$\mathbf{P}[y_t = i] - \mathbf{P}[y_t = j] \geq \frac{k^2b}{2n} \frac{\partial p_k(\mathbf{0})}{\partial x_1}.$$

Using this inequality together with the following facts (i) $\mathbf{P}[y_t = v] = \mathbf{P}[y_t = i]$ for every $v \in N_1$, (ii) $\mathbf{P}[y_t = v] = \mathbf{P}[y_t = j]$ for every $v \in N_2$, (iii) $\sum_{v \in N} \mathbf{P}[y_t = v] = 1$, and (iv) $|N_1| = |N_2| = n/2$, it can be readily shown that

$$\mathbf{P}[y_t = i] \geq \frac{1}{n} + \frac{k^2b}{4n} \frac{\partial p_k(\mathbf{0})}{\partial x_1},$$

which establishes (85). By the same arguments one can establish (86).

Appendix K. Proof of Theorem 20

Suppose that n is a positive even integer and θ is the parameter vector such that $\theta_i = b$ for $i \in N_1$ and $\theta_i = -b$ for $i \in N_2$, where $N_1 = \{1, 2, \dots, n/2\}$ and $N_2 = \{n/2 + 1, \dots, n\}$. Let θ' be the parameter vector that is identical to θ except for swapping the first and the last item, i.e. $\theta'_i = b$ for $i \in N'_1$ and $\theta'_i = -b$ for $i \in N'_2$, where $N'_1 = \{n, 2, \dots, n/2\}$ and $N'_2 = \{n/2 + 1, \dots, n - 1, 1\}$.

We denote with $\mathbf{P}_\theta[A]$ and $\mathbf{P}_{\theta'}[A]$ the probabilities of an event A under hypothesis that the generalized Thurstone model is according to parameter θ and θ' , respectively. We denote with \mathbf{E}_θ and $\mathbf{E}_{\theta'}$ the expectations under the two respective distributions.

Given observed data $(\mathbf{S}, \mathbf{y}) = (S_1, y_1), \dots, (S_m, y_m)$, we denote the log-likelihood ratio statistic $L(\mathbf{S}, \mathbf{y})$ as follows

$$L(\mathbf{S}, \mathbf{y}) = \sum_{t=1}^m \log \left(\frac{p_{y_t, S_t}(\theta') \rho_t(S_t)}{p_{y_t, S_t}(\theta) \rho_t(S_t)} \right), \quad (96)$$

where $\rho_t(S)$ is the probability that S is drawn at time t .

The proof follows the following two steps:

Step 1: We show that for given $\delta \in [0, 1]$, for the existence of an algorithm that correctly classifies all the items with probability at least $1 - \delta$, it is necessary that the following condition holds

$$\mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \geq \log(n/\delta)] \geq \frac{1}{2}. \quad (97)$$

Step 2: We show that

$$\mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})] \leq 36 \frac{m}{n} \left(k^2 b \frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \quad (98)$$

$$\sigma_{\theta'}^2[L(\mathbf{S}, \mathbf{y})] \leq 144 \frac{m}{n} \left(k^2 b \frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2 \quad (99)$$

where $\sigma_{\theta'}^2[L(\mathbf{S}, \mathbf{y})]$ denotes the variance of random variable $L(\mathbf{S}, \mathbf{y})$ under a generalized Thurstone model with parameter θ' .

By Chebyshev's inequality, for every $g \in \mathbf{R}$,

$$\mathbf{P}_{\theta'}[|L(\mathbf{S}, \mathbf{y}) - \mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})]| \geq |g|] \leq \frac{\sigma_{\theta'}^2[L(\mathbf{S}, \mathbf{y})]}{g^2}.$$

Using this for $g = \log(n/\delta) - \mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})]$, it follows that (97) implies the following condition:

$$\begin{aligned} \log(n/\delta) - \mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})] &\leq |\log(n/\delta) - \mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})]| \\ &\leq \sqrt{2\sigma_{\theta'}^2[L(\mathbf{S}, \mathbf{y})]}. \end{aligned}$$

Further combining with (98) and (99), we obtain

$$m \geq \frac{1}{62} \frac{1}{b^2 k^4 (\partial p_k(\mathbf{0}) / \partial x_1)^2} n (\log(n) + \log(1/\delta))$$

which is the condition asserted in the theorem.

