Parameter Estimation for Generalized Thurstone Choice Models

Milan Vojnovic MILANV@MICROSOFT.COM
Microsoft Research, Cambridge, UK

Se-Young Yun T-SEYUN@MICROSOFT.COM
Microsoft Research, Cambridge, UK

Abstract

We consider the maximum likelihood parameter estimation problem for a generalized Thurstone choice model, where choices are from comparison sets of two or more items. We provide tight characterizations of the mean square error, as well as necessary and sufficient conditions for correct classification when each item belongs to one of two classes. These results provide insights into how the estimation accuracy depends on the choice of a generalized Thurstone choice model and the structure of comparison sets. We find that for a priori unbiased structures of comparisons, e.g., when comparison sets are drawn independently and uniformly at random, the number of observations needed to achieve a prescribed estimation accuracy depends on the choice of a generalized Thurstone choice model. For a broad set of generalized Thurstone choice models, which includes all popular instances used in practice, the estimation error is shown to be largely insensitive to the cardinality of comparison sets. On the other hand, we found that there exist generalized Thurstone choice models for which the estimation error decreases much faster with the cardinality of comparison sets. We report results of empirical evaluations using schedules of contests as observed in some popular sport competitions and online crowdsourcing systems.

1. Introduction

We consider the problem of estimating the strengths of items based on observed choices of items, where each choice is from a subset of two or more items. This accommodates pair comparisons as a special case, where each comparison set consists of two items. In general, the outcome of each comparison is a top-1 list that singles out one item from given set of compared items. There are many applications in practice that are accommodated by this framework, e.g., single-winner contests in crowdsourcing services such as TopCoder or Taskcn, or hiring decisions where one applicant gets hired among those who applied for a job, e.g., in online labour marketplaces such as Fiverr and Upwork, as well as numerous sports competitions and online gaming platforms.

In particular, we consider the choices according to a generalized Thurstone choice model. This model accommodates several well known models, e.g. Luce's choice model, and Bradley-Terry model for pair comparisons; see discussion of related work in Section 1.1. A generalized Thurstone choice model is defined by a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbf{R}^n$, where θ_i represents the strength of item i, and a cumulative distribution function F. For every given non-empty subset of items S, the choice is assumed to be an item in S that exhibits the best performance, where the performance of each item $i \in S$ is defined as the sum of the strength parameter θ_i and an independent sample from the cumulative distribution function F. Many well known models of choice are special instances of generalized Thurstone choice models for specific choices of F; see a catalogue of examples in Section 2.3.

In this paper, our goal is to characterize the accuracy of a parameter estimator of a generalized Thurstone choice model. In particular, we want to understand how is the estimation accuracy affected by the choice of a generalized Thurstone model, and the structure of the comparison sets. Our results show that the choice of a generalized Thurstone model can have a substantial effect on the parameter estimation accuracy.

More specifically, our main contributions in this paper can be summarized as follows.

We provide tight lower and upper bounds for the mean square error of the maximum likelihood parameter estimator (Section 3). These results provide necessary and suffi-

This document is Microsoft Research Technical Report MSR-TR-2016-2, February 2016.

cient conditions for the estimation of the parameter within a prescribed accuracy. Moreover, they reveal how the choice of a generalized Thurstone choice model and the structure of comparison sets affect the estimation accuracy. In particular, we find that a key parameter is an eigenvalue gap of a pair-weight matrix. This pair-weight matrix is defined such that each element of this matrix that corresponds to a pair of items is equal to a weighted sum of the number of co-participations of the given pair of items in comparison sets of different cardinalities. The weight associated with a comparison set is a decreasing function of the cardinality of the comparison set, which depends on the choice of the generalized Thurstone choice model.

As a corollary, we derive tight characterizations of the mean square error for the case when all comparison sets are of equal cardinalities and the comparison sets are unbiased, e.g., each comparison set is sampled independently, uniformly at random without replacement from the set of all items. Such comparison sets are in spirit of tournament schedules like round-robin schedules that are common in various sports competitions. We also consider the parameter estimation problem for a generalized Thurstone choice model where each item is either of a high or a low class (Section 4). We establish necessary and sufficient conditions for correct classification of all items, when comparison sets have equal cardinalities and are drawn independently, uniformly at random without replacement from the set of all items. These conditions are shown to match those derived from the bounds for the mean square error up to constant factors.

These results provide a clear picture about the effect of a choice of a generalized Thurstone choice model and the cardinality of comparison sets. Perhaps surprisingly, we find that for a large set of special instances of generalized Thurstone choice models, which includes all popular cases used in practice, the mean square error decreases with the cardinality of comparison sets, but rather weakly. In particular, the mean square error is shown to be largely insensitive to the cardinality of comparison sets of three or more items. On the other hand, we exhibit instances of generalized Thurstone choice models for which the mean square error decreases much faster with the cardinality of comparison sets; in particular, decreasing inversely proportionally to the square of the cardinality (Section 5).

We present experimental results using both simulation and real-world data (Section 6). In particular, we validate the claim that the mean square error can be significantly affected by the choice of a generalized Thurstone model, and evaluate the eigenvalue gap for pair-weight matrices for comparison sets as observed in several real-world datasets.

1.1. Related Work

The original Thurstone choice model was proposed by (Thurstone, 1927) as a model of comparative judgement for pair comparisons. The key property of this model is that each item is assumed to be associated with a performance random variable defined as the sum of a strength parameter and a noise random variable. Specifically, in the original Thurstone model, the noise is assumed to be a Gaussian random variable. This amounts to the winning probability of one item against another item in a pair comparison that is a cumulative Gaussian distribution function of the difference of their corresponding strength parameters. Similar model but with winning probabilities according to a logistic cumulative distribution function was originally studied by (Zermelo, 1929), and following the work by (Bradley & Terry, 1952; 1954) is often referred to as the Bradley-Terry model. A generalization of this model to comparisons of two or more items was studied by (Luce, 1959) and is referred to as the Luce's choice model (Luce, 1959). Other models of choice have also been studied, e.g., Dawkins' choice model (Dawkins, 1969). Relationships between the Luce's choice model and generalized Thurstone choice models were studied in (Yellott, 1977). Some of these models underlie the design of popular rating systems, e.g., Elo rating system (Elo, 1978) that was originally designed and has been used for rating skills of chess players but also for various other sport competitions, and TrueSkill (Graepel et al., 2006) that is used by a popular online gaming platform. All these models are instances of a generalized Thurstone model, and are special instances of generalized linear models, see, e.g., (Nelder & Wedderburn, 1972), (McCullagh & Nelder, 1989), and Chapter 9 in (Murphy, 2012). See Chapter 9 (Vojnović, 2016) for an exposition to the principles of rating systems.

Several studies argued that different models of pair comparisons yield empirically equivalent performance, e.g. (Stern, 1992), suggesting that the choice of a generalized Thurstone model does not matter much in practice. Our results show that there can be a significant fundamental difference between generalized Thurstone choice models with respect to the parameter estimation accuracy.

More recent work has focused on characterizing the parameter estimation error and deriving efficient computational methods for parameter estimation for different models of pair comparisons, e.g., (Negahban et al., 2012) and (Rajkumar & Agarwal, 2014) for pair comparisons according to Bradley-Terry model, and (Hajek et al., 2014) for full ranking outcomes according to a generalized Thurstone model with double-exponential distribution of noise. Our work is different in that we consider the parameter estimation error for generalized Thurstone choice models that allow for comparisons of two or more items and different distribu-

tions of individual performances.

2. Problem Formulation and Notation

2.1. Basic Definitions

We denote with $N = \{1, 2, \ldots, n\}$ the set of all items. The input data consists of a sequence of $m \geq 1$ observations (S_1, y_1), (S_2, y_2), ..., (S_m, y_m), where for each observation t, $S_t \subseteq N$ is a subset of items, and y_t is the item observed to be chosen from S_t; we refer to S_t as a *comparison set* and to y_t as a *choice*.

For every $S \subseteq N$ and $i \in S$, we denote with $w_{i,S}$ the number of observations such that the comparison set is S and the chosen item is i. In particular, for pair comparisons, we denote with $w_{i,j}$ the number of observations such that the comparison set is $\{i, j\}$ and the chosen item is i.

2.2. Generalized Thurstone Choice Model

A generalized Thurstone choice model, denoted as $\mathcal{T}_F$, is defined by a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ that takes value in a parameter set $\Theta_n \subseteq \mathbf{R}^n$, and a cumulative distribution function F of a zero-mean random variable that takes value in $\mathbf{R}$. Here θ_i represents the strength of item $i \in N$. We denote with f the density function of the cumulative distribution function F.

According to $\mathcal{T}_F$, the observations are such that for each observation, conditional on that the comparison set of this observation is S, the choice is item $i \in S$ with probability[1]

$$p_{i,S}(\theta) = p_{|S|}(\theta_i - \theta_{S\setminus\{i\}}) \tag{1}$$

where

$$p_k(\mathbf{x}) = \int_{\mathbf{R}} f(z) \prod_{l=1}^{k-1} F(x_l + z) dz, \text{ for } \mathbf{x} \in \mathbf{R}^{k-1}. \tag{2}$$

A generalized Thurstone model of choice $\mathcal{T}_F$ follows from the following probabilistic generative model. For every observation with comparison set S, each item in this set is associated with independent random variables $(X_i,\ i \in S)$ that represent individual performances of these items, where each X_i is a sum of θ_i and a zero-mean noise random variable ε_i with cumulative distribution function F. The choice $i \in S$ is the item that exhibits the largest performance, i.e. $p_{i,S}(\theta) = \mathbf{P}[X_i \geq \max_{j \in S} X_j]$, which corresponds to the asserted expression in (1).

Note that the probability distribution of choice depends only on the differences between the strength parameters. Hence, the probability distribution of choice for a parameter vector θ is equal to that under the parameter vector $\theta + c \cdot \mathbf{1}$, for any constant c, where $\mathbf{1}$ is the all-one vector. To allow for identifiability of the parameter vector, we admit the assumption that θ is such that $\sum_{i=1}^n \theta_i = 0$.

[1] Hereinafter, θ_A denotes the vector $\theta_A = (\theta_i, i \in A)$ for a non-empty set $A \subseteq N$, and, for brevity, with a slight abuse of notation, $a - \theta_A$ denotes the vector $(a - \theta_i, i \in A)$, for $a \in \mathbf{R}$.

