Convergence Rates of Gradient Descent and MM Algorithms for Bradley-Terry Models

Milan Vojnovic
LSE, UK

Se-Young Yun
KAIST, Korea

Kaifang Zhou
LSE, UK

Abstract

We present tight convergence rate bounds for gradient descent and MM algorithms for maximum likelihood (ML) estimation and maximum a posteriori probability (MAP) estimation of a popular Bayesian inference method, for Bradley-Terry models of ranking data. Our results show that MM algorithms have the same convergence rate, up to a constant factor, as gradient descent algorithms with optimal constant step size. For the ML estimation objective, the convergence is linear with the rate crucially determined by the algebraic connectivity of the matrix of item pair co-occurrences in observed comparison data. For the MAP estimation objective, we show that the convergence rate is also linear, with the rate determined by a parameter of the prior distribution in a way that can make convergence arbitrarily slow for small values of this parameter. The limit of small values of this parameter corresponds to a flat, non-informative prior distribution.

1 Introduction

Statistical models of ranking data play an important role in a wide range of applications, including learning to rank in information retrieval (Burges et al. (2006); Li (2011)), skill rating in games such as Chess and other sports (Elo (1978)), online platforms such as online gaming platforms (Herbrich et al. (2006)), and evaluation of machine learning algorithms by comparing them with each other (Balduzzi et al. (2018)).

A common class of statistical models of ranking data are *generalized Bradley-Terry models*, which accommodate paired comparisons with win-lose outcomes (Zermelo (1929); Bradley and Terry (1952); Bradley (1954)), paired comparisons with win-lose-draw outcomes (e.g. Rao and Kupper (1967)), choices from comparison sets of two or more items (e.g. Luce choice model by Luce (1959)), full ranking outcomes for comparison sets of two or more items (e.g. Plackett-Luce ranking model by Plackett (1975)), as well as group comparisons (e.g. Huang et al. (2006, 2008)). These models can be derived from suitably defined latent variable models, where items are associated with independent latent performance random variables, which is in the spirit of the well-known Thurstone model of comparative judgment by Thurstone (1927).

According to the Bradley-Terry model of paired comparisons with win-lose outcomes, each comparison of items i and j has an independent outcome: either i wins against j ($i \succ j$) or j wins against i ($j \succ i$). The distribution of the comparison outcome is given by

$$\Pr[i \succ j] = \frac{e^{w_i}}{e^{w_i} + e^{w_j}}$$

where $\mathbf{w} = (w_1, w_2, \dots, w_n)^\top \in \mathbb{R}^n$ is a parameter vector. The Bradley-Terry model of paired comparisons was studied by many, e.g. Ford (1957); Dykstra, Jr. (1956, 1960); Simons and Yao (1999) and is covered in books on categorical data analysis, e.g., Agresti (2002).

An iterative optimization algorithm for the maximum likelihood (ML) estimation (MLE) of the Bradley-Terry model has been known since the work by Zermelo (1929). Lange et al. (2000) showed that this algorithm belongs to the class of MM algorithms. Here MM refers to either minorize-maximization or majorize-minimization, depending on whether the optimization problem is maximization or minimization of an objective function. Lange (2016) provided a book account on MM optimization algorithms and Hunter and Lange (2004) provided a tutorial. Mairal (2015) established results on the convergence of incremental MM algorithms. A good feature of MM algorithms is not requiring to set a step size like for other iterative algorithms such as gradient descent with or without acceleration.

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

In a seminal paper, Hunter (2004) derived MM algorithms for generalized Bradley-Terry models as well as sufficient conditions for their convergence to ML estimators using the framework of MM optimization algorithms. For the Bradley-Terry model of paired comparisons, a necessary and sufficient condition for the existence of a ML estimator is that the directed graph whose vertices correspond to items and edges represent outcomes of paired comparisons is connected. In other words, the set of items cannot be partitioned in two sets such that none of the items in one partition won against an item in other partition.

