Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence

Nihar B. Shah **Sivaraman Balakrishnan** **Joseph Bradley**
Abhay Parekh **Kannan Ramchandran** **Martin Wainwright**

UC Berkeley

Abstract

Consider the problem of identifying the underlying qualities of a set of items based on measuring noisy comparisons between pairs of items. The Bradley-Terry-Luce (BTL) and Thurstone models are the most widely used parametric models for such pairwise comparison data. Working within a standard minimax framework, this paper provides sharp upper and lower bounds on the optimal error in estimating the underlying qualities under the BTL and the Thurstone models. These bounds are are topology-aware, meaning that they change qualitatively depending on the comparison graph induced by the subset of pairs being compared. Thus, in settings where the subset of pairs may be chosen, our results provide some principled guidelines for making this choice. Finally, we compare these error rates to those under cardinal measurement models and show that the error rates in the ordinal and cardinal settings have identical scalings apart from constant pre-factors. We use this result to investigate the relative merits of cardinal and ordinal measurement schemes.

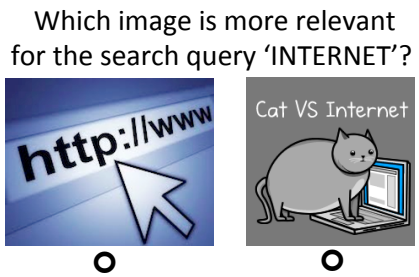
1 Introduction

In an increasing range of applications, it is of interest to elicit judgements from non-expert humans. Elicitation of preferences of consumers about products, either directly or indirectly, is a common practice [GCD81]. The data gathering process has been facilitated by the emergence of several new "crowdsourcing" platforms, such as Amazon Mechanical Turk, that have become powerful, low-cost tools for collecting human judgements [KDC+11, LRR11, vMM+08]. Crowdsourcing is employed not only for collection of preferences, but also for collecting data: for instance, rating responses of an online search engine to search queries [Kaz11], or counting the number of malaria parasites in an image of a blood smear [LOAF12]. Crowdsourcing has also become an indispensable tool for labeling data for training machine learning algorithms [HDY+12, RYZ+10, DDS+09]. Competitive sports implicitly elicit comparative qualities between individuals or teams [Ros07, HMG07]. Peer-grading in massive open online courses (MOOCs) [PHC+13] is an application gaining increasing popularity.

A common method of elicitation is through pairwise comparisons. For instance, the decision of a consumer to choose one product over another constitutes a pairwise comparison between the two products. Workers in a crowdsourcing setup are often asked to compare pairs of items: for instance, they might be asked to identify the better of two possible results of a search engine, as shown in Figure 1a. Competitive sports such as chess or basketball also involve sequences of pairwise comparisons of players or teams.

One use of pairwise comparisons is to estimate the inherent "qualities" or "weights" of the items being compared (e.g., skill levels of chess players, relevance of search engine results, etc.) The data obtained from pairwise comparisons can be modeled as a noisy sample of these latent (real-valued) weights. Noise can arise from a variety of sources. When objective questions are posed to human subjects, noise can arise from their differing levels of expertise. In a sports competition, many sources of randomness can influence the outcome of any particular match between a pair of competitors. Thus, one important goal is to estimate the latent qualities based on noisy data in the form of pairwise comparisons. A related problem is that of experimental design: assuming that we can choose the subset of pairs to be compared (e.g., in designing a chess tournament), what is the best such choice? Char-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.



(a) Asking for a pairwise comparisons.

(b) Asking for a numeric score.

Figure 1: An example of eliciting judgements from people: rating the relevance of the result of a search query.

acterizing the fundamental difficulty of estimating the weights will allow us to make this choice judiciously. These tasks are the primary focus of this paper.

In more detail, we consider the two most popular models for pairwise comparisons: the Thurstone (Case V) model [Thu27], and the Bradley-Terry-Luce (BTL) model [BT52, Luc59]. The Thurstone (Case V) model has been used in a variety of both applied [Swe73, Ros07, HMG07] and theoretical papers [B+05, Kra08, Nos85]. The BTL model has been similarly popular in both theory and practice [Nos85, AWL+98, KR82, HH10, LRS12, GCD81, KZ87]. Both models involve a latent real number as the weight of each item, and the outcome of each comparison is some noisy version of the pairwise comparison between the underlying scores of the two items.

With this context, the contributions of this paper are three-fold. First, we derive upper and lower bounds on the minimax estimation rates under the two models. Our upper and lower bounds on the squared ℓ_2 estimation error agree up to constant factors: to the best of our knowledge, despite the voluminous literature on these two models, this provides the first sharp characterization of the associated minimax rates. Moreover, our error guarantees provide guidance to the practitioner in assessing the minimax number of pairwise comparisons to be made in order to guarantee a prespecified error. Our second contribution is to derive bounds that are *topology-aware*, meaning that they depend on the comparison graph induced by the subset of pairs that are compared. Our theoretical analysis reveals that the spectral gap of the graph Laplacian plays a fundamental role, and provides guidelines for the practitioner on how to choose the subset of comparisons to be made. Third, we employ our sharp bounds to investigate when it is better to compare than to score. When eliciting data, one often has the liberty to ask for either cardinal values or for pairwise comparisons from the human subjects. These two options are illustrated in Figure 1. One would like to adopt the approach that would lead to a better estimate. One may be tempted to think that cardinal elicitation methods are superior, since each cardinal measurement gives a real-valued number whereas an ordinal measurement provides at most one bit of information. Our bounds show that, perhaps surprisingly, the scaling of the error in the cardinal and ordinal settings is identical up to constant pre-factors. As we demonstrate, this result allows for a comparison of cardinal and ordinal data elicitation methods in terms of the per-measurement noise alone, independent of the number of measurements and the number of items.