2.3. Special Generalized Thurstone Choice Models

Several special generalized Thurstone models of choice are given as follows. (i) Gaussian noise with variance σ^2: $f(x) = \exp(-x^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)$. (ii) Double-exponential distribution of noise with parameter $\beta > 0$: $F(x) = \exp(-\exp(-(x + \beta\gamma)/\beta))$, where γ is the Euler-Mascheroni constant, which has variance $\sigma^2 = \pi^2\beta^2/6$. (iii) Laplace distribution of noise with parameter β: $F(x) = \frac{1}{2}e^{\frac{x}{\beta}}$, for $x < 0$, and $F(x) = 1 - \frac{1}{2}e^{-\frac{x}{\beta}}$, for $x \geq 0$, which has variance $\sigma^2 = 2\beta^2$. (iv) Uniform distribution of noise on $[-a, a]$: $f(x) = 1/(2a)$, for $x \in [-a, a]$, which has variance $\sigma^2 = a^2/3$.

For the special case of a generalized Thurstone model $\mathcal{T}_F$ with a double-exponential distribution of noise and a comparison set of cardinality k, we have

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

Hence, for a comparison set $S \subseteq N$,

$$p_{|S|}(\theta_i - \theta_{S\setminus\{i\}}) = \frac{e^{\theta_i/\beta}}{\sum_{l \in S} e^{\theta_l/\beta}}, \text{ for } i \in S,$$

which corresponds to the well-known Luce's choice model.

In particular, for pair comparisons, we have the following two well known cases: (i) for the Gaussian distribution of noise, we have $p_2(x) = \Phi(x/(\sqrt{2}\sigma))$ where Φ is the cumulative distribution function of a standard normal random variable, and (ii) for the double-exponential distribution of noise, we have $p_2(x) = 1/(1 + e^{-x/\beta})$, which is a special case of the Luce's choice model and is commonly referred as the Bradley-Terry model.

2.4. Maximum Likelihood Estimation

For given input observations, the log-likelihood function, up to an additive constant, is equal to

$$\ell(\theta) = \sum_{S \subseteq N} \sum_{i \in S} w_{i,S} \log(p_{|S|}(\theta_i - \theta_{S\setminus\{i\}})). \tag{3}$$

The maximum likelihood estimator of the parameter vector θ is defined as a parameter vector $\hat{\theta}$ that maximizes the log-likelihood function over the set of parameters Θ_n, i.e. $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta_n} \ell(\theta)$. In particular, for pair comparisons, we can write the log-likelihood function as follows:

$$\ell(\theta) = \sum_{i=1}^n \sum_{j=1}^n w_{i,j} \log\left(p_2(\theta_i - \theta_j)\right). \tag{4}$$

2.5. Some Key Definitions

We shall see that for the maximum likelihood parameter estimation problem, a special type of a matrix plays an important role. For every pair of items $\{i, j\}$ and a positive integer k, let $m_{i,j}(k)$ denote the number of observed comparison sets of cardinality k each containing the pair of items $\{i, j\}$. Let $w : \{1, 2, \ldots, m\} \to \mathbf{R}_+$ be a decreasing function, we refer to as a *weight function*, which is given. We define the *pair-weight matrix* $\mathbf{M} = [m_{i,j}] \in \mathbf{R}_+^{n \times n}$ as follows:

$$m_{i,j} = \begin{cases} \frac{n}{m} \sum_{k \geq 2} w(k) m_{i,j}(k), & \text{if } i \neq j \\ 0, & \text{if } i = j. \end{cases} \tag{5}$$

Note that if all comparison sets are of cardinality k, then each non-diagonal element (i, j) of the pair-weight matrix is equal to, up to a multiplicative factor, the number of observed comparison sets that contain the pair of items $\{i, j\}$. For pair comparisons, this corresponds to the number of pair comparisons. The normalization factor n/m corresponds to a normalization with the mean number of comparison sets per item.

We say that a set of comparison sets is *unbiased*, if for each positive integer k and pair of items $\{i, j\}$, there is a common number of comparison sets of cardinality k that contain the pair of items $\{i, j\}$.[2] Let $\mu(k)$ be the fraction of of comparison sets of cardinality k. Then, for any unbiased set of comparison sets, for every positive integer k and pair of items $\{i, j\}$, it must hold

$$m_{i,j}(k) = \frac{\binom{n-2}{k-2}}{\binom{n}{k}} \mu(k) m = \frac{k(k-1)}{n(n-1)} \mu(k) m.$$

Hence, for every pair of items $\{i, j\}$, it holds that

$$m_{i,j} = \frac{1}{n-1} \sum_{k \geq 2} w(k) k(k-1) \mu(k). \tag{6}$$

We shall use the notation $\overline{\mathbf{M}}$ to denote the expected value of a pair-weight matrix $\mathbf{M}$, where the expectation is with respect to the distribution over the set of comparison sets. We say that comparison sets are *a priori unbiased* if $\overline{\mathbf{M}}$ is an unbiased matrix. For example, sampling each comparison set independently by uniform random sampling without replacement from the set of all items results in an a priori unbiased set of comparison sets. Note that any unbiased set of comparison sets is a priori unbiased.

[2] An example of unbiased comparison sets is a fixture of games in some popular sport competitions that consists of games between pairs of teams such that each team plays against each other team equal number of times; e.g., fixtures of games in national football leagues like the one in Section 6.

We shall show that for the parameter estimation accuracy, the following parameters play an important role:

$$\gamma_{F,k} = \frac{1}{k^3(k-1)(\partial p_k(\mathbf{0})/\partial x_1)^2} \tag{7}$$

where

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx. \tag{8}$$

We shall see that the algebraic connectivity of pair-weight matrices is a key factor that determines the estimation accuracy, for a suitable choice of the weight function that depends on the generalized Thurstone choice model $\mathcal{T}_F$. In particular, we shall see that the weight function should be set as defined by[3]

$$w(k) = \left(k \frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \tag{9}$$

2.6. Additional Notation

For a matrix $\mathbf{A}$, we denote with $\lambda_i(\mathbf{A})$ its i-th smallest eigenvalue. We denote with $\Lambda_{\mathbf{A}}$ the Laplacian matrix of matrix $\mathbf{A}$, i.e., $\Lambda_{\mathbf{A}} = \mathrm{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$.

For any symmetric, non-negative, and irreducible matrix $\mathbf{A}$, its *Fiedler value* is defined as the smallest non-zero eigenvalue of the Laplacian matrix $\Lambda_{\mathbf{A}}$, i.e., equal to $\lambda_2(\Lambda_{\mathbf{A}})$.

3. Mean Square Error

In this section, we derive upper and lower bounds for the mean square error for the maximum likelihood parameter estimator of a generalized Thurstone choice model. For a generalized Thurstone choice model $\mathcal{T}_F$ with parameter $\theta^\star$, for any estimator $\hat{\theta}$, the mean square error $\mathrm{MSE}(\hat{\theta}, \theta^\star)$ is defined by

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) = \frac{1}{n} \|\hat{\theta} - \theta^\star\|_2^2. \tag{10}$$

3.1. Pair Comparisons

In this section, we consider generalized Thurstone models $\mathcal{T}_F$ for pair comparisons, with the parameter set $\Theta_n = [-b, b]^n$, for $b \geq 0$.

We define $G_D = (N, E_D)$ to be a directed graph, with edge $(i, j) \in E_D$ if and only if $w_{i,j} > 0$; and the undirected graph $G_U = (N, E_U)$ where edge $(i, j) \in E_U$ if and only if $w_{i,j} + w_{j,i} > 0$. Let $\mathbf{M}$ be the pair-weight matrix with the weight function $w(k) = 1/k^2$.

[3] For example, for the Luce's choice model this amounts to $w(k) = 1/(\beta k)^2$; for a large class of generalized Thurstone choice models, the weight function is such that $w(k) = \Theta(1/k^2)$; see discussion in Section 5.

G1. G_D is strongly connected, i.e., for every pair of vertices i and j, there exist paths in each direction between i and j.

G2. G_U is connected, i.e., for every pair of vertices i and j, there exists a path that connects them.

Note that if G_D is strongly connected, then G_U is connected, and the converse does not necessarily hold. Note also that when G_U is connected, i.e., condition **G2** holds true, then, $\lambda_2(\Lambda_{\mathbf{M}}) > 0$.

When $\log(p_2(x))$ is strictly concave for $x \in [-2b, 2b]$, i.e., $\max_{x\in[-2b,2b]} \frac{d^2}{dx^2} \log(p_2(x)) < 0$, $\ell(\theta)$ is strictly concave under **G2**.

Let us define $c_{F,b} = A/B$ where

$$A = \max_{x\in[-2b,2b]} \frac{d}{dx} \log(p_2(x))$$

and

$$B = \min_{x\in[-2b,2b]} \left| \frac{d^2}{dx^2} \log(p_2(x)) \right|.$$

Theorem 1. *Suppose that observations are according to a generalized Thurstone model $\mathcal{T}_F$ with parameter $\theta^\star \in [-b, b]^n$, for $n \geq 2$. If $\log(p_2(x))$ is a strictly concave function and* **G2** *holds, then with probability at least $1 - 2/n$, the maximum likelihood estimator $\hat{\theta}$ satisfies*

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq c_{F,b}^2 \frac{n(\log(n) + 2)}{\lambda_2(\Lambda_{\mathbf{M}})^2} \frac{1}{m}. \tag{11}$$

The result in Theorem 1 generalizes the characterization of the mean square error in (Negahban et al., 2012) and (Hajek et al., 2014) for the Bradley-Terry model to a generalized Thurstone choice model for pair comparisons.

Since the Bradley-Terry model is a generalized Thurstone choice model with noise according to the double-exponential distribution, we have $p_2(x) = 1/(1 + e^{-x/\beta})$, for which we derive $A = 1/[\beta(1 + e^{-2b/\beta})]$ and $B = e^{-2b/\beta}/[\beta^2(1+e^{-2b/\beta})^2]$, and hence $c_{F,b} = \beta(e^{2b/\beta} + 1)$.