A Bayesian inference method for generalized Bradley-Terry models was proposed by Caron and Doucet (2012), showing that classical MM algorithms can be reinterpreted as special instances of Expectation-Maximization (EM) algorithms associated with suitably defined latent variables and proposed some original extensions. This amounts to MM algorithms for maximum a posteriori probability (MAP) parameter estimation, for a specific family of prior distributions. This prior distribution is a product-form distribution with Gamma(α, β) marginal distributions, with the shape parameter $\alpha \geq 1$ and the rate parameter $\beta > 0$. Unlike to the ML estimation, the MAP estimator is always guaranteed to exist, for any observed data.

The MM algorithms for fitting model parameters of generalized Bradley-Terry models are implemented in open source software packages, including CRAN packages BradleyTerry2 by Turner and Firth (2012) and BradleyTerryScalable by Kaye and Firth (2017), as well as in the Python package Choix by Maystre (2018).

While the conditions for convergence of MM algorithms for generalized Bradley-Terry models are well understood, to the best of our knowledge, not much is known about their convergence rates for either ML or MAP estimation. In this paper, we close this gap by providing tight characterizations of convergence rates. Our results identify key properties of input data that determine the convergence rate, and in the case of MAP estimation, how the convergence rate depends on parameters of the prior distribution. Our results reveal that the MM algorithm, commonly used for MAP estimation for generalized Bradley-Terry models, can have a slow convergence for some prior distributions.

Recent research on statistical models of paired comparisons focused on characterization of the accuracy of parameter estimators and development of new, scalable parameter estimation methods, e.g., Guiver and Snelson (2009); Wauthier et al. (2013); Hajek et al. (2014); Rajkumar and Agarwal (2014); Chen and Suh (2015); Shah et al. (2016); Vojnovic and Yun (2016); Khetan and Oh (2016a); Negahban et al. (2017); Borkar et al. (2016). Although some recent algorithms show empirically faster convergence rate than the MM, e.g., Negahban et al. (2017); Maystre and Grossglauser (2015); Agarwal et al. (2018), it is hard to apply them for MAP estimation. We thus restrict our attention to the MM and the gradient descent algorithms which are able to solve both MLE and MAP optimization problems.

Summary of our contributions

We present tight characterizations of the rate of convergence of gradient descent and MM algorithms for ML and MAP estimation for generalized Bradley-Terry models. Our results show that both gradient descent and MM algorithms have linear convergence with the rates of convergence that differ only in constant factors. Note that an optimization algorithm with linear convergence is considered to be fast and many algorithms cannot guarantee a linear convergence. We provide explicit characterizations of the rate of convergence bounds that provide insights into key properties of the observed comparison data that determine the rate of convergence.

More specifically, we show that the rate of convergence critically depends on the properties of matrix $\mathbf{M}$, defined as the matrix of item pair co-occurrences in the observed comparison data. We found that the two key properties are: (a) maximum number of paired comparisons per item (denoted as $d(\mathbf{M})$) and (b) the algebraic connectivity of matrix $\mathbf{M}$ (denoted as $a(\mathbf{M})$). Intuitively, $a(\mathbf{M})$ quantifies how well is the graph of paired comparisons connected. Formally, $a(\mathbf{M})$ is the Fiedler value (eigenvalue) by Fiedler (1973), defined as the second smallest eigenvalue of the Laplacian matrix $\mathbf{L}_\mathbf{M} = \mathbf{D}_\mathbf{M} - \mathbf{M}$, where $\mathbf{D}_\mathbf{M}$ is a diagonal matrix whose diagonal elements are the row sums of $\mathbf{M}$. The Fiedler value of a matrix of paired comparison counts has been shown to play a key role in determining the MLE accuracy, e.g., Hajek et al. (2014); Shah et al. (2016); Khetan and Oh (2016b); Vojnovic and Yun (2016); Negahban et al. (2017). These works characterize the required number of samples to estimate the unknown value of the true model parameter. Note that this is different from the problem of characterizing the accuracy of an iterative optimization algorithm for computing the MLE value, which is studied in this paper.