2 Problem formulation

We begin with some background followed by a precise formulation of the problem.

2.1 Generative models

Given a collection of d items to be evaluated, suppose that each item has a certain numeric *weight*, and a comparison of any pair of items is generated via a comparison of the two weights in the presence of noise. We represent the weights as a vector $w^* \in \mathbb{R}^d$, so item $j \in [d]$ has weight w_j^*. Now suppose that we make n pairwise comparisons: if comparison $i \in [n]$ pertains to items (a_i, b_i), then it can be described by a differencing vector $x_i \in \mathbb{R}^d$, with entry a_i equal to 1, entry b_i equal to -1, and the remaining entries set to 0.

In terms of this notation, the Thurstone (Case V) model [Thu27] is based on making n i.i.d. observations of the form

$$y_i = \text{sign}\Big\{\langle x_i, w^* \rangle + \epsilon_i\Big\}, \quad \text{for } i \in [n], \tag{Thurstone}$$

where $\epsilon_i \sim N(0, \sigma^2)$ is i.i.d. observation noise. On the other hand, the Bradley-Terry-Luce (BTL) model [BT52, Luc59] involves obtaining samples $y_i \in \{-1, 1\}$ drawn independently from the distribu-

tion

$$\mathbb{P}\big[y_i = 1; x_i, w^*\big] = \frac{1}{1 + \exp\left(\frac{-\langle x_i,\, w^*\rangle}{\sigma}\right)} \quad \text{for } i \in [n]. \tag{BTL}$$

In both models, the parameter σ plays the role of a noise parameter, with a higher value of σ leading to more uncertainty in the comparisons. In each case, the value of σ is assumed to be known. Note that both THURSTONE and BTL models are invariant to shifts in w^*, that is, they do not differentiate between the vector w^* and the shifted vector w^*+1, where 1 is the all-ones vector. Therefore, we assume that $\langle 1,\, w^*\rangle = 0$ in order to enforce identifiability of the vector of weights.

While our primary focus is analysis of the pairwise-comparison setting, for comparison purposes we also analyze analogous *cardinal* settings where each observation is real valued. In the CARDINAL model we consider, each observation consists of a numeric evaluation of a single item,

$$y_i = \langle u_i,\, w^*\rangle + \epsilon_i \qquad \text{for } i \in [n], \qquad \text{(CARDINAL)}$$

where u_i in this case is a coordinate vector with one of its entries equal to 1 and remaining entries equal to 0, and ϵ_i is independent Gaussian noise $N(0, \sigma^2)$. One may alternatively elicit cardinal values of the differences between pairs of items

$$y_i = \langle x_i,\, w^*\rangle + \epsilon_i \qquad \text{for } i \in [n], \quad \text{(PAIRED LINEAR)}$$

where ϵ_i are i.i.d. $N(0, \sigma^2)$. We term this model the PAIRED LINEAR model.

2.2 Fixed design and the graph Laplacian

Let us begin by analyzing the estimation error when a fixed subset of pairs is chosen for comparison. Of interest to us will be the *comparison graph* defined by these chosen pairs, with each pair inducing an edge in the graph. Edge weights are determined by the fraction of times a given pair is compared. The analysis in the sequel reveals the central role played by the Laplacian of this weighted graph. Note that we are operating in a fixed-design setup where the graph is constructed offline and does not depend on the observations.

In the ordinal models, the i^{th} measurement is related to the difference between the two items being compared, as defined by the measurement vector $x_i \in \mathbb{R}^d$. We let $X \in \mathbb{R}^{n \times d}$ denote the measurement matrix with the vector x_i^T as its i^{th} row. The Laplacian matrix L associated with this differencing matrix is given by

$$L := \frac{1}{n} X^T X \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T. \tag{1}$$

By construction, for any vector $v \in \mathbb{R}^d$, we have $v^T L v = \sum_{j \neq k} L_{jk}(v_j - v_k)^2$, where L_{jk} is the fraction of the measurement vectors $\{x_i\}_{i=1}^n$ in which items (j, k) are compared.

The Laplacian matrix is positive semidefinite, and has at least one zero-eigenvalue, corresponding to the all-ones eigenvector. The Laplacian matrix induces a graph $G(L)$ on the vertex set $V = \{1, \ldots, d\}$, in which a given pair (j, k) is included as an edge if and only if $L_{jk} \neq 0$, and the weight on an edge (j, k) equals L_{jk}. Throughout our analysis, we assume that this graph is connected, since otherwise, the quality score vector w is not identifiable. Note that the Laplacian matrix L induces a seminorm[1] on $\mathbb{R}^d$, given by

$$\|u - v\|_L := \sqrt{(u - v)^T L (u - v)}. \tag{2}$$

A major focus is on the *minimax risk* in terms of the Laplacian seminorm.