Proof of Step 1. Let us define the following two events

$$A = \{|N_1 \setminus \widehat{N}_1| = 1\} \cap \{|N_2 \setminus \widehat{N}_2| = 1\}$$

and

$$B = \{\widehat{N}_1 = N'_1\} \cap \{\widehat{N}_2 = N'_2\}.$$

Let B^c denote the complement of event B .

Note that

$$\mathbf{P}_\theta[B] = \mathbf{P}_\theta[B|A]\mathbf{P}_\theta[A] = \left(\frac{2}{n}\right)^2 \mathbf{P}_\theta[A] \leq \frac{4}{n^2}\delta$$

where the second equation holds because $B \subseteq A$ and every possible partition in A has the same probability under θ .

For every $g \in \mathbf{R}$, we have

$$\mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \leq g] = \mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \leq g, B] + \mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \leq g, B^c].$$

Now, note

$$\begin{aligned} \mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \leq g, B] &= \mathbf{E}_{\theta'}[\mathbf{1}(L(\mathbf{S}, \mathbf{y}) \leq g, B)] \\ &= \mathbf{E}_\theta[e^{L(\mathbf{S}, \mathbf{y})} \mathbf{1}(L(\mathbf{S}, \mathbf{y}) \leq g, B)] \\ &\leq \mathbf{E}_\theta[e^g \mathbf{1}(L(\mathbf{S}, \mathbf{y}) \leq g, B)] \\ &= e^g \mathbf{P}_\theta[L(\mathbf{S}, \mathbf{y}) \leq g, B] \\ &\leq e^g \mathbf{P}_\theta[B] \\ &\leq e^g \frac{4}{n^2} \delta \end{aligned}$$

where in the second equation we use the standard change of measure argument.

Since the algorithm correctly classifies all the items with probability at least $1 - \delta$, we have

$$\mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \leq g, B^c] \leq \mathbf{P}_{\theta'}[B^c] \leq \delta. \quad (100)$$

For $g = \log(n/\delta)$, from (100) and (100), it follows that

$$\mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \leq \log(n/\delta)] \leq \delta + \frac{4}{n} \leq \frac{1}{2}$$

where the last inequality is by the conditions of the theorem.

Proof of Step 2. If the observed comparison sets S_1, S_2, \dots, S_m are such that $S_t \cap \{1, n\} = \emptyset$, for every observation t , then we obviously have

$$\log \left(\frac{p_{y_t, S_t}(\theta')}{p_{y_t, S_t}(\theta)} \right) = 0, \text{ for all } t.$$

We therefore consider the case when $S_t \cap \{1, n\} \neq \emptyset$.

Using (88), (89), and (90), we have for every S and $i \in S$,

$$|p_{i,S}(\theta') - p_{i,S}(\theta)| \leq 2kb \frac{\partial p_k(\mathbf{0})}{\partial x_1} + 4\beta b k \leq 3kb \frac{\partial p_k(\mathbf{0})}{\partial x_1} \quad (101)$$

where the last inequality is obtained from the condition of this theorem.

From (101), for every comparison set S such that $S \cap \{1, n\} \neq \emptyset$, we have

$$\begin{aligned} \sum_{i \in S} (p_{i,S}(\theta') - p_{i,S}(\theta))^2 &\leq \sum_{i \in \{1, n\} \cap S} (p_{i,S}(\theta') - p_{i,S}(\theta))^2 + \left(\sum_{i \in S \setminus \{1, n\}} p_{i,S}(\theta') - p_{i,S}(\theta) \right)^2 \\ &\leq 2 \left(3kb \frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2, \end{aligned} \quad (102)$$

which is because for every comparison set S such that $1 \in S$,

$$p_{1,S}(\theta') \leq \frac{1}{k} \leq p_{1,S}(\theta) \text{ and } p_{i,S}(\theta') \geq p_{i,S}(\theta) \quad \forall i \neq 1;$$

and, for every comparison set S such that $n \in S$,

$$p_{n,S}(\theta') \geq \frac{1}{k} \geq p_{n,S}(\theta) \text{ and } p_{i,S}(\theta') \leq p_{i,S}(\theta) \quad \forall i \neq n.$$