Our results imply the following bounds on the convergence time, defined as the number of iterations for an iterative optimization algorithm to reach the value of the underlying objective function that is within a given error tolerance parameter $\epsilon > 0$ of the optimum value.

For the ML objective, the convergence time satisfies

$$T^{\mathrm{ML}} = O\left(\frac{d(\mathbf{M})}{a(\mathbf{M})} \log(1/\epsilon)\right). \qquad (1.1)$$

This reveals that the rate of convergence critically depends on the connectivity of the graph of paired comparisons in the observed data.

On the other hand, for the MAP estimation, we show that the convergence time satisfies

$$T^{\mathrm{MAP}} = O\left(\left(\frac{d(\mathbf{M})}{\beta}+1\right)\log(1/\epsilon)\right) \tag{1.2}$$

where, recall, $\beta > 0$ is the rate parameter of the Gamma prior distribution. This bound is shown to be tight for some input data instances. This shows that the MAP estimation can be arbitrarily slow by taking small enough parameter β. The small values of parameter β correspond to less informative prior distributions.

Our results identify a slow rate of convergence issue for gradient descent and MM algorithms in the case of MAP estimation. While the MAP estimation alleviates the issue of non-existence of a ML estimator when the graph of paired comparisons is disconnected, it can have a much slower convergence than ML when the graph of paired comparisons is connected. Perhaps surprisingly, the rate of convergence has a discontinuity at $\beta = 0$, in the sense that for $\alpha = 1$ and $\beta = 0$, the MM algorithm for the MAP estimation corresponds to the classic MM algorithm for ML estimation, and in this case, the convergence bound (1.1) holds, while for the MAP estimation, the convergence time grows arbitrarily large as β approaches 0 from above.

2 Preliminaries

In this section we define some basic concepts from convex optimization analysis that we use in the paper.

Gradient Descent. We consider standard gradient descent algorithm with constant step size $\eta > 0$:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)}). \tag{2.1}$$

MM algorithms. The MM algorithm for minimizing function f is defined by minimizing a surrogate function that *majorizes* function f. A surrogate function $g(\mathbf{x};\mathbf{y})$ is said to be a majorant function of f if $f(\mathbf{x}) \leq g(\mathbf{x};\mathbf{y})$ and $f(\mathbf{x}) = g(\mathbf{x};\mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. The MM algorithm is defined by the iteration:

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}} g(\mathbf{x};\mathbf{x}^{(t)}). \tag{2.2}$$

For maximizing a function f, we can analogously define the MM algorithm as minimization of a surrogate function g that *minorizes* function f. Majorization surrogate functions are used for minimization of convex functions, and minorization surrogate functions are used for maximization of concave functions.

Strong convexity Function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be γ-*strongly convex* on $\mathcal{X}$ if it satisfies the following subgradient inequality: for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^\top(\mathbf{x}-\mathbf{y}) - \frac{\gamma}{2}||\mathbf{x}-\mathbf{y}||^2.$$

Function f is γ-strongly convex on $\mathcal{X}$ if, and only if, $f(\mathbf{x}) - \frac{\gamma}{2}||\mathbf{x}||^2$ is convex on $\mathcal{X}$.

Smoothness Function f is said to be μ-*smooth* on $\mathcal{X}$ if its gradient vector ∇f is μ-Lipschitz on $\mathcal{X}$:

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq \mu||\mathbf{x}-\mathbf{y}||, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

For any function f that is μ-smooth on $\mathcal{X}$, the following property holds (see e.g. Lemma 3.4 in Bubeck (2015)):

$$\begin{aligned}&|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top(\mathbf{x}-\mathbf{y})|\\ &\leq \frac{\mu}{2}||\mathbf{x}-\mathbf{y}||^2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}.\end{aligned} \tag{2.3}$$

Polyak-Lojasiewicz (PL) inequality Polyak (1963) Function f is said to satisfy the Polyak-Lojasiewicz inequality on $\mathcal{X}$ if there exists $\gamma > 0$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\gamma}||\nabla f(\mathbf{x})||^2 \text{ for all } \mathbf{x} \in \mathcal{X} \tag{2.4}$$

where $\mathbf{x}^*$ is a minimizer of f. When the PL inequality holds on $\mathcal{X}$ for a specific value of γ, we say that γ-PL inequality holds on $\mathcal{X}$. If f is γ-strongly convex on $\mathcal{X}$, then f satisfies the γ-PL inequality on $\mathcal{X}$.