2.3 Minimax framework

Finally, we review the standard notion of minimax risk used in this paper. For a given family of generative models, each weight vector w induces an associated distribution $\mathbb{P}_w$. We let $w(\mathcal{P})$ denote the set of allowed vectors w, and $\mathcal{P}$ denote the family of induced distributions. For a given weight vector w and collection of comparison vectors $\{x_i\}_{i=1}^n$, suppose that we observe n i.i.d. samples $\{y_i\}_{i=1}^n$ generated according to $\mathbb{P}_w$. Our goal is to estimate the unknown weight vector, and an estimator $\widehat{w}$ is any measurable mapping from the observations $\{y_i\}_{i=1}^n$ to the space $w(\mathcal{P})$.

For a given seminorm ρ, we consider the minimax risk given by

$$\mathfrak{M}_n(w(\mathcal{P}); \rho^2) := \inf_{\widehat{w}} \sup_{w^* \in w(\mathcal{P})} \mathbb{E}[\rho(\widehat{w}, w^*)^2], \tag{3}$$

where the expectation is taken over the samples $\{y_i\}_{i=1}^n$. The minimax risk characterizes the performance of the best estimator, as measured in the seminorm ρ squared, in a worst-case sense over the family $w(\mathcal{P})$.

In this paper, we analyze the minimax risk for two choices of seminorm ρ, namely the Laplacian seminorm $\|\widehat{w} - w^*\|_L$ from (2), and the Euclidean norm $\|\widehat{w} - w^*\|_2$. We denote these risks by $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_L^2\big)$ and $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2^2\big)$, respectively.

3 Sharp bounds on the minimax risk

This section presents the main results of the paper: sharp minimax bounds on the estimation error under

[1] A seminorm differs from a norm in that the seminorm of a non-zero element is allowed to be zero.

the pairwise comparison models introduced earlier in Section 2.1. Theorem 1 below bounds this minimax risk in each of the three models. In all of the statements, we use $c_{1\ell}, c_{2\ell}, c_{1u}, c_{2u}, c_1, c_2$ to denote positive numerical constants, independent of the sample size n, number of items d and other problem-dependent parameters. For a subset of the results, we assume that each coordinate of the weight vector w^* is bounded as

$$\|w^*\|_\infty \le B \tag{4}$$

for some constant B. We use $L^\dagger$ to denote the Moore-Penrose pseudoinverse of L.

Theorem 1 (Bounds on minimax rates)**.** *(a) For the paired linear model, the minimax rate is bounded as*

$$c_{1\ell}\,\sigma^2\frac{d}{n} \le \mathfrak{M}_n\big(\textsc{Paired Linear}; \|\cdot\|_L^2\big) \le c_{1u}\,\sigma^2\frac{d}{n}.$$

(b) For the Thurstone model with B-bounded weight vector (4)*, and sample size $n \ge \frac{c_2\sigma^2\kappa\mathrm{tr}(L^\dagger)}{B^2}$, the minimax rate is bounded as*

$$c_{2\ell}\sigma^2\frac{\kappa d}{n} \le \mathfrak{M}_n\big(\textsc{Thurstone}; \|\cdot\|_L^2\big) \le \frac{c_{2u}}{\kappa^2}\sigma^2\frac{d}{n},$$

where $\kappa := \Phi(2B/\sigma)(1-\Phi(2B/\sigma))$.

(c) For the BTL model with B-bounded weight vector (4) *and sample size $n \ge \frac{c_3\sigma^2\mathrm{tr}(L^\dagger)}{B^2}$, the minimax rate is bounded as*

$$c_{3\ell}\sigma^2\frac{d}{n} \le \mathfrak{M}_n\big(\mathrm{BTL}; \|\cdot\|_L^2\big) \le c_{3u}e^{\frac{4B}{\sigma}}\sigma^2\,\frac{d}{n}.$$

We defer detailed proofs of this and subsequent results to the Appendix. The upper bounds follow from an analysis of the maximum likelihood estimator. Interestingly, maximum likelihood estimation in each of these cases turns out to be a convex optimization problem (see, for instance, [TG11] for a proof in the Thurstone case). On the other hand, the lower bounds are based on a combination of information-theoretic techniques and carefully constructed packings of the parameter set $w(\mathcal{P})$. The main technical difficulty is in constructing a packing in the seminorm induced by the Laplacian L.

We note that the minimax bounds in the Thurstone and the BTL models depend on $\|w^*\|_\infty$. The bounds must necessarily be governed by $\|w^*\|_\infty$ since it can be shown that the minimax error under an unbounded $\|w^*\|_\infty$ will be infinite. Informally, this is related to the difficulty of estimating very small (or very large) probabilities that can arise in the two models for large $\|w^*\|_\infty$.