From (101) and the assumption of the theorem, we have

$$\min_S \min_{i \in S} p_{i,S}(\theta) = \min_{S: n \in S} p_{n,S}(\theta) \geq \frac{1}{k} - 3kb \frac{\partial p_k(\mathbf{0})}{\partial x_1} \geq \frac{1}{2k}. \quad (103)$$

For simplicity of notation, let

$$D = 3kb \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (104)$$

Then, for all S such that $S \cap \{1, n\} \neq \emptyset$, we have

$$\sum_{i \in S} p_{i,S}(\theta') \log \left(\frac{p_{i,S}(\theta')}{p_{i,S}(\theta)} \right) \leq 2kD^2 \quad (105)$$

which is obtained from

- (i) $p_{i,S}(\theta) \geq 1/(2k)$ for all $i \in S$ that holds by (103),
- (ii) $\sum_{i \in S} (p_{i,S}(\theta') - p_{i,S}(\theta))^2 = 2D^2$ from (102),
- (iii) $a \log \frac{a}{b} \leq \frac{(a-b)^2}{2b} + a - b$.

Similarly to (105), from (i) and (ii) and $a \left(\log \frac{a}{b} \right)^2 \leq \frac{(a-b)^2}{a \wedge b} \left(1 + \frac{|a-b|}{3(a \wedge b)} \right)$, we have

$$\sum_{i \in S} p_{i,S}(\theta') \left(\log \left(\frac{p_{i,S}(\theta')}{p_{i,S}(\theta)} \right) \right)^2 \leq 8kD^2. \quad (106)$$

Since

$$\mathbf{P}_{\theta'}[\{S_t \cap \{1, n\} \neq \emptyset\}] = 1 - \frac{\binom{n-2}{k}}{\binom{n}{k}} \leq 2 \frac{k}{n}$$

and according to the model, the input observations are independent, from (105) and (106), we have

$$\begin{aligned}
 \mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})] &= m \mathbf{E}_{\theta'} \left[\log \left(\frac{p_{y_1, S_1}(\theta')}{p_{y_1, S_1}(\theta)} \right) \right] \\
 &= m \sum_{S: S \cap \{1, n\} \neq \emptyset} \mathbf{P}_{\theta'}[S_1 = S] \sum_{y \in S} p_{y, S}(\theta') \left[\log \left(\frac{p_{y, S}(\theta')}{p_{y, S}(\theta)} \right) \right] \\
 &\leq 4 \frac{m}{n} k^2 D^2
 \end{aligned} \tag{107}$$

and

$$\begin{aligned}
 \sigma_{\theta'}^2[L(\mathbf{S}, \mathbf{y})] &= m \sigma_{\theta'}^2 \left[\log \left(\frac{p_{y_1, S_1}(\theta')}{p_{y_1, S_1}(\theta)} \right) \right] \\
 &\leq m \mathbf{E}_{\theta'} \left[\left(\log \left(\frac{p_{y_1, S_1}(\theta')}{p_{y_1, S_1}(\theta)} \right) \right)^2 \right] \\
 &= m \sum_{S: S \cap \{1, n\} \neq \emptyset} \mathbf{P}_{\theta'}[S_1 = S] \sum_{y \in S} p_{y, S}(\theta') \left[\left(\log \left(\frac{p_{y, S}(\theta')}{p_{y, S}(\theta)} \right) \right)^2 \right] \\
 &\leq 16 \frac{m}{n} k^2 D^2.
 \end{aligned} \tag{108}$$

Appendix L. Characterizations of $\partial p_k(\mathbf{0})/\partial x_1$

In this section, we note several different representations of the parameter $\partial p_k(\mathbf{0})/\partial x_1$.