3 Convergence rates for the Bradley-Terry model

In this section, we provide characterizations of the rate of convergence for gradient descent and MM algorithms for ML and MAP estimation for the Bradley-Terry model of paired comparisons. We first establish general convergence theorems that hold for any strongly convex and smooth function f, which characterize the rate of convergence in terms of the strong-convexity and smoothness parameters of f, and a parameter of the surrogate function of the MM algorithm. The rate of convergence bounds are then derived for Bradley-Terry model by applying these theorems.

The results of this section can be extended to other instances of generalized Bradley-Terry models, including the Rao-Kupper model of paired comparisons with tie outcomes, the Luce choice model, and the Plackett-Luce ranking model. These extensions are established by following the same steps as for the Bradley-Terry model of paired comparisons. The differences lie in the characterization of the strong-convexity and smoothness parameters. The results provide characterizations of the convergence rates that are equivalent to those for the Bradley-Terry model of paired comparisons up to constant factors. We provide details in our supplementary material, Appendix C.

3.1 General convergence theorems

It is well known that for any μ-smooth function satisfying γ-PL inequality, the gradient descent algorithm (2.1) with a suitable choice of the step size η has a linear convergence with the rate of convergence $1-\gamma/\mu<1$. This result is due to Nesterov (2012) and a simple proof can be found in Boyd and Vandenberghe (2004), Chapter 9.3. The linear convergence and the rate of convergence bound follow from the following basic result.

Theorem 3.1 (gradient descent)**.** *Suppose that f is a convex function that is μ-smooth on $\mathcal{X}_\mu$ and that satisfies the γ-PL inequality on $\mathcal{X}_\gamma$. Let $\mathbf{x}^*$ be the minimizer of f and $\mathbf{x}^{(t+1)}$ be according to the gradient descent algorithm (2.1) with step size $\eta=1/\mu$. Then, if $\mathbf{x}^{(t)}\in\mathcal{X}_\gamma$ and $\mathbf{x}^{(t+1)}\in\mathcal{X}_\mu$, we have*

$$f(\mathbf{x}^{(t+1)})-f(\mathbf{x}^*)\leq\left(1-\frac{\gamma}{\mu}\right)(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)).$$

Proof of Theorem 3.1 is provided in Appendix A.1.

We next show a new theorem that establishes linear convergence of MM algorithms for any smooth and strongly convex function f that has a surrogate function g satisfying a condition relative to f.

Theorem 3.2 (MM)**.** *Suppose that f is a convex function that is μ-smooth on $\mathcal{X}_\mu$ and that satisfies the γ-PL inequality on $\mathcal{X}_\gamma\subseteq\mathcal{X}_\mu$, and g is a majorant surrogate function of f such that for some $\delta>0$, $g(\mathbf{x};\mathbf{y})-f(\mathbf{x})\leq\frac{\delta}{2}||\mathbf{x}-\mathbf{y}||^2$ for all $\mathbf{x},\mathbf{y}\in\mathcal{X}_\mu$. Let $\mathbf{x}^*$ be the minimizer of f and $\mathbf{x}^{(t+1)}$ be according to the MM algorithm (2.2). Then, if $\mathbf{x}^{(t)}\in\mathcal{X}_\gamma$ and $\mathbf{x}^{(t)}-\frac{1}{\mu+\delta}\nabla f(\mathbf{x}^{(t)})\in\mathcal{X}_\mu$, we have*

$$f(\mathbf{x}^{(t+1)})-f(\mathbf{x}^*)\leq\left(1-\frac{\gamma}{\mu+\delta}\right)(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)).$$

Proof of Theorem 3.2 is provided in Appendix A.2.