Negahban et al. [NOS14] also provided minimax bounds for the BTL model in the special case of differencing vectors $\{x_i\}_{i=1}^n$ chosen uniformly at random. They focused on this case to complement their analysis of a random walk-based algorithm. In their analysis, there is a gap between the achievable rate of the MLE, and the lower bound. In contrast, our analysis eliminates this discrepancy and shows that MLE is an optimal estimator (up to constant factors) in the terms of the minimax rate $\mathfrak{M}_n(\quad; \|\cdot\|_L^2)$. In independent and concurrent work Hajek et al. [HOX14] consider the problem of estimation in the Plackett-Luce model, which extends the BTL model to comparisons of two or more items. They derive bounds on the minimax error rates under this model which are tight up to logarithmic factors. In contrast, our results are tight up to constants and, as we emphasize in the following section, provide deeper insights into the role of the topology of the comparison graph.

4 Role of graph topology

In certain applications, one may have the liberty to decide which pairs are compared. The results of Section 3 demonstrated the role played by the Laplacian of the comparison graph in the estimation error. We now employ these results to derive guidelines towards designing the comparison graph, i.e., towards answering the question: "If one can make d measurements, then which pairs should be compared?".

Let us focus on upper bounds in the ordinal setting, and consider estimation error in the squared ℓ_2 norm. As in Theorem 1, we assume that the graph induced by the comparisons is connected. Apart from model-specific constants, the minimax risks also share the same scaling—namely

$$\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2^2\big) \asymp \sigma^2\frac{d}{n\lambda_2(L)}, \tag{6}$$

where $\lambda_2(L)$ is the second smallest eigenvalue of the Laplacian matrix L. In order to derive this expression, we used the fact that $\langle w, 1\rangle = 0$.

As a graph Laplacian, the second smallest eigenvalue is determined by the topology of the chosen comparisons. In order to illustrate, let us consider five canonical examples: the barbell graph, the complete graph, a bounded degree expander, the path graph and the lattice graph. In each case, we assume that the samples are distributed evenly along the edges of a fixed graph, and that the sample size n is sufficiently large. Using standard matrix concentration inequalities, it is straightforward to extend our analysis to the setting of random chosen comparisons from a fixed graph (see for instance [Oli09]). The properties of the Laplacian

matrices of these graphs can be found in [BH11] and other texts on the subject.

(a) *Barbell graph:* For an even number d of vertices, the barbell graph consists of two cliques of $d/2$ disjoint sets of vertices with a single edge between them. Suppose $n \geq \binom{d/2}{2} + 1$. In this case we obtain that $\lambda_2(L) = \Theta(\frac{1}{d^3})$ and the squared ℓ_2 error scales as $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2^2\big) \asymp \frac{d^4}{n}$.

(b) *Complete graph:* In the regime $n \geq \binom{d}{2}$, we have $\lambda_2(L) = \frac{d}{\binom{d}{2}}$, so that the squared ℓ_2 error scales as $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2^2\big) \asymp \frac{d^2}{n}$.

(c) *Degree-k expander:* A similar argument as in the previous case shows that if $n \geq kd$, then the error scales as $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2^2\big) \asymp \frac{d^2}{n}$.

(d) *Path graph:* For the path graph, we have $\lambda_2(L) = \Theta(1/d^3)$ and hence $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2^2\big) \asymp \frac{d^4}{n}$.

(e) *2D lattice:* In this case we obtain $\lambda_2(L) = \Theta(\frac{1}{d^2})$, and $\mathfrak{M}_n\big(w(\mathcal{P}); \|\cdot\|_2\big) \asymp \frac{d^3}{n}$.

To summarize, we see the squared ℓ_2 error scaling as $\frac{d^2}{n}$ for the complete graph and the degree-k expander. We conjecture that this is in fact the *best possible* scaling. Observe that the degree-k expander requires a sample size lower bounded as $n \geq kd$ while the complete graph requires $n \geq \binom{d}{2}$, so in practice, we should prefer a low-degree expander (at least for low sample sizes). On the other hand, for other graphs—including the path, lattice and barbell graphs—the error scaling is considerably worse, showing that these are are poor choices for the topology of comparisons.

5 Cardinal versus ordinal measurements

In this section, we compare two approaches towards eliciting data: a score-based "cardinal" approach and a comparison-based "ordinal" approach. In a cardinal approach, evaluators directly enter numeric scores as their answers (Figure 1b), while an ordinal approach involves comparing (pairs of) items (Figure 1a).

There are obvious advantages and disadvantages associated with either approach. On one hand, the cardinal approach allows for very fine measurements. For instance, the cardinal measurements in Figure 1 can take any value between 0 and 100, whereas an ordinal measurement is binary. One might be tempted to go even further and argue that ordinal measurements necessarily give less information, for one can always convert a set of cardinal measurements into ordinal, simply by ordering the measurements by value. If this conversion were valid, the data processing inequality [CT12], would then guarantee that estimators based on ordinal data can never outperform estimators based on cardinal data. However, this conversion assumes that cardinal and ordinal measurements are suffer from the same type of statistical fluctuation. In contrast, ordinal measurements avoid calibration issues that are frequently encountered in cardinal measurements [TG11], such as the evaluators' inherent (and possibly time-varying) biases, or tendencies to give inflated or conservative evaluations. Ordinal measurements are also recognized to be easier or faster for humans to make [Bar03, SBC05], allowing for more evaluations with the same amount of time, effort and cost.