Condition (11) implies that for $\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq \epsilon^2$ to hold for given $\epsilon > 0$, it suffices that

$$m \geq \frac{1}{\epsilon^2} c_{F,b}^2 \frac{1}{\lambda_2(\Lambda_{\mathbf{M}})^2} n(\log(n) + 2). \tag{12}$$

The Fiedler value $\lambda_2(\Lambda_{\mathbf{M}})$ reflects how well is the pair-weight matrix $\mathbf{M}$ connected. If each pair is compared an equal number of times, then from (6), we have $m_{i,j} = 1/(2(n-1))$ for $i \neq j$, and in this case, $\lambda_2(\Lambda_{\mathbf{M}}) = \cdots = \lambda_n(\Lambda_{\mathbf{M}}) = n/(2(n-1))$. Hence, from the condition in (12), it suffices that

$$m \geq \frac{4}{\epsilon^2} c_{F,b}^2 \, n(\log(n) + 2).$$

3.2. Arbitrary Cardinalities of Comparisons Sets

In this section, we derive upper and lower bounds for the mean square error when each comparison set consists of two or more items. Let K denote the set of distinct values of cardinalities of comparison sets observed in input data, or that can occur with a strictly positive probability if comparison sets are sampled from a distribution.

We consider a generalized Thurstone choice model $\mathcal{T}_F$ that satisfies the following assumptions:

A1 There exist $\overline{A}_{F,b} \geq \underline{A}_{F,b} > 0$ such that for all $S \subseteq N$ with $|S| \in K$ and $\{y, i, j\} \subseteq S$,

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

and, for all $\theta \in [-b, b]^n$, it holds

$$\underline{A}_{F,b} \leq \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log(p_{y,S}(\theta))}{\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log(p_{y,S}(\mathbf{0}))} \leq \overline{A}_{F,b}.$$

A2 There exist $\overline{B}_{F,b} \geq \underline{B}_{F,b} > 0$ such that for all $\theta \in [-b, b]^n$, $S \subseteq N$ with $|S| \in K$ and $y \in S$,

$$\underline{B}_{F,b} \leq \frac{p_{y,S}(\theta)}{p_{y,S}(\mathbf{0})} \leq \overline{B}_{F,b}.$$

A3 There exist $\overline{C}_{F,b} \geq \underline{C}_{F,b} > 0$ such that for all $\theta \in [-b, b]^n$, $S \subseteq N$ with $|S| \in K$ and $y \in S$,

$$\underline{C}_{F,b} \leq \frac{\|\nabla p_{y,S}(\theta)\|_2}{\|\nabla p_{y,S}(\mathbf{0})\|_2} \leq \overline{C}_{F,b}.$$

Let $D_{F,b} = \overline{C}_{F,b}/(\underline{A}_{F,b}\underline{B}_{F,b})$.[4]

The following theorem establishes an upper bound for the mean square error.

Theorem 2. *Assume* **A1**, **A2** *and* **A3**. *Let $\overline{\mathbf{M}}_F$ be the pair-weight matrix with the weight function (9). Suppose that*

$$m \geq 32 \frac{\sigma_{F,K}}{\underline{B}_{F,b}} \frac{1}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} n \log(n),$$

then, with probability at least $1 - 3/n$,

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq 32 D_{F,b}^2 \sigma_{F,K} \frac{n(\log(n) + 2)}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2} \frac{1}{m}$$

where $\sigma_{F,K} = 1/\min_{k\in K} \gamma_{F,k}$.

[4] (i) In particular, if F is the double-exponential distribution, then $\partial^2 \log(p_{y,S})/\partial\theta_i\partial\theta_j = p_{i,S}(\theta)p_{j,S}(\theta)/\beta^2 \geq 0$ and it is admissible to take $\underline{A}_{F,b} = e^{-4b/\beta}$, $\overline{A}_{F,b} = e^{4b/\beta}$, $\underline{B}_{F,b} = e^{-2b/\beta}$, $\overline{B}_{F,b} = e^{2b/\beta}$, $\underline{C}_{F,b} = e^{-4b/\beta}$, $\overline{C}_{F,b} = 4$, and $\sigma_{F,K} = 1/\beta^2$; (ii) In general, in the limit as b goes to 0, all the lower- and upper-bound parameters in **A1**, **A2**, **A3** go to 1. Thus, in this limit, they are non-essential for the results presented in this section.

If, in addition to the assumptions of Theorem 2, all comparison sets are of cardinality $k \geq 2$, then, the statement of the theorem holds with

$$\frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} = \left(1 - \frac{1}{k}\right) \frac{1}{\lambda_2(\Lambda_{\overline{\mathbf{M}}})}$$

and

$$\frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2} = \left(1 - \frac{1}{k}\right)^2 \gamma_{F,k} \frac{1}{\lambda_2(\Lambda_{\overline{\mathbf{M}}})^2}$$

where $\gamma_{F,k}$ is defined in (7), and $\overline{\mathbf{M}}$ is the pair-weight matrix with the weight function $w(k) = 1/k^2$.

If, in addition, each comparison set is sampled independently, uniformly at random without replacement from the set of all items, then the statement of Theorem 2 holds with

$$\frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} = 1 - \frac{1}{n} \text{ and } \frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2} = \left(1 - \frac{1}{n}\right)^2 \gamma_{F,k}.$$

In the following theorem, we establish a lower bound.

Theorem 3. *Any unbiased estimator $\hat{\theta}$ satisfies*

$$\mathbf{E}[\mathrm{MSE}(\hat{\theta}, \theta^\star)] \geq \frac{1}{\overline{A}_{F,b}\overline{B}_{F,b}} \left(\sum_{i=2}^{n} \frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})}\right) \frac{1}{m}.$$

If all comparison sets are of cardinality k, then any unbiased estimator $\hat{\theta}$ satisfies the inequality in Theorem 3 with

$$\sum_{i=2}^{n} \frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})} = \left(1 - \frac{1}{k}\right) \gamma_{F,k} \sum_{i=2}^{n} \frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}})}.$$

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 the inequality in Theorem 3 with[5]

$$\sum_{i=2}^{n} \frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})} = \gamma_{F,k} \left(1 - \frac{1}{n}\right)^2 n.$$

4. Classification of Items of Two Classes

In this section, we consider a generalized Thurstone choice model $\mathcal{T}_F$ with parameter θ that takes value in $\Theta_n = \{-b, b\}^n$, for parameter $b > 0$. This is a special case where each item is either of two classes: a low or a high class. We consider a classification problem, where the goal is to correctly classify each item as either of low or high class, based on observed input data of choices.

[5]This tells us that under the given assumptions, for the mean square error to be smaller than a constant, it is necessary that the number of comparisons satisfies $m = \Omega(\gamma_{F,k} n)$.

Suppose that $\theta_i = b$ for all $i \in N_1$ and $\theta_i = -b$ for all $i \in N_2$ where $N_1 \cup N_2 = N$ and $|N_1| = |N_2| = n/2$. Without loss of generality, assume that $N_1 = \{1, \ldots, n/2\}$ and $N_2 = \{n/2 + 1, \ldots, n\}$.

We consider a *point score ranking method* that outputs an estimate $\hat{N}_1$ of the set of items of high class and $\hat{N}_2$ that contains the remaining items, which is defined by the following algorithm:

1. Observe outcomes of m observations and associate each item with a point score defined as the number of comparison sets in which this item is the chosen item.
2. Sort items in decreasing order of the point scores.
3. Output $\hat{N}_1$ defined as the set of top $n/2$ items (with uniform random tie break) and $\hat{N}_2$ defined as the set of remaining items.

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

$$b \max_{\mathbf{x} \in [-2b, 2b]^{k-1}} \|\nabla^2 p_k(\mathbf{x})\|_2 \leq \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \tag{13}$$

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

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

the point score ranking method correctly identifies the classes of all items with probability at least $1 - \delta$.

The bound of the theorem is tight as established in the following theorem.

Theorem 5. *Suppose that $b \leq 1/(6k^2 \partial p_k(\mathbf{0})/\partial x_1)$ and that condition (13) holds. Then, for every even number of items such that $n \geq 16$, and $\delta \in (0, 1/4]$, for any algorithm to correctly classify all items with probability at least $1 - \delta$, it is necessary that*

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

5. Discussion of Results

In this section, we discuss how the number of observations needed for given parameter estimation error tolerance depends on the cardinality of comparison sets. We found in Section 3.2 and Section 4 that for a priori unbiased schedules of comparisons, where each comparison set is of cardinality k and is drawn independently, uniformly at random from the set of all items, the required number of observations to bring down the mean square error or correctly classify items of two classes with high probability, the number of observations is of the order $\gamma_{F,k}$, defined in (7).

The values of parameters $\partial p_k(\mathbf{0})/\partial x_1$ and $\gamma_{F,k}$ for our example generalized Thurstone choice models $\mathcal{T}_F$ in Section 2.3 are presented in Table 1.

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	$\frac{1}{2a(k-1)}$	$4a^2 \frac{k-1}{k^3}$

Note that for both double-exponential and Laplace distributions 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, $\gamma_{F,k} = \Theta(1/k^2)$.

In general, the value of parameter $\gamma_{F,k}$ admits the following lower and upper bounds.

Proposition 6. *For the value of parameter $\gamma_{F,k}$, the following two claims hold:*

1. *For every cumulative distribution function F with an even and continuously differentiable density function, we have $\gamma_{F,k} = O(1)$.*

2. *For every cumulative distribution function F with a density function such that $f(x) \leq C$ for all $x \in \mathbf{R}$, for a constant $C > 0$, $\gamma_{F,k} = \Omega(1/k^2)$.*

We observe that both double-exponential and Laplace distributions of noise are extremal in achieving the upper bound of $O(1)$ for the value of parameter $\gamma_{F,k}$, asymptotically for large k. On the other hand, a uniform distribution of noise is extremal in achieving the lower bound $\Omega(1/k^2)$ for the value of parameter $\gamma_{F,k}$. More generally, we can show that $\gamma_{F,k} = \Theta(1/k^2)$ for any cumulative distribution function F with the density function such that $f(x) \geq C$ for every point x of its support, for a constant $C > 0$.

6. 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 square error can depend on the cardinality of comparison sets in different ways depending on the choice of a generalized Thurstone model, which is suggested by our theory. Our second goal is to evaluate Fiedler values of different pair-weight matrices observed in practice, which we found to play an important role.

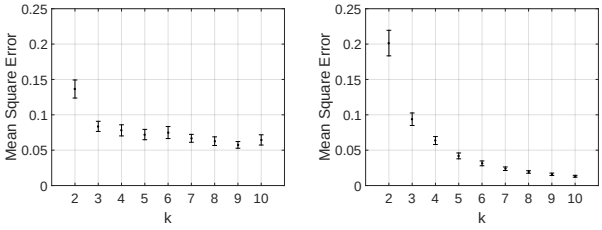

Figure 1. Mean square error for two different generalized Thurstone choice models $\mathcal{T}_F$: (left) F is a double-exponential distribution, and (right) F is a uniform distribution. The vertical bars denote 95% confidence intervals. The results confirm two qualitatively different relations with the cardinality of comparison sets as suggested by the theory.