First, note that

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \frac{1}{k-1} \int_{\mathbf{R}} f(x) dF(x)^{k-1}. \tag{109}$$

The integral corresponds to $\mathbf{E}[f(X)]$ where X is a random variable whose distribution is equal to that of a maximum of $k-1$ independent and identically distributed random variables with cumulative distribution F .

Second, suppose that F is a cumulative distribution function with its support contained in $[-a, a]$, and that has a differentiable density function f . Then, we have

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = A_{F,k} + B_{F,k} \tag{110}$$

where

$$A_{F,k} = \frac{1}{k-1} f(a)$$

and

$$B_{F,k} = \frac{1}{k(k-1)} \int_{-a}^a (-f'(x)) dF(x)^k.$$

The identity (110) is shown to hold as follows. Note that

$$\frac{d^2}{dx^2} F(x)^k = \frac{d}{dx} (k F(x)^{k-1} f(x)) = k(k-1) F(x)^{k-2} f(x)^2 + k F(x)^{k-1} f'(x).$$

By integrating over $[-a, a]$, we obtain

$$\frac{d}{dx} F(x)^k|_{-a}^a = k(k-1) \int_{-a}^a f(x)^2 F(x)^{k-2} dx + k \int_{-a}^a f'(x) F(x)^{k-1} dx.$$

Combining with the fact

$$\frac{d}{dx} F(x)^k|_{-a}^a = k f(x) F^{k-1}(x)|_{-a}^a = k f(a),$$

we obtain (110).

Note that $B_{F,k} = \mathbf{E}[-f'(X)]/(k(k-1))$ where X is a random variable with distribution that corresponds to that of a maximum of k independent samples from the cumulative distribution function F . Note also that if, in addition, f is an even function, then (i) $B_{F,k} \geq 0$ and (ii) $B_{F,k}$ is increasing in k .

Third, for any cumulative distribution function F with an even density function f , we have $F(-x) = 1 - F(x)$ for all $x \in \mathbf{R}$. In this case, we have the identity

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_0^\infty f(x)^2 (F(x)^{k-2} + (1 - F(x))^{k-2}) dx. \quad (111)$$

Appendix M. Proof of Lemma 21

The upper bound follows by noting that that $B_{F,k}$ in (110) is such that $B_{F,k} = \Omega(1/k^2)$. Hence, it follows that

$$\gamma_{F,k} = O(1).$$

The lower bound follows by noting that for every cumulative distribution function F such that there exists a constant $C > 0$ such that $f(x) \leq C$ for all $x \in \mathbf{R}$,

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx \leq C \int_{\mathbf{R}} f(x) F(x)^{k-2} dx = C \frac{1}{k-1}.$$

Hence, $\gamma_{F,k} \geq (1/C)(k-1)/k^3 = \Omega(1/k^2)$.

Appendix N. Derivations of parameter $\gamma_{F,k}$

We derive explicit expressions for parameter $\gamma_{F,k}$ for our example generalized Thurstone choice models introduced in Section 2

Recall from (19) that we have that

$$\gamma_{F,k} = \frac{1}{(k-1)k^3} \frac{1}{(\partial p_k(\mathbf{0})/\partial x_1)^2}$$

where

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx$$

Gaussian distribution A cumulative distribution function F is said to have a type-3 domain of maximum attraction if the maximum of r independent and identically distributed random variables with cumulative distribution function F has as a limit a double-exponential cumulative distribution function:

$$e^{-e^{-\frac{x-a_r}{b_r}}}$$

where

$$a_r = F^{-1}\left(1 - \frac{1}{r}\right)$$

and

$$b_r = F^{-1}\left(1 - \frac{1}{er}\right) - F^{-1}\left(1 - \frac{1}{r}\right).$$

It is a well known fact that any Gaussian cumulative distribution function has a type-3 domain of maximum attraction. Let Φ denote the cumulative distribution function of a standard normal random variable, and let ϕ denotes its density.