The lack of clarity regarding when to use a cardinal versus an ordinal approach forms the motivation of this section. Can we make as reliable estimates from paired comparisons as from numeric scores? How much lower does the noise have to be for comparative measurements to be preferred over cardinal measurements? The answers to these questions will help in determining how responses should be elicited.

In order to compare the cardinal and ordinal methods of data elicitation, we focus on a setting with evenly budgeted measurements. In accordance with the fixed-design setup assumed throughout the paper, we choose the vectors x_i a priori. We consider the Gaussian-noise models Thurstone and Cardinal. In order to capture the fact that the amount of noise is different in the cardinal and ordinal settings, we will denote the standard deviation of the noise in the cardinal setting as σ_c, and retain our notation of σ for the noise in the ordinal setting. In order to bring the two models on the same footing, we measure the error in terms of the squared ℓ_2-norm.

Let Φ denote the standard Gaussian c.d.f., and define

$$b_\ell(\sigma, B) := c_{2\ell}\Phi(2B/\sigma)(1 - \Phi(2B/\sigma)),$$
$$b_u(\sigma, B) := \frac{c_{2u}}{\big(\Phi(2B/\sigma)(1 - \Phi(2B/\sigma))\big)^2},$$
$$b(\sigma, B) := \left\lceil \frac{c_2\Phi(2B/\sigma)(1 - \Phi(2B/\sigma))\sigma^2}{B^2} \right\rceil.$$

Observe that b_ℓ, b_u and b are independent of the parameters n and d.

With these preliminaries in place, we now compare the minimax error in the estimation under the cardinal and ordinal settings.

Theorem 2. *Let $\|w^*\|_\infty \leq B$ for some known value B, and suppose n is a multiple of $d(d-1)b(\sigma, B)$, and that in the* Cardinal *model we observe each coordinate n/d times for a known noise parameter σ_c. Then*

the minimax risk is given by

$$\mathfrak{M}_n\big(\textsc{Cardinal} ; \|\cdot\|_2^2\big) = \sigma_c^2 \frac{d^2}{n}.$$

Suppose that in the Thurstone *model we observe each pair* $n/\binom{d}{2}$ *times with known noise parameter* σ*. Then the minimax risk is sandwiched as*

$$\sigma^2 b_\ell(\sigma, B)\frac{d^2}{n} \le \mathfrak{M}_n\big(\textsc{Thurstone} ; \|\cdot\|_2^2\big) \le \sigma^2 b_u(\sigma, B)\frac{d^2}{n}.$$

In the cardinal case, when each coordinate is measured the same number of times, the Cardinal model reduces to the well-studied normal location model, for which the MLE is known to be the minimax estimator and its risk is straightforward to characterize (see Lehmann and Casella [LC98], for instance). In the ordinal case, the result follows from the general treatment in Section 3.

Let us now return to the question deciding between the cardinal and the ordinal methods of data elicitation. Suppose that we believe the Gaussian-noise models to be reasonably correct, and the per-observation errors σ and σ_c under the two settings are known or can be separately measured. Theorem 2 shows that the scaling of the minimax error in the cardinal and the ordinal settings is identical in terms of the problem parameters n and d. Our result thus allows for the choice to be made based only on the parameters (σ, σ_c, B) and not on n and d: the ordinal approach incurs a lower minimax error when $b_u(\sigma, B)\sigma^2 < \sigma_c^2$ while the cardinal approach is better off in terms of minimax error whenever $b_\ell(\sigma, B)\sigma^2 > \sigma_c^2$. Tightening the (σ, B)-dependent constants in the bounds would lead to a sharp decision boundary between the cardinal and the ordinal approaches.

6 Experiments and simulations

In this section we describe experiments on the crowdsourcing platform Amazon Mechanical Turk (MTurk), MTurk.com, and simulations using synthetically generated data. We summarize the experiments and enumerate the results in this section, and refer the reader to Appendix C for more details. Throughout this section, estimation procedures are executed via maximum likelihood under the Thurstone model. In simulations with synthetic data, the true vector w^* is generated by first drawing a d-length vector from $N(0, I)$ and then shifting it to ensure that $\langle w^*, 1\rangle = 0$. In the synthetic case, the ML estimator is supplied with the correct value of σ, and in the data obtained from experiments from MTurk, the estimator is supplied the best-fitting value of σ obtained via 3-fold cross-validation.

6.1 Dependence on topology

We investigate the dependence of the squared ℓ_2 estimation error on the topology of the comparison graph. We consider the following five topologies: path, barbell, complete, expander and a 2D-lattice. For the expander graph, we use the Margulis-Gabber-Galil construction [Mar73, GG81] to form an 8-regular expander graph. For any chosen graph topology, the n difference vectors are selected as one edge each drawn uniformly at random (without replacement) from the comparison graph. Recall that our theory from Section 4 predicts the complete and expander graphs to perform the best, and the path and barbell graphs to fare the worst. Also recall that our theory predicts the error $\frac{\|w^* - \widehat{w}\|_2^2}{d}$ to scale as $1/n$ in the complete and expander topologies.