6.1. MSE versus Cardinality of Comparison Sets

We consider the following simulation experiment. We fix the values of the number of items n and the number of comparisons m, and consider a choice of a generalized Thurstone model $\mathcal{T}_F$ for the value of parameter $\theta^\star = \mathbf{0}$. We consider comparison sets of the same cardinality of value k that are independent uniform random samples from the set of all items. For every fixed value of k, we run 100 repetitions to estimate the mean square 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 setting of parameters $n = 10$ and $m = 100$. The results clearly demonstrate that the mean square error exhibits qualitatively different relations with the cardinality of comparison sets for the two generalized Thurstone models. Our theoretical results in Section 3.2 suggest that the mean square error should decrease with the cardinality of comparison sets as $1/(1 - 1/k)$ for the double-exponential distribution, 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.

6.2. Fiedler Values of Pair-weight Matrices

We found that Fiedler value of a pair-weight matrix is an important factor that determines the mean square error in Section 3.1 and Section 3.2. Here we evaluate Fiedler value for different pair-weight matrices of different schedules of comparisons. Throughout this section, we use the definition of a pair-weight matrix in (5) 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

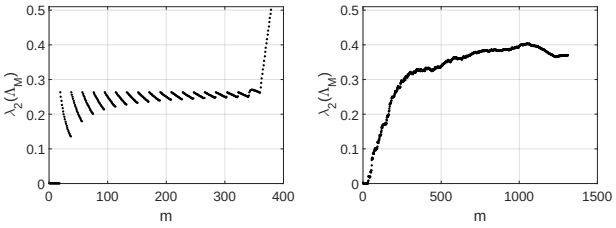

Figure 2. Fiedler value of the pair-weight matrices for the game fixtures of two sports in the season 2014-2015: (left) football Barclays premier league, and (right) basketball NBA league.

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 games, and 81 playoff games.[6] We evaluate Fiedler value of pair-weight 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 pair-weight 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 pair-weight 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 Tackcn. We refer to coders in TopCoder and workers in Taskcn as users. We consider contests of different categories observed in year 2012; more information about datasets is provided in Appendix. We present results only for one category of tasks for each system, which are representative. In both these systems, the participation in contests is according to choices made by users.

[6]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.

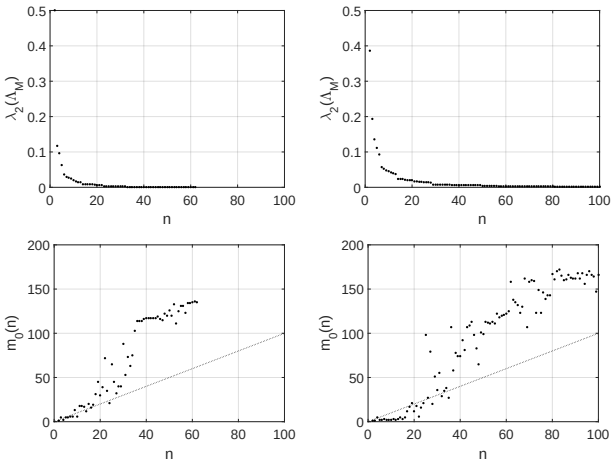

Figure 3. (Left) Topcoder data restricted to top-n coders and (Right) same as left but for Taskcn, for Design and Website task categories, respectively. The top plots show the Fiedler value and the bottom plots show the minimum number of contests to observe a strictly positive Fiedler value.

For each set of tasks of given category, we conduct the following analysis. We consider a thinned dataset that consists only of a set of top-n users with respect to the number 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 pair-weight 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 pair-weight 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 pair-weight 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 pair-weight matrices tends to decrease with the number of top users considered.

7. Conclusion

The results of this paper elucidate how the parameter estimation accuracy for a generalized Thurstone choice model depends on the given model and the structure of comparison sets. They show that a key factor is an eigenvalue gap of a pair-weight matrix that reflects its algebraic connectivity, which depends in a particular way on the given model.

It is shown that for a large class of generalized Thurstone choice models, including all popular instances used in practice, there is a diminishing returns decrease of the estimation error with the cardinality of comparison sets, which is rather slow for comparison sets of three of more items. This offers a guideline for the designers of schedules of competitions to ensure that the schedule has a well-connected pair-weight matrix and to expect limited gains from comparison sets of large sizes.

References

Boyd, Stephen. Convex optimization of graph Laplacian eigenvalues. In *Proceedings of the International Congress of Mathematicians*, pp. 1311–1319, 2006.

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: I. method of paired comparisons. *Biometrika*, 39(3/4):324–345, Dec 1952.

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: II. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4): 502–537, Dec 1954.

Dawkins, Richard. A threshold model of choice behaviour. *Animal Behaviour*, 17(Part 1):120–133, February 1969.

Elo, Arpad E. *The Rating of Chessplayers*. Ishi Press International, 1978.

Graepel, Thore, Minka, Tom, and Herbrich, Ralf. Trueskill(tm): A bayesian skill rating system. In *Proc. of NIPS 2006*, volume 19, pp. 569–576, 2006.

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

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

Luce, R. Duncan. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman & Hall, New York, 2 edition, 1989.

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Negahban, Sahand, Oh, Sewoong, and Shah, Devavrat. Iterative ranking from pair-wise comparisons. In *Proc. of NIPS 2012*, pp. 2483–2491, 2012.

Nelder, J. A. and Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.

Rajkumar, Arun and Agarwal, Shivani. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proc. of ICML 2014*, pp. 118–126, 2014.

Stern, Hal. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1): 103–117, 1992.

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 34(2):273–286, 1927.

Tropp, Joel A. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Vojnović, Milan. *Contest Theory: Incentive Mechanisms and Ranking Methods*. Cambridge University Press, 2016.

Yellott, John I. 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.

Zermelo, E. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Math. Z.*, 29:436–460, 1929.

A. Some Basic Facts

A.1. Cramér-Rao Inequality

Let $\mathrm{cov}[Y]$ denote the covariance matrix of a multivariate random variable Y, i.e., $\mathrm{cov}[Y] = \mathbf{E}[(Y - \mathbf{E}[Y])(Y - \mathbf{E}[Y])^\top]$.

Proposition 7 (Cramér-Rao inequality)**.** *Suppose that X is a multivariate random variable with distribution $p(x;\theta)$, for parameter $\theta \in \Theta_n$, and let $\mathbf{T}(X) = (T_1(X), \ldots, T_r(X))^\top$ be any unbiased estimator of $\psi(\theta) = (\psi_1(\theta), \ldots, \psi_r(\theta))^\top$, i.e., $\psi(\theta) = \mathbf{E}[\mathbf{T}(X)]$. Then, we have*

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

where $\frac{\partial \psi(\theta)}{\partial \theta}$ is the Jacobian matrix of ψ and $F(\theta)$ is the Fisher information matrix with (i,j) element defined by

$$F_{i,j}(\theta) = \mathbf{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j}(-\log(p(X;\theta)))\right].$$

A.2. Azuma-Hoeffding's Inequality for Vectors

The inequality is known as the Azuma-Hoeffding's inequality for multivariate random Variables, which was established in Theorem 1.8 (Hayes).

Proposition 8 (Azuma-Hoeffding's inequality)**.** *Suppose that $S_m = \sum_{t=1}^m X_t$ is a martingale where $X_1, X_2, \ldots, X_m$ take values in $\mathbf{R}^n$ and are such that $\mathbf{E}[X_t] = \mathbf{0}$ and $\|X_t\|_2 \leq D$ for all t, for $D > 0$. Then, for every $x > 0$,*

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

A.3. Chernoff's Inequality for Matrices

The inequality is known as the Chernoff's inequalities for random matrices; e.g. stated as Theorem 5.1.1 in (Tropp, 2015).

Proposition 9 (Matrix Chernoff's inequality)**.** *Let $X_1, X_2, \ldots, X_m$ be a finite sequence of independent, random, Hermitian matrices with dimension d. Assume that*

$$0 \leq \lambda_1(X_i) \quad \textit{and} \quad \|X_i\|_2 \leq \alpha \quad \textit{for all} \quad i.$$

Let

$$\beta_{\min} = \lambda_1\left(\sum_{i=1}^m \mathbf{E}[X_i]\right)$$

and

$$\beta_{\max} = \lambda_d\left(\sum_{i=1}^m \mathbf{E}[X_i]\right).$$

Then, for $\varepsilon \geq 0$,

$$\mathbf{P}\left[\lambda_d\left(\sum_{i=1}^m X_i\right) \geq (1+\varepsilon)\beta_{\max}\right] \leq d\left(\frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\beta_{\max}/\alpha} \quad \textit{for } \varepsilon \geq 0 \tag{14}$$

and, for $\epsilon \in [0,1)$,

$$\mathbf{P}\left[\lambda_1\left(\sum_{i=1}^m X_i\right) \leq (1-\varepsilon)\beta_{\min}\right] \leq d\left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\beta_{\min}/\alpha} \quad \textit{for } \varepsilon \in [0,1). \tag{15}$$

We have the following corollary:

Corollary 10. *Under the assumptions of Proposition 9, for $\varepsilon \in [0,1)$,*

$$\mathbf{P}\left[\lambda_1\left(\sum_{i=1}^m X_i\right) \leq (1-\varepsilon)\beta_{\min}\right] \leq de^{-\frac{\varepsilon^2 \beta_{min}}{2\alpha}}$$

Proof. This follows from (15) and the following fact

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

□

A.4. A Chernoff's Tail Bound

The following tail bound follows from the Chernoff's bound and is proved in Appendix L.3.

Proposition 11. *Suppose that X is a sum of m independent Bernoulli random variables each with mean p, then if $q \leq p \leq 2q$,*

$$\mathbf{P}[X \leq qm] \leq \exp\left(-\frac{(q-p)^2}{4q}m\right) \tag{16}$$

and, if $p \leq q$,

$$\mathbf{P}[X \geq qm] \leq \exp\left(-\frac{(q-p)^2}{4q}m\right). \tag{17}$$

A.5. Properties of Laplacian Matrices

If $\mathbf{A}$ is a symmetric non-negative matrix and the diagonal of $\mathbf{A}$ is zero, we have the following properties (Boyd, 2006):

$$0 = \lambda_1(\Lambda_{\mathbf{A}}) \leq \cdots \leq \lambda_n(\Lambda_{\mathbf{A}})$$

and

$$\lambda_{i+1}(\Lambda_{\mathbf{A}}) = \lambda_i(\mathbf{Q}_{\mathbf{1}}^\top \Lambda_{\mathbf{A}} \mathbf{Q}_{\mathbf{1}}) \text{ for } i = 1, 2, \ldots, n-1 \quad (18)$$

where $\mathbf{Q_1} \in \mathbf{R}^{n \times n-1}$ denotes a matrix whose columns are orthonormal to the all-one vector $\mathbf{1}$.