Note that

$$\begin{aligned} \int_{\mathbf{R}} \phi(x) d\Phi(x)^r &\sim \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{-\frac{x^2}{2}} d(e^{-e^{-\frac{x-a_r}{b_r}}}) \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{1}{2}(a_r+b_r \log(1/z))^2} e^{-z} dz \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_r^2} \int_0^\infty z^{a_r b_r} e^{-\frac{1}{2}b_r^2 \log(1/z)^2} e^{-z} dz \\ &\leq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_r^2} \int_0^\infty z^{a_r b_r} e^{-z} dz \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_r^2} \Gamma(a_r b_r + 1). \end{aligned}$$

Now, note that

$$a_r \sim \sqrt{2 \log(r)} \text{ and } b_r = \Theta(1), \text{ for large } r.$$

It is readily checked that $e^{-a_r^2/2} \sim 1/r$ and $\Gamma(a_r b_r + 1) = O(r^\epsilon)$ for every constant $\epsilon > 0$. Hence, we have that

$$\int_{\mathbf{R}} \phi(x) d\Phi(x)^r = O(1/r^{1-\epsilon})$$

and thus, $\partial p_k(\mathbf{0})/\partial x_1 = O(1/k^{2-\epsilon})$. Hence,

$$\gamma_{F,k} = \Omega(1/k^{2\epsilon}).$$

Double-exponential distribution Note that $f(x) = \frac{1}{\beta} e^{-\frac{x+\beta\gamma}{\beta}} F(x)$. Hence, we have

$$\begin{aligned}
 \frac{\partial p_k(\mathbf{0})}{\partial x_1} &= \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx \\
 &= \frac{1}{\beta^2} \int_{\mathbf{R}} e^{-2\frac{x+\beta\gamma}{\beta}} F(x)^k dx \\
 &= \frac{1}{\beta} \int_0^\infty z e^{-kz} dz \\
 &= \frac{1}{\beta k^2}.
 \end{aligned}$$

Laplace distribution Let $\beta = \sigma/\sqrt{2}$. Note that

$$F(x) = 1 - \frac{1}{2}e^{-x/\beta} \text{ and } f(x) = \frac{1}{2\beta}e^{-x/\beta}, \text{ for } x \in \mathbf{R}_+.$$

$$\begin{aligned}
 A &= \int_0^\infty f(x)^2 F(x)^{k-2} dx \\
 &= \int_0^\infty \left(\frac{1}{2\beta}\right)^2 e^{-2x/\beta} \left(1 - \frac{1}{2}e^{-x/\beta}\right)^{k-2} dx \\
 &= \frac{1}{2\beta} \int_{1/2}^1 2(1-z)z^{k-2} dz \\
 &= \frac{1}{\beta} \left(\frac{1}{k-1} \left(1 - \frac{1}{2^{k-1}}\right) - \frac{1}{k} \left(1 - \frac{1}{2^k}\right) \right) \\
 &= \frac{1}{\beta k(k-1)} \left(1 - \frac{k}{2^{k-1}} + \frac{k-1}{2^k}\right)
 \end{aligned}$$

and

$$\begin{aligned}
 B &= \int_0^\infty f(x)^2 (1 - F(x))^{k-2} dx \\
 &= \int_0^\infty \left(\frac{1}{2\beta}\right)^2 e^{-2x/\beta} \frac{1}{2^{k-2}} e^{-(k-2)x/\beta} dx \\
 &= \frac{1}{\beta^2 2^k} \int_0^\infty e^{-kx/\beta} dx \\
 &= \frac{1}{\beta k 2^k}.
 \end{aligned}$$

Combining with (111), we obtain

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = A + B = \frac{1}{\beta k(k-1)} \left(1 - \frac{1}{2^{k-1}}\right).$$

Uniform distribution Note that

$$\begin{aligned}
 \frac{\partial p_k(\mathbf{0})}{\partial x_1} &= \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx \\
 &= \frac{1}{(2a)^2} \int_{-a}^a \left(\frac{x+a}{2a} \right)^{k-2} dx \\
 &= \frac{1}{2a} \int_0^1 z^{k-2} dz \\
 &= \frac{1}{2a(k-1)}.
 \end{aligned}$$