6.1.1 Synthetic simulations

We first performed simulations using data generated synthetically from the Thurstone model. Figure 2 plots the estimation error under various topologies of the comparison graph. Observe in the figure that the error is the lowest under the complete graph, and the highest under the barbell and the path graphs. This observation is consistent with our theoretical predictions.

6.1.2 Experiments on Mechanical Turk

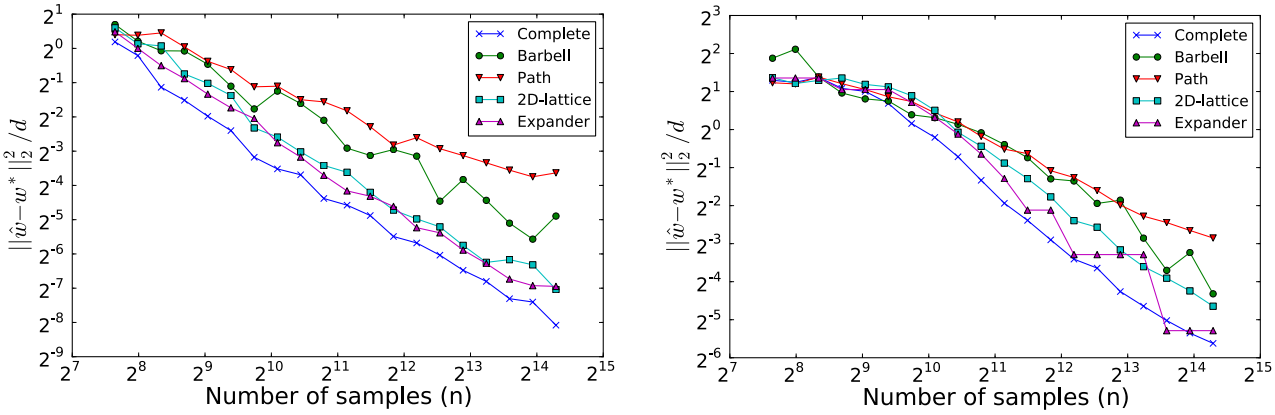
We conducted three experiments that required the workers to make ordinal choices. The experiments involved (i) identifying the bigger of a pair of circles, (ii) identifying the older of two people from their photographs, (iii) identifying the pair of cities which are farther apart. For each experiment, we recruited 140 workers on MTurk, and assigned them to one of the five topologies uniformly at random. Figure 3 plots the squared ℓ_2 estimation error for the three experiments under the five topologies considered. We see that the relative errors are generally consistent with our theory, with the complete graph exhibiting the best performance and the path graph faring the worst.

6.2 Cardinal vs. ordinal

We now consider the problem of choosing between the cardinal and the ordinal means of data elicitation.

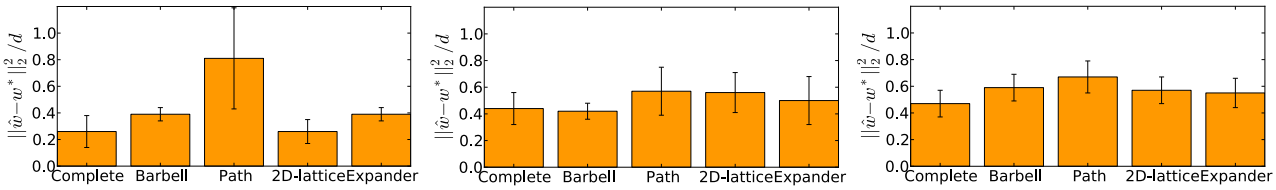
6.2.1 Measuring Per-observation Error

We conducted seven different experiments on MTurk to investigate the possibility of a data-processing inequality between the elicited cardinal and ordinal responses: Are responses elicited in ordinal form equivalent to data obtained by first eliciting cardinal responses and then subtracting pairs of items? Our experiments lead



(a) $d = 64$ (b) $d = 256$

Figure 2: Estimation error under different topologies in the simulations using synthetic data.



(a) Area of circle (b) Age from photograph (c) City distances

Figure 3: Estimation error under different topologies in the experiments conducted on MTurk.

us to conclude that this is generally not the case: converting cardinally collected data into ordinal (by subtracting pairs of responses) generally led to a higher amount of noise as compared to that in data that is elicited directly in ordinal form.

For each of the seven experiments, we recruited 100 workers, and assigned each worker randomly to either the ordinal or the cardinal version of the task. For the experiments in which we had access to "ground truth" solutions, we directly computed the fraction of responses that were incorrect in the ordinal and the cardinal-converted-to-ordinal data. For the remaining experiments, we computed the "error" as the fraction of responses that disagreed with each other. Note that we did *not* run any estimation procedure on the data: we only measured the noise in the raw responses.

The results are tabulated in Figure 4a. If the cardinal measurements could always be converted to ordinal measurements with the same noise level as directly eliciting ordinal responses, then it would be unlikely for the amount of error in the ordinal setting to be smaller than that in the cardinal setting. Figure 4a shows that converting cardinal data to an ordinal form often results in a higher (and sometimes significantly higher) per-sample error in the (raw) responses than direct elicitation of ordinal evaluations. This absence of data-processing inequality may be explained by the argument that the inherent evaluation process in the humans is not the same in the cardinal and ordinal cases: humans do *not* perform an ordinal evaluation by first performing cardinal evaluations and then comparing them [Bar03,SBC05]. One can thus assume that in many applications, we will have $\sigma < \sigma_c$.