From (18), we have for all symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ with zero diagonals, it holds that

$$\Lambda_{\mathbf{A}} \succeq \Lambda_{\mathbf{B}} \quad \text{when} \quad \mathbf{A} \succeq \mathbf{B}, \quad (19)$$

where $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive semi-definite matrix.

B. $\mathcal{T}_F$ Log-likelihoods

Let $H_{i,j}(\theta, S)$ be defined for $\theta \in \Theta_n$, $S \subseteq N$ and $i \in S$, $j \in S$, as follows

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

Lemma 12. *For every comparison set $S \subseteq N$, we have*

$$H_{i,j}(\mathbf{0}, S) = \begin{cases} k^2(k-1)\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 & \text{if} \quad i = j \\ -k^2\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 & \text{if} \quad i \neq j. \end{cases} \quad (20)$$

Proof of the last above lemma is provided in Appendix L.1.

Lemma 13. *Let $S \subseteq N$ and $y \in S$ and let k be the cardinality of set S. Then, it holds*

1. $\mathbf{1}^\top \nabla^2(-\log(p_{y,S}(\mathbf{0}))) = 0$, *and*
2. $\frac{1}{k}\sum_{y \in S} \nabla^2(-\log(p_{y,S}(\mathbf{0}))) = \Lambda_{\mathbf{M}_S} k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2$

where $\mathbf{M}_S$ denotes a matrix that has all (i,j) elements such that $\{i,j\} \subseteq S$ equal to 1, and all other elements equal to 0.

Proof of the last above lemma follows easily from that of Lemma 12.

Lemma 14. *If for a comparison set $S \subseteq N$ of cardinality k, $\nabla^2(-\log(p_{y,S}(\mathbf{0})))$ is a positive semi-definite matrix, then it holds that*

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

Proof of the last above lemma is given in Appendix L.2.

C. Proof of Theorem 1

Let $\Delta = \hat{\theta} - \theta^\star$. By the Taylor expansion, we have

$$\ell(\hat{\theta}) \leq \ell(\theta^\star) + \nabla \ell(\theta^\star)^\top \Delta + \frac{1}{2} \max_{\alpha \in [0,1]} \Delta^\top \nabla^2 \ell(\theta^\star + \alpha\Delta)\Delta. \quad (21)$$

Note that Δ is orthogonal to the all-one vector, i.e., $\sum_{i=1}^n \Delta_i = 0$.

By the Cauchy-Schwartz inequality, we have

$$\nabla \ell(\theta^\star)^\top \Delta \leq \|\nabla \ell(\theta^\star)\|_2 \|\Delta\|_2. \quad (22)$$

Since $\hat{\theta}$ is a maximum likelihood estimator, we have

$$\ell(\hat{\theta}) - \ell(\theta^\star) \geq 0. \quad (23)$$

From (21), (22) and (23),

$$-\max_{\alpha \in [0,1]} \Delta^\top \nabla^2 \ell(\theta^\star + \alpha\Delta)\Delta \leq 2\|\nabla \ell(\theta^\star)\|_2 \|\Delta\|_2. \quad (24)$$

Now, note that for every $\theta \in \mathbf{R}^n$ and $i, j \in N$,

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

Hence, for every $\theta \in \Theta_n$ and $\mathbf{x} \in \mathbf{R}^n$, we have

$$\begin{aligned} &\mathbf{x}^\top \nabla^2 \ell(\theta) \mathbf{x} \\ &= \sum_{i=1}^n \sum_{j \neq i} w_{i,j}(x_i - x_j)^2 \frac{d^2}{dx^2}\log(p_2(\theta_i - \theta_j)) \\ &\leq -\sum_{i=1}^n \sum_{j \neq i} w_{i,j} B (x_i - x_j)^2 \\ &= -B\frac{4m}{n}\mathbf{x}^\top \Lambda_{\mathbf{M}} \mathbf{x} \\ &\leq -B\frac{4m}{n}\|\mathbf{x}\|_2^2 \lambda_2(\Lambda_{\mathbf{M}}), \end{aligned} \quad (25)$$

where we deduce the last inequality from (18).

From (24) and (25), we obtain

$$\frac{2Bm}{n}\|\Delta\|_2^2 \lambda_2(\Lambda_{\mathbf{M}}) \leq \|\nabla \ell(\theta^\star)\|_2 \|\Delta\|_2. \quad (26)$$

We bound $\|\nabla \ell(\theta^\star)\|_2$ using the Azuma-Hoeffding's inequality for multivariate random variables in Proposition 8.

Note that $\nabla \ell(\theta^\star)$ is a sum of m independent random vectors having zero-mean, where each comparison of a pair of items (i,j) results in a vector of value $\nabla \log(p_2(\theta_i^\star - \theta_j^\star))$ with probability $p_2(\theta_i^\star - \theta_j^\star)$ and of value $\nabla \log(p_2(\theta_j^\star - \theta_i^\star))$ with probability $p_2(\theta_j^\star - \theta_i^\star)$. Note that for every pair of items (i,j), $\|\nabla \log(p_2(\theta_i^\star - \theta_j^\star))\|_2 \leq A\sqrt{2}$.

By the Azuma-Hoeffding's inequality in Proposition 8, it follows that

$$\mathbf{P}\left[\|\nabla \ell(\theta^\star)\|_2 \geq 2A\sqrt{m(\log(n)+2)}\right] \leq \frac{2}{n}. \quad (27)$$

Finally, from (26), (22) and (27), with probability $1-2/n$, it holds

$$\|\Delta\|_2 \leq \frac{An\sqrt{(\log(n)+2)}}{B\lambda_2(\Lambda_{\mathbf{M}})\sqrt{m}}.$$

D. Proof of Theorem 2

This proof follows the main steps of the proof of Theorem 1. Let $\Delta = \hat{\theta} - \theta^\star$. By the same arguments as in the proof of Theorem 1, we have that equation (24) holds, i.e.,

$$-\max_{\alpha\in[0,1]} \Delta^\top \nabla^2 \ell(\theta^\star + \alpha\Delta)\Delta \leq 2\|\nabla \ell(\theta^\star)\|_2 \|\Delta\|_2. \quad (28)$$

Since $\nabla^2(-\ell(\theta))$ is a Laplacian matrix, from assumption **A1** and (19), we have

$$\nabla^2(-\ell(\theta)) \succeq \underline{A}_{F,b} \nabla^2(-\ell(\mathbf{0})) \text{ for all } \theta \in [-b,b]^n. \quad (29)$$

From (28) and (29), we obtain

$$\begin{aligned} &\lambda_1\left(\mathbf{Q_1}^\top \nabla^2(-\ell(\mathbf{0}))\mathbf{Q_1}\right) \underline{A}_{F,b} \|\Delta\|_2 \\ &\leq 2\|\nabla \ell(\theta^\star)\|_2, \end{aligned} \quad (30)$$

which follows by the fact that $\hat{\theta}$ is orthogonal to $\mathbf{1}$.

We state two lemmas whose proofs are given at the end of this section, in Appendix D.1 and D.2.

Lemma 15. *Suppose that*

$$m \geq 32 \frac{\sigma_{F,K}}{\underline{B}_{F,b}\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} n\log(n)$$

then, with probability at least $1-1/n$,

$$\lambda_1\left(\mathbf{Q_1}^\top \nabla^2(-\ell(\mathbf{0}))\mathbf{Q_1}\right) \geq \frac{\underline{B}_{F,b} m}{2n} \lambda_2(\Lambda_{\overline{\mathbf{M}}_F}).$$

and

Lemma 16. *With probability at least* $1-2/n$, *it holds that*

$$\|\nabla \ell(\theta^\star)\|_2 \leq \overline{C}_{F,b}\sqrt{\sigma_{F,K}}\sqrt{2m(\log(n)+2)}.$$

From (30) and the bounds in Lemma 15 and Lemma 16, it follows that if

$$m \geq 32 \frac{\sigma_{F,K}}{\underline{B}_{F,b}\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} n\log(n),$$

then, with probability at least $1-3/n$,

$$\|\Delta\|_2 \leq 32\left(\frac{\overline{C}_{F,b}}{\underline{A}_{F,b}\underline{B}_{F,b}}\right)^2 \sigma_{F,K} \frac{n(\log(n)+2)}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2} \frac{1}{m}.$$

D.1. Proof of Lemma 15

From the definition of the log-likelihood function $\ell(\theta)$, $\mathbf{Q_1}^\top \nabla^2(-\ell(\mathbf{0}))\mathbf{Q_1}$ is a sum of a sequence of random matrices $\{\mathbf{Q_1}^\top \nabla^2(-\log(p_{y_t,S_t}(\mathbf{0})))\mathbf{Q_1}\}_{1\leq t\leq m}$, i.e.,

$$\mathbf{Q_1}^\top \nabla^2(-\ell(\mathbf{0}))\mathbf{Q_1} = \sum_{t=1}^m \mathbf{Q_1}^\top \nabla^2(-\log(p_{y_t,S_t}(\mathbf{0})))\mathbf{Q_1}.$$

From assumption **A1** and (18), for every observation t,

$$\lambda_1\left(\mathbf{Q_1}^\top \nabla^2(-\log(p_{y_t,S_t}(\mathbf{0})))\mathbf{Q_1}\right) \geq 0.$$

We can thus apply the matrix Chernoff's inequality, given in Proposition 9, once we find a lower bound for $\lambda_1\left(\mathbf{E}\left[\mathbf{Q_1}^\top \nabla^2(-\ell(\mathbf{0}))\mathbf{Q_1}\right]\right)$ and an upper bound for $\left\|\mathbf{Q_1}^\top\left(\nabla^2 \log(p_{y_t,S_t}(\mathbf{0}))\right)\mathbf{Q_1}\right\|_2$ for every observation t.