6.2.2 Estimation error

For sake of completeness, we also computed the estimation error in the cardinal and ordinal settings. We consider data from the three experiments for which we have access to the ground truth. We normalize the true vector to have $\|w^*\|_\infty = 1$ and set $B = 1$. For each of the three experiments, we execute 100 iterations of the following procedure. Select five workers from the cardinal and five from the ordinal pool of workers uniformly at random. (The number five is chosen based on practical systems [WIP11, PHC+13].) We run the maximum-likelihood estimator of the CARDINAL model on the data from the five workers selected from the cardinal pool, and the maximum-likelihood estimator of the THURSTONE model on the data from the five workers of the ordinal pool. Note that unlike Section 6.2.1, the cardinal data here is *not* converted to ordinal.

The results are plotted in Figure 4b. To put the results in perspective of the rest of the paper, let us also recall the per-sample errors in these experiments from Figure 4a. Observe that in the experiment of estimating distances, the per-sample error in the cardinal data was significantly higher than the ordinal data. This is reflected in the results of Figure 4b where the es-

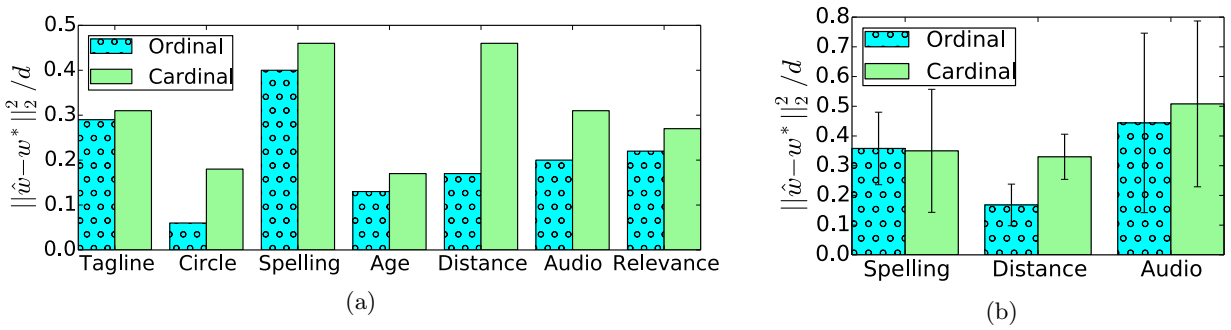


Figure 4. Results from experiments run on MTurk comparing the ordinal and cardinal methods of eliciting responses: (a) Fraction of incorrect responses. (b) Estimation error.

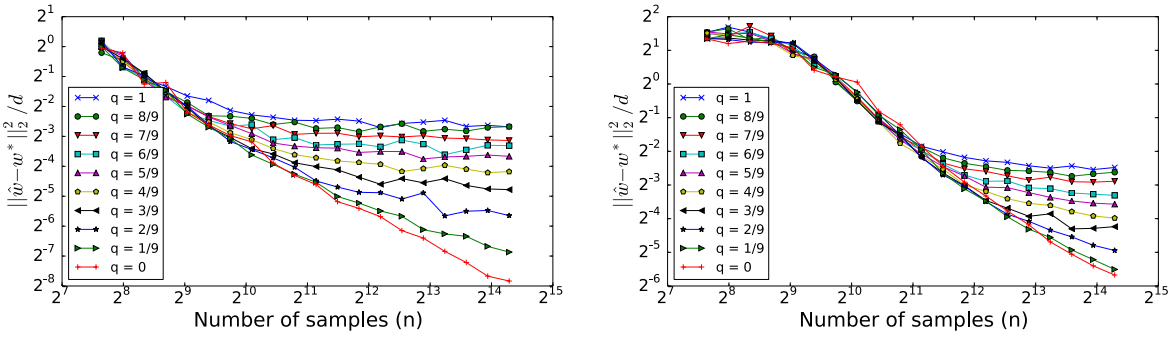


Figure 5: Estimation error under a misspecified model (simulations from synthetic data).

timator on the ordinal data outperforms (in terms of the squared ℓ_2 error) than the estimator on the cardinal data. On the other hand, the task of identifying the number of spelling mistakes involved a per-sample noise that was comparable across the two settings, and hence the estimator on the cardinal data scores over the ordinal one. Our theory needs to tighten the constants in order to address this regime.

6.3 Model misspecification

We investigated the effects of model mismatches via synthetic simulations. In the data generation process, every data point was generated from the BTL model with a probability $\epsilon \in [0, 1]$ and from the THURSTONE model with a probability $(1 - \epsilon)$. We set $\sigma = 1$ under both models. Inference was performed assuming the entire data was generated from THURSTONE, but using the correct values of σ and B. Figure 5 plots the error observed as ϵ was varied from 0 to 1. Observe that when $\epsilon = 0$, the estimation error drops linearly with a slope of -1 (on the log-log scale) as predicted by our theory. On the other hand, when ϵ is reasonably high, the error reduces much slower as n increases. An analytical investigation of model misspecification under the THURSTONE and BTL models is a topic for future work.