We have the following sequence of relations

$$\begin{aligned} &\mathbf{E}_{\theta^\star}\left[\nabla^2(-\log(\ell(\mathbf{0})))\right] \\ =\ &\sum_{t=1}^m \mathbf{E}_{\theta^\star}\left[\nabla^2(-\log(p_{y_t,S_t}(\mathbf{0})))\right] \\ =\ &\sum_{t=1}^m \sum_{y\in S_t} p_{y,S_t}(\theta^\star)\nabla^2(-\log(p_{y,S_t}(\mathbf{0}))) \\ \succeq\ &\underline{B}_{F,b}\sum_{t=1}^m \sum_{y\in S_t} \frac{1}{|S_t|}\nabla^2(-\log(p_{y,S_t}(\mathbf{0}))) \\ =\ &\underline{B}_{F,b}\sum_{t=1}^m \sum_{y\in S_t} \frac{1}{|S_t|}\Lambda_{\mathbf{M}_{S_t}}|S_t|^2\left(\frac{\partial p_{|S_t|}(\mathbf{0})}{\partial x_1}\right)^2 \quad (31) \\ =\ &\underline{B}_{F,b}\frac{m}{n}\Lambda_{\overline{\mathbf{M}}_F} \quad (32) \end{aligned}$$

where (31) follows Lemma 13 and $\mathbf{M}_S$ denotes a matrix that has all (i,j) elements such that $\{i,j\} \subseteq S$ equal to 1, and all other elements equal to 0.

From (32), we have

$$\begin{aligned} &\lambda_1\left(\mathbf{E}\left[\mathbf{Q_1}^\top \nabla^2(-\ell(\mathbf{0}))\mathbf{Q_1}\right]\right) \\ \geq\ &\underline{B}_{F,b}\frac{m}{n}\lambda_1\left(\mathbf{Q_1}^\top \Lambda_{\overline{\mathbf{M}}_F}\mathbf{Q_1}\right) \\ =\ &\underline{B}_{F,b}\frac{m}{n}\lambda_2(\Lambda_{\overline{\mathbf{M}}_F}), \quad (33) \end{aligned}$$

where the last equality holds by (18).

From Lemma 14, for every observation t,

$$\left\|\nabla^2 \log(p_{y_t,S_t}(\mathbf{0}))\right\|_2 \leq \frac{2}{\gamma_{F,|S_t|}} \leq 2\sigma_{F,K}. \tag{34}$$

Using the matrix Chernoff's inequality in Corollary 10 with $\varepsilon = 1/2$, $\beta_{min} \geq \underline{B}_{F,b}\frac{m}{n}\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})$ by (33) and $\alpha \leq \sigma_{F,K}$ by (34), we obtain the assertion of the lemma.

D.2. Proof of Lemma 16

For every comparison set $S \subseteq N$ and $i \in S$, we have

$$\frac{\partial \log p_{i,S}(\theta)}{\partial \theta_i} = -\frac{1}{p_{i,S}(\theta)} \sum_{v \in S\setminus\{i\}} \frac{\partial p_{v,S}(\theta)}{\partial \theta_i} \tag{35}$$

and, for all $j \in S \setminus \{i\}$,

$$\frac{\partial \log p_{j,S}(\theta)}{\partial \theta_i} = \frac{1}{p_{j,S}(\theta)} \frac{\partial p_{j,S}(\theta)}{\partial \theta_i}. \tag{36}$$

From (35) and (36), we have

$$\mathbf{E}\left[\nabla \log p_{y,S}(\theta^\star)\right] = \mathbf{0} \tag{37}$$

and

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

By assumption **A3**, every $S \subseteq N$ such that $|S| \in K$,

$$\begin{aligned} \|\nabla \log p_{y,S}(\theta^\star)\|_2^2 &\leq \overline{C}_{F,b}^2 \|\nabla \log p_{y,S}(\mathbf{0})\|_2^2 \\ &\leq \overline{C}_{F,b}^2 \sigma_{F,K}. \end{aligned} \tag{38}$$

Using (37) and (38) with the Azuma-Hoeffding inequality for multivariate random variables in Proposition 8, we obtain that with probability at least $1 - 2/n$,

$$\|\nabla \ell(\theta^\star)\|_2 \leq \overline{C}_{F,b}\sqrt{\sigma_{F,K}}\sqrt{2m(\log(n)+2)}.$$

E. Remark for Theorem 2

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

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

Since

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

and

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

we have that

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

and

$$e^{-4b/\beta} \leq \underline{A}_{F,b} \leq \overline{A}_{F,b} \leq e^{4b/\beta}, \tag{39}$$

$$e^{-2b/\beta} \leq \underline{B}_{F,b} \leq \overline{B}_{F,b} \leq e^{2b/\beta}, \tag{40}$$

$$e^{-4b/\beta} \leq \underline{C}_{F,b} \leq \overline{C}_{F,b} \leq 4. \tag{41}$$

F. Proof of Theorem 3

The proof of the theorem follows from the well-known Cramér-Rao inequality, which is given in Proposition 7.

Since $\sum_{i=1}^n \theta_i = 0$, we define $\psi_i(\theta) = \theta_i - \frac{1}{n}\sum_{l=1}^n \theta_l$. Note that $\sum_{i=1}^n \psi_i(\theta) = 0$. Then,

$$\frac{\partial \boldsymbol{\psi}(\theta)}{\partial \theta} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top. \tag{42}$$

Let $F(\theta)$ be the Fisher information matrix of the random vector $X = (\mathbf{S}, \mathbf{y})$ where $\mathbf{S} = (S_1, S_2, \ldots, S_m)$ are the comparison sets and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ are the choices of comparisons.

Then, we have that the (i,j) element of matrix $F(\theta)$ is given by

$$F(\theta) = \sum_{t=1}^m \mathbf{E}\left[\nabla^2(-\log(p_{y_t,S_t}(\theta)))\right]. \tag{43}$$

From the assumptions **A1**, **A2**, and Lemma 13, we have

$$\begin{aligned}
&\mathbf{E}\left[\nabla^2(-\log(p_{y_t,S_t}(\theta)))\middle|\, S_t = S\right] \\
&= \sum_{y\in S} p_{y,S}(\theta)\nabla^2(-\log(p_{y,S}(\theta))) \\
&\preceq \sum_{y\in S} \frac{\overline{B}_{F,b}}{|S|}\nabla^2(-\log(p_{y,S}(\theta))) \\
&\preceq \sum_{y\in S} \frac{\overline{A}_{F,b}\overline{B}_{F,b}}{|S|}\nabla^2(-\log(p_{y,S}(\mathbf{0}))) \\
&= \overline{A}_{F,b}\overline{B}_{F,b}\left(|S|\frac{\partial p_{|S|}(\mathbf{0})}{\partial x_1}\right)^2 \Lambda_{\mathbf{M}_S} \qquad (44)
\end{aligned}$$

where we use (19) for the two inequalities and $\mathbf{M}_S$ that has each element (i,j) such that $\{i,j\} \subseteq S$ equal to 1 and all other elements equal to 0.

From (43) and (44),

$$F(\theta) \preceq \overline{A}_{F,b}\overline{B}_{F,b}\frac{m}{n}\Lambda_{\overline{\mathbf{M}}_F}. \qquad (45)$$

For a $n \times n$ matrix $\mathbf{A} = [a_{i,j}]$, let $\mathrm{tr}(\mathbf{A})$ denote its trace, i.e. $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^n a_{i,i}$. Note that

$$\begin{aligned}
\mathbf{E}[\|\hat{\theta}-\theta\|_2^2] &= \mathrm{tr}(\mathrm{cov}[\mathbf{T}(X)]) \\
&= \sum_{i=1}^n \lambda_i\left(\mathrm{cov}[\mathbf{T}(X)]\right).
\end{aligned}$$

By the Cramér-Rao bound and (45), we have

$$\begin{aligned}
\frac{1}{n}\mathbf{E}[\|\hat{\theta}-\theta\|_2^2] &\geq \frac{1}{n}\sum_{i=1}^n \lambda_i\left(\frac{\partial\boldsymbol{\psi}(\theta)}{\partial\theta}F^{-1}(\theta)\frac{\partial\boldsymbol{\psi}(\theta)}{\partial\theta}^\top\right) \\
&= \frac{1}{n}\sum_{i=1}^{n-1}\lambda_i(\mathbf{Q_1}^\top F^{-1}(\mathbf{0})\mathbf{Q_1}) \\
&= \frac{1}{n}\sum_{i=1}^{n-1}\frac{1}{\lambda_i(\mathbf{Q_1}^\top F(\mathbf{0})\mathbf{Q_1})} \\
&\geq \frac{1}{\overline{A}_{F,b}\overline{B}_{F,b}m}\sum_{i=1}^{n-1}\frac{1}{\lambda_i(\mathbf{Q_1}^\top\Lambda_{\overline{\mathbf{M}}_F}\mathbf{Q_1})} \\
&= \frac{1}{\overline{A}_{F,b}\overline{B}_{F,b}m}\sum_{i=2}^{n}\frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})},
\end{aligned}$$

where the last equality is obtained from (18).

G. Proof of Theorem 4

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

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

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

$$\begin{aligned}
p^e &\leq \mathbf{P}\left[\bigcup_{l\in N_1}\left\{R_l \leq \frac{m}{n}\right\}\cup\bigcup_{l\in N_2}\left\{R_l\geq\frac{m}{n}\right\}\right] \\
&\leq \sum_{l\in N_1}\mathbf{P}\left[R_l\leq\frac{m}{n}\right]+\sum_{l\in N_2}\mathbf{P}\left[R_l\geq\frac{m}{n}\right]. \qquad (46)
\end{aligned}$$

Let i and j be arbitrarily fixed items such that $i \in N_1$ and $j \in N_2$. We will show that for every observation t,

$$\mathbf{P}[y_t = i] \geq \frac{1}{n} + \frac{bk^2}{4n}\frac{\partial p_k(\mathbf{0})}{\partial x_1} \qquad (47)$$

and

$$\mathbf{P}[y_t = j] \leq \frac{1}{n} - \frac{bk^2}{4n}\frac{\partial p_k(\mathbf{0})}{\partial x_1}. \qquad (48)$$

From the Chernoff's bound in Lemma 11, we have the following bounds.

Using (16) for the random variable R_i, we obtain

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

Using (17) and using the same arguments, we obtain

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

Combining with (46), it follows that

$$p^e \leq \delta.$$

In the remainder of the proof we show that inequalities (47) and (48) hold.