7 Conclusions

We derive topology-aware minimax error bounds under two widely studied preference-elicitation models, and demonstrated their usefulness in guiding the selection of comparisons and in guiding the choice of the elicitation paradigm (cardinal versus ordinal) when these options are available. One potential direction for future work would be to investigate improved data collection mechanisms, for instance adaptive schemes where we focus our effort on the hardest comparisons. A second direction would be to characterize the precise thresholds for making the choice between the cardinal and ordinal approaches. Finally, the Thurstone and BTL models are parametric idealizations that have proved useful in a wide variety of applications. In future work, it would be interesting to investigate more flexible non-parametric pairwise comparison models (see for instance, the paper [Cha12]).

Acknowledgements

This work was partially supported by the AFOSR grant FA9550-14-1-0016, and NSF grants CIF-31712-23800 and DMS-1107000 to MJW. In addition, the work of NS was partially supported by a Microsoft Research fellowship.

References

[AWL+98] Donald R Atkinson, Bruce E Wampold, Susana M Lowe, Linda Matthews, and Hyun-Nie Ahn. Asian American preferences for counselor characteristics: Application of the Bradley-Terry-Luce model to paired comparison data. *The Counseling Psychologist*, 26(1):101–123, 1998.

[B+05] Tom Bramley et al. A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, 6(2):202–223, 2005.

[Bar03] William Barnett. The modern theory of consumer behavior: Ordinal or cardinal? *The Quarterly Journal of Austrian Economics*, 6(1):41–65, 2003.

[BH11] Andries E Brouwer and Willem H Haemers. *Spectra of graphs*. Springer, 2011.

[BT52] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pages 324–345, 1952.

[Cha12] Sourav Chatterjee. Matrix estimation by universal singular value thresholding, 2012.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, pages 248–255. IEEE, 2009.

[GCD81] Paul E Green, J Douglas Carroll, and Wayne S DeSarbo. Estimating choice probabilities in multiattribute decision making. *Journal of Consumer Research*, pages 76–84, 1981.

[GG81] Ofer Gabber and Zvi Galil. Explicit constructions of linear-sized superconcentrators. *Journal of Computer and System Sciences*, 22(3):407–420, 1981.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[HH10] Sandra Heldsinger and Stephen Humphry. Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2):1–19, 2010.

[HMG07] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 19:569, 2007.

[HOX14] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. *arXiv preprint arXiv:1406.5638*, 2014.

[Kaz11] Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in information retrieval*, pages 165–176. Springer, 2011.

[KDC+11] Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, Mariusz Jaskolski, and David Baker. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[KR82] Kenneth J Koehler and Harold Ridpath. An application of a biased version of the Bradley-Terry-Luce model to professional basketball results. *Journal of Mathematical Psychology*, 25(3), 1982.

[Kra08] Paul FM Krabbe. Thurstone scaling as a measurement method to quantify subjective health outcomes. *Medical care*, 46(4):357–365, 2008.

[KZ87] Zahid Y Khairullah and Stanley Zionts. An approach for preference ranking of alternatives. *European journal of operational research*, 28(3):329–342, 1987.

[LC98] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics, 1998.

[LOAF12] Miguel Angel Luengo-Oroz, Asier Arranz, and John Frean. Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of medical Internet research*, 14(6), 2012.

[LRR11] ASID Lang and Joshua Rio-Ross. Using Amazon Mechanical Turk to transcribe historical handwritten documents. *The Code4Lib Journal*, 2011.

[LRS12] Peter John Loewen, Daniel Rubenson, and Arthur Spirling. Testing the power of arguments in referendums: A Bradley–Terry approach. *Electoral Studies*, 31(1):212–221, 2012.

[Luc59] R Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley, 1959.

[Mar73] Grigorii Aleksandrovich Margulis. Explicit constructions of concentrators. *Problemy Peredachi Informatsii*, 9(4):71–80, 1973.

[Nos85] Robert M Nosofsky. Luce's choice model and Thurstone's categorical judgment model compared: Kornbrot's data revisited. *Attention, Perception, & Psychophysics*, 37(1):89–91, 1985.

[NOS14] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pair-wise comparisons. *arXiv preprint arXiv:1209.1688*, 2014.

[Oli09] Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[PHC+13] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. In *International Conference on Educational Data Mining*, 2013.

[Ros07] Daniel Ross. Arpad Elo and the Elo rating system, 2007. http://en.chessbase.com/post/arpad-elo-and-the-elo-rating-system.

[RYZ+10] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010.

[SBC05] Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

[Swe73] John Swets. The relative operating characteristic in psychology. *Science*, 182(4116), 1973.

[TG11] Kristi Tsukida and Maya R Gupta. How to analyze paired comparison data. Technical report, DTIC Document, 2011.

[Thu27] Louis L Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

[vMM+08] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. Recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[WIP11] Jing Wang, Panagiotis G Ipeirotis, and Foster Provost. Managing crowdsourcing workers. In *The 2011 Winter Conference on Business Intelligence*, pages 10–12, 2011.