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

$$\begin{aligned}
&\mathbf{P}[y_t = i] - \mathbf{P}[y_t = j] \\
&= \sum_{A\in\mathcal{A}}\mathbf{P}[S_t = A\cup\{i\}]D_{i,j}(A) \\
&\quad + \sum_{B\in\mathcal{B}}\mathbf{P}[S_t = B\cup\{i,j\}]D_{i,j}(B) \qquad (49)
\end{aligned}$$

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

$$p_k(\mathbf{x}) \geq p_k(\mathbf{0}) + \nabla p_k(\mathbf{0})^\top \mathbf{x} - \frac{1}{2}\beta \|\mathbf{x}\|_2^2 \quad (50)$$

$$p_k(\mathbf{x}) \leq p_k(\mathbf{0}) + \nabla p_k(\mathbf{0})^\top \mathbf{x} + \frac{1}{2}\beta \|\mathbf{x}\|_2^2 \quad (51)$$

where

$$\beta = \max_{\mathbf{x} \in [-2b, 2b]^{k-1}} \|\nabla^2 p_k(\mathbf{x})\|_2. \quad (52)$$

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

$$D_{i,j}(A) \geq 2(k-1)b \frac{\partial p_k(\mathbf{0})}{\partial x_1} - 4(k-1)b^2\beta. \quad (53)$$

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 (53), for every $\theta_A \in \{-b, b\}^{k-1}$,

$$D_{i,j}(A) \geq (k-1)b \frac{\partial p_k(\mathbf{0})}{\partial x_1} \geq \frac{kb}{2} \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (54)$$

By the same arguments, we can show that

$$D_{i,j}(B) = p_k(\boldsymbol{b} - \theta_B^{(-b)}) - p_k(-\boldsymbol{b} - \theta_B^{(b)}) \geq \frac{kb}{2} \frac{\partial p_k(\mathbf{0})}{\partial x_1} \quad (55)$$

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

Since the comparison sets are sampled uniformly at random without replacement, note that

$$\mathbf{P}[S_t = A \cup \{i\}] = \frac{\binom{n-1}{k-1}}{\binom{n}{k}}, \text{ for all } A \in \mathcal{A} \quad (56)$$

and

$$\mathbf{P}[S_t = B \cup \{i,j\}] = \frac{\binom{n-2}{k-2}}{\binom{n}{k}}, \text{ for all } B \in \mathcal{B}. \quad (57)$$

From (49), (54), (55), (56) and (57), we have

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

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

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

which establishes (47). By the same arguments one can establish (48).

H. Proof of Theorem 5

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 $N_1 = \{1, 2, \ldots, n/2\}$ and $N_2 = \{n/2+1, \ldots, n\}$. 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 $N'_1 = \{n, 2, \ldots, n/2\}$ and $N'_2 = \{n/2+1, \ldots, n-1, 1\}$.

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 $\mathbf{E}_\theta$ and $\mathbf{E}_{\theta'}$ the expectations under the two respective distributions.

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

$$L(\mathbf{S}, \mathbf{y}) = \sum_{t=1}^{m} \log \left(\frac{p_{y_t, S_t}(\theta') \rho_t(S_t)}{p_{y_t, S_t}(\theta) \rho_t(S_t)} \right), \quad (58)$$

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

$$\mathbf{P}_{\theta'}[L(\mathbf{S}, \mathbf{y}) \geq \log(n/\delta)] \geq \frac{1}{2}. \quad (59)$$

Step 2: We show that

$$\mathbf{E}_{\theta'}[L(\mathbf{S},\mathbf{y})] \leq 36\frac{m}{n}\left(k^2 b\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 \quad (60)$$

$$\sigma^2_{\theta'}[L(\mathbf{S},\mathbf{y})] \leq 144\frac{m}{n}\left(k^2 b\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 \quad (61)$$

where $\sigma^2_{\theta'}[L(\mathbf{S},\mathbf{y})]$ 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 $g \in \mathbf{R}$,

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

Using this for $g = \log(n/\delta) - \mathbf{E}_{\theta'}[L(\mathbf{S},\mathbf{y})]$, it follows that (59) implies the following condition:

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

Further combining with (60) and (61), we obtain

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

which is the condition asserted in the theorem.

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

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

and

$$B = \{\hat{N}_1 = N_1'\} \cap \{\hat{N}_2 = N_2'\}.$$

Let B^c denote the complement of the event B.

Note that

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

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

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

Now, note

$$\begin{aligned}\mathbf{P}_{\theta'}[L(\mathbf{S},\mathbf{y}) \leq g, B] =& \mathbf{E}_{\theta'}[\mathbf{1}(L(\mathbf{S},\mathbf{y}) \leq g, B)] \\ =& \mathbf{E}_{\theta}[e^{L(\mathbf{S},\mathbf{y})}\mathbf{1}(L(\mathbf{S},\mathbf{y}) \leq g, B)] \\ \leq& \mathbf{E}_{\theta}[e^{g}\mathbf{1}(L(\mathbf{S},\mathbf{y}) \leq g, B)] \\ =& e^g \mathbf{P}_\theta[L(\mathbf{S},\mathbf{y}) \leq g, B] \\ \leq& e^g \mathbf{P}_\theta[B] \\ \leq& e^g \frac{4}{n^2}\delta \quad (62)\end{aligned}$$

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

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

$$\mathbf{P}_{\theta'}[L(\mathbf{S},\mathbf{y}) \leq g, B^c] \leq \mathbf{P}_{\theta'}[B^c] \leq \delta. \quad (63)$$

For $g = \log(n/\delta)$, from (62) and (63), 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 $S_1, S_2, \ldots, S_m$ 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 (50), (51), and (52), we have for every S and $i \in S$,

$$\begin{aligned}&|p_{i,S}(\theta') - p_{i,S}(\theta)| \\ \leq& 2kb\frac{\partial p_k(\mathbf{0})}{\partial x_1} + 4\beta bk \\ \leq& 3kb\frac{\partial p_k(\mathbf{0})}{\partial x_1}, \quad (64)\end{aligned}$$

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

From (64), for every comparison set S such that $S \cap \{1,n\} \neq \emptyset$, we have

$$\begin{aligned}&\sum_{i\in S}\left(p_{i,S}(\theta') - p_{i,S}(\theta)\right)^2 \\ &\leq \sum_{i\in\{1,n\}\cap S}\left(p_{i,S}(\theta') - p_{i,S}(\theta)\right)^2 + \\ &\quad \left(\sum_{i\in S\setminus\{1,n\}} p_{i,S}(\theta') - p_{i,S}(\theta)\right)^2 \\ &\leq 2\left(3kb\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2, \quad (65)\end{aligned}$$

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

$$p_{1,S}(\theta') \le \frac{1}{k} \le p_{1,S}(\theta) \quad \text{and}$$
$$p_{i,S}(\theta') \ge p_{i,S}(\theta) \quad \forall i \ne 1;$$

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

$$p_{n,S}(\theta') \ge \frac{1}{k} \ge p_{n,S}(\theta) \quad \text{and}$$
$$p_{i,S}(\theta') \le p_{i,S}(\theta) \quad \forall i \ne n.$$

From (64) and the assumption of the theorem, we have

$$\begin{aligned}\min_S \min_{i\in S} p_{i,S}(\theta) &= \min_{S:n\in S} p_{n,S}(\theta) \\ &\ge \frac{1}{k} - 3kb\frac{\partial p_k(\mathbf{0})}{\partial x_1} \\ &\ge \frac{1}{2k}. \end{aligned} \tag{66}$$

For simplicity of notation, let

$$D = 3kb\frac{\partial p_k(\mathbf{0})}{\partial x_1}. \tag{67}$$

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

$$\sum_{i\in S} p_{i,S}(\theta') \log\left(\frac{p_{i,S}(\theta')}{p_{i,S}(\theta)}\right) \le 2kD^2 \tag{68}$$

which is obtained from

(i) $p_{i,S}(\theta) \ge 1/(2k)$ for all $i \in S$ that holds by (66),

(ii) $\sum_{i\in S}(p_{i,S}(\theta') - p_{i,S}(\theta))^2 = 2D^2$ from (65),

(iii) $a \log \frac{a}{b} \le \frac{(a-b)^2}{2b} + a - b$.

Similarly to (68), from (i) and (ii) and $a\left(\log\frac{a}{b}\right)^2 \le \frac{(a-b)^2}{a\wedge b}\left(1 + \frac{|a-b|}{3(a\wedge b)}\right)$, we have

$$\sum_{i\in S} p_{i,S}(\theta')\left(\log\left(\frac{p_{i,S}(\theta')}{p_{i,S}(\theta)}\right)\right)^2 \le 8kD^2. \tag{69}$$

Since

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

and according to the model, the input observations are independent, from (68) and (69), we have

$$\begin{aligned}&\mathbf{E}_{\theta'}[L(\mathbf{S}, \mathbf{y})] \\ =\ & m\mathbf{E}_{\theta'}\left[\log\left(\frac{p_{y_1,S_1}(\theta')}{p_{y_1,S_1}(\theta)}\right)\right] \\ =\ & m \sum_{S:S\cap\{1,n\}\ne\emptyset} \mathbf{P}_{\theta'}[S_1 = S] \\ & \sum_{y\in S} p_{y,S}(\theta')\left[\log\left(\frac{p_{y,S}(\theta')}{p_{y,S}(\theta)}\right)\right] \\ \le\ & 4\frac{m}{n}k^2D^2 \end{aligned} \tag{70}$$

and

$$\begin{aligned}&\sigma^2_{\theta'}[L(\mathbf{S}, \mathbf{y})] \\ =\ & m\sigma^2_{\theta'}\left[\log\left(\frac{p_{y_1,S_1}(\theta')}{p_{y_1,S_1}(\theta)}\right)\right] \\ \le\ & m\mathbf{E}_{\theta'}\left[\left(\log\left(\frac{p_{y_1,S_1}(\theta')}{p_{y_1,S_1}(\theta)}\right)\right)^2\right] \\ =\ & m \sum_{S:S\cap\{1,n\}\ne\emptyset} \mathbf{P}_{\theta'}[S_1 = S] \\ & \sum_{y\in S} p_{y,S}(\theta')\left[\left(\log\left(\frac{p_{y,S}(\theta')}{p_{y,S}(\theta)}\right)\right)^2\right] \\ \le\ & 16\frac{m}{n}k^2D^2. \end{aligned} \tag{71}$$

I. 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

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \frac{1}{k-1}\int_{\mathbf{R}} f(x)dF(x)^{k-1}. \tag{72}$$

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

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = A_{F,k} + B_{F,k} \tag{73}$$

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 (73) is shown to hold as follows. Note that

$$\begin{aligned} & \frac{d^2}{dx^2}F(x)^k \\ = \; & \frac{d}{dx}(kF(x)^{k-1}f(x)) \\ = \; & k(k-1)F(x)^{k-2}f(x)^2 + kF(x)^{k-1}f'(x). \end{aligned}$$

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

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

Combining with the fact

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

we obtain (73).

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

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_0^{\infty} f(x)^2(F(x)^{k-2} + (1-F(x))^{k-2})dx. \tag{74}$$

J. Proof of Proposition 6

The upper bound follows by noting that that $B_{F,k}$ in (73) 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 $x \in \mathbf{R}$,

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

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

K. 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 (7) 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

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

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

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

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 (74), 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

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

L. Some Remaining Proofs

L.1. Proof of Lemma 12

Consider a set $S \subseteq N$ such that $|S| = k$, for an arbitrary integer $2 \le k \le n$. Without loss of generality, consider $S = \{1, 2, \ldots, k\}$. Let $\mathbf{x}_l(\theta) = \theta_i - \theta_{S\setminus\{l\}}$, for $l \in S$. For simplicity, with a slight abuse of notation, we write x_l in lie of $x_i(\theta)$, for $l \in S$. We first consider the case $i \neq j$. By straightforward derivation, we have

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

We separately consider three different cases.

Case 1: i, j, l are all distinct. Note that

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j}(-\log(p_k(\mathbf{x}_l)))|_{\theta=\mathbf{0}} = I_1 \tag{75}$$

where

$$I_1 = -k \frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} + k^2 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2.$$

Case 2: $i \neq l$ and $j = l$. In this case, we characterize the following quantity for $\theta = \mathbf{0}$,

$$\begin{aligned}
&\frac{\partial^2}{\partial\theta_i \partial\theta_j}(-\log(p_k(\mathbf{x}_j))) \\
= &-\frac{1}{p_k(\mathbf{x}_j)} \frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i \partial\theta_j} + \frac{1}{p_k(\mathbf{x}_j)^2} \frac{\partial p_k(\mathbf{x}_j)}{\partial\theta_i} \frac{\partial p_k(\mathbf{x}_j)}{\partial\theta_j}
\end{aligned} \tag{76}$$

For every $u \in S$, $p_k(\mathbf{x}_u)$ does not change its value by changing the parameter θ to value $\theta + \Delta\theta$, for every constant $\Delta\theta \in \mathbf{R}$. Hence, by the full differential, we have

$$\frac{\partial p_k(\mathbf{x}_u)}{\partial\theta_j} = -\sum_{v \in S\setminus\{j\}} \frac{\partial p_k(\mathbf{x}_u)}{\partial\theta_v}. \tag{77}$$

Using (77), we have

$$\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i\partial\theta_j} = -\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i^2} - \sum_{v\in S\setminus\{i,j\}} \frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i\partial\theta_v}. \quad (78)$$

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, we can derive

$$\begin{aligned}
&\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i^2}|_{\theta=\mathbf{0}} \\
=& \int_{\mathbf{R}} f(z)f'(z) \prod_{v\in S\setminus\{i,j\}} F(z)^{k-2}dz \\
=& f(z)^2F(z)^{k-1}|_{-\infty}^{\infty} - \int_{\mathbf{R}} f(z)(f(z)F(z)^{k-2})'dz \\
=& -\int_{\mathbf{R}} f(z)f'(z)F(z)^{k-1} - (k-2)\int_{\mathbf{R}} f(z)^2F(z)^{k-3}dz \\
=& -\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i^2}|_{\theta=\mathbf{0}} - (k-2)\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2}
\end{aligned}$$

from which it follows that

$$\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i^2}|_{\theta=\mathbf{0}} = -\frac{k-2}{2}\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2}. \quad (79)$$

From (78) and (79), we obtain

$$\frac{\partial^2 p_k(\mathbf{x}_j)}{\partial\theta_i\partial\theta_j}|_{\theta=\mathbf{0}} = -\frac{k-2}{2}\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2}. \quad (80)$$

Using (77), we have

$$\frac{\partial p_k(\mathbf{x}_j)}{\partial\theta_j}|_{\theta=\mathbf{0}} = (k-1)\frac{\partial p_k(\mathbf{0})}{\partial x_1}. \quad (81)$$

Combining (76), (80) and (81), we have

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}(-\log(p_k(\mathbf{x}_j)))|_{\theta=\mathbf{0}} = I_2 \quad (82)$$

where

$$I_2 = \frac{k(k-2)}{2}\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2} - k^2(k-1)\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2.$$

Case 3: $i = l$ and $j \neq l$. By symmetry, from Case 2, we have

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}(-\log(p_k(\mathbf{x}_i)))|_{\theta=\mathbf{0}} = I_2. \quad (83)$$

Final step Putting the pieces together, from (75), (82), and (83), we have for $\theta = \mathbf{0}$,

$$\begin{aligned}
H_{i,j}(\theta, S) &= \sum_{l\in S\setminus\{i,j\}} p_k(\mathbf{x}_l)\frac{\partial^2}{\partial\theta_i\partial\theta_j}(-\log(p_k(\mathbf{x}_l))) \\
&\quad + p_k(\mathbf{x}_i)\frac{\partial^2}{\partial\theta_i\partial\theta_j}(-\log(p_k(\mathbf{x}_i))) \\
&\quad + p_k(\mathbf{x}_l)\frac{\partial^2}{\partial\theta_i\partial\theta_j}(-\log(p_k(\mathbf{x}_j))) \\
&= \frac{k-2}{k}I_1 + \frac{1}{k}I_2 + \frac{1}{k}I_2 \\
&= -k^2\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2. \quad (84)
\end{aligned}$$

Now, we consider the case $i = j$. Using same argument as in (77), we have

$$\frac{\partial^2(-\log(p_k(\mathbf{x}_l)))}{\partial\theta_i^2} = -\sum_{v\in S\setminus\{i\}} \frac{\partial^2(-\log(p_k(\mathbf{x}_l)))}{\partial\theta_i\partial\theta_v}.$$

Hence,

$$H_{i,i}(\theta, S) = -\sum_{v\in S\setminus\{i\}}\sum_{l\in S} p_k(\mathbf{x_l})\frac{\partial^2(-\log(p_k(\mathbf{x}_l)))}{\partial\theta_i\partial\theta_v}.$$

Combining with $H_{i,i}(\theta, S) = -\sum_{v\in S\setminus\{i\}} H_{i,v}(\theta, S)$ and the result established in (84), we have for $\theta = \mathbf{0}$,

$$H_{i,i}(\theta, S) = k^2(k-1)\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2.$$

L.2. Proof of Lemma 14

Without loss of generality, let $y = 1$ and $S = \{1, \ldots, k\}$. Then, we have

$$\begin{aligned}
\frac{1}{k^2}\frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial\theta_1\partial\theta_2} &= -(k-1)\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 \\
&\quad + \frac{k-2}{2k}\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2} \quad (85)
\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{k^2}\frac{\partial^2(-\log(p_{1,S}(\mathbf{0})))}{\partial\theta_i\partial\theta_j} &= -\frac{1}{k}\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2} \\
&\quad + \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2. \quad (86)
\end{aligned}$$

From assumption **A1** and (85),

$$\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1\partial x_2} \leq \frac{2k(k-1)}{k-2}\left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2. \quad (87)$$

Note that it holds that

$$\frac{\partial^2 p_k(\mathbf{0})}{\partial x_1 \partial x_2} \geq 0. \tag{88}$$

Combining (85), (86), (87), and (88), we have

$$\frac{\partial^2(-\log(p_{y,S}(\mathbf{0})))}{\partial\theta_i \partial\theta_j} \geq -k^3 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2 \quad \forall \quad i \neq j.$$

From the above inequality and (19),

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

Therefore, we conclude

$$\|\nabla^2(-\log(p_{y,S}(\mathbf{0})))\|_2 \leq k^4 \left(\frac{\partial p_k(\mathbf{0})}{\partial x_1}\right)^2,$$

which holds because $\|\Lambda_{\mathbf{M}_S}\|_2 = k$.

L.3. Proof of Proposition 11

We prove only (17) as the proof of (16) follows by similar arguments.

By 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

$$h(s) = qs - \log(1-p+pe^s)$$

Now, using the elementary fact $\log(1-x) \leq -x$, we have

$$h(s) \geq qs + p - pe^s.$$

Take $s = s^* := \log(q/p)$, then,

$$h(s^*) \geq q\log\left(\frac{q}{p}\right) + p - q.$$

Now, let $\epsilon = q - p$, and note that

$$q\log\left(\frac{q}{p}\right) + p - q \quad := \quad g(\epsilon)$$

where

$$g(\epsilon) = q\log\left(\frac{q}{q-\epsilon}\right) - \epsilon.$$

Since

$$g'(\epsilon) = \frac{q}{q-\epsilon} - 1 = \frac{\epsilon}{q-\epsilon} \geq \frac{1}{2q}\epsilon$$

we have

$$g(\epsilon) = \int_0^\epsilon g'(x)dx \geq \frac{1}{4q}\epsilon^2$$

Hence, it follows that

$$h(s^*) \geq \frac{1}{4q}(p-q)^2$$

and, thus,

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

M. Experimental Results

Table 2. Summary statistics for TopCoder and Taskcn datasets. The rightmost two columns contain, repectively, mean and median values of comparison sets' cardinalities.

Category	# contests	# workers	mean	median
TopCoder				
Design	209	62	1.99	2
Development	198	171	3.07	2
Specification	75	39	2.39	2
Architecture	238	55	1.75	2
Taskcn				
Website	131	636	9.87	6
Design	1,967	6,891	27.3	18
Coding	31	284	27.1	18
Writing	420	15,575	46.11	19

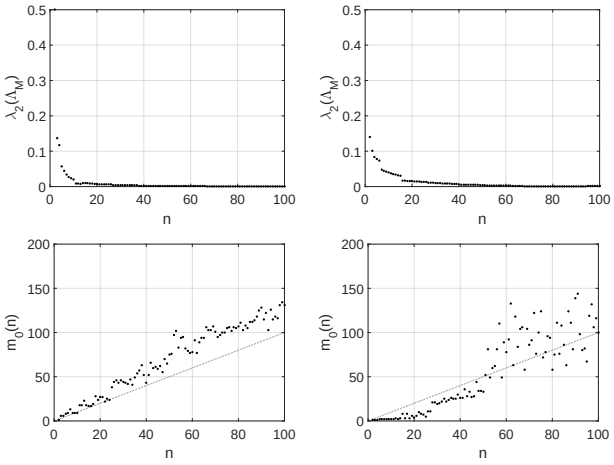

Figure 4. Same as in Figure 4 but for different categories (Development and Writing).