From Theorems 3.1 and 3.2, we observe that the MM algorithm has the same rate of convergence bound as the gradient descent algorithm except for the smoothness parameter μ being enlarged for value δ. If $\delta\leq c\mu$, for a constant $c>0$, then the MM algorithm has the same rate of convergence bound as the gradient descent algorithm, up to a constant factor.

A different set of sufficient conditions for linear convergence of MM algorithms can be found in Proposition 2.7 in Mairal (2015). These conditions require that g is a first-order surrogate function of f (Appendix B provides definition and more discussion). The rate of convergence bound derived from Theorem 3.2 can be tighter than a bound derived from Proposition 2.7 in Mairal (2015).

3.2 Maximum likelihood estimation

For the Bradley-Terry model, the log-likelihood function is given by

$$\ell(\mathbf{w})=\sum_{i=1}^n\sum_{j\neq i}d_{i,j}\left(w_i-\log\left(e^{w_i}+e^{w_j}\right)\right)\qquad(3.1)$$

where $d_{i,j}$ is the number of observed paired comparisons such that $i\succ j$.

The negative log-likelihood function has the following properties. Let $\lambda_i(\mathbf{A})$ denote the i-th smallest eigenvalue of matrix $\mathbf{A}$.

Lemma 3.1. *The negative log-likelihood function for the Bradley-Terry model is γ-strongly convex on $\mathcal{W}_{\omega,0}=\mathcal{W}_\omega\cap\{\mathbf{w}\in\mathbb{R}^n:\mathbf{w}^\top\mathbf{1}=0\}$, where $\mathcal{W}_\omega=\{\mathbf{w}\in\mathbb{R}^n:||\mathbf{w}||_\infty\leq\omega\}$, and μ-smooth on $\mathbb{R}^n$ with $\gamma=c_\omega\lambda_2(\mathbf{L_M})$ and $\mu=\frac{1}{4}\lambda_n(\mathbf{L_M})$ where $c_\omega=1/(e^{-\omega}+e^\omega)^2$.*

Proof of Lemma 3.1 is provided in Appendix A.3.

By Lemma 3.1, the smoothness parameter μ is proportional to the largest eigenvalue of the Laplacian matrix $\mathbf{L_M}$. By the well-known Gershgorin circle theorem, e.g. Theorem 7.2.1 in Golub and Loan (2013), we have $\lambda_n(\mathbf{L_M})\leq 2d(\mathbf{M})$. Thus, we can take $\mu=d(\mathbf{M})/2$. We will express all our convergence time results in terms of $d(\mathbf{M})$ instead of $\lambda_n(\mathbf{L_M})$. This is a tight characterization up to constant factors. When $\mathbf{M}$ is a graph adjacency matrix, then $\lambda_n(\mathbf{L_M})\geq d(\mathbf{M})+1$ (Grone et al. (1990)). In the context of paired comparisons, $d(\mathbf{M})$ has an intuitive interpretation as the maximum number of observed paired comparisons involving an item.

The following lemma is instrumental for deducing that a function f satisfies the γ-PL inequality from a γ-strong convexity condition on f.

Lemma 3.2. *Suppose that $\mathcal{X}$ is a convex set such that f is γ-strongly convex on $\mathcal{X}_0=\mathcal{X}\cap\{\mathbf{x}\in\mathbb{R}^n:\mathbf{x}^\top\mathbf{1}=0\}$. Assume that f is such that (C1) $f(\Pi_c(\mathbf{x}))=f(\mathbf{x})$ and (C2) $\nabla f(\Pi_c(\mathbf{x}))=\nabla f(\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}$ and all $\Pi_c(\mathbf{x})=\mathbf{x}+c\mathbf{1}$ for $c\in\mathbb{R}$. Then, f satisfies the γ-PL inequality on $\mathcal{X}$.*

The proof of Lemma 3.2 follows by noting that if f is γ-strongly convex on $\mathcal{X}_0$, then it satisfies the γ-PL inequality on $\mathcal{X}_0$. Since for every $\mathbf{x}\in\mathcal{X}$, $\mathbf{x}=\mathbf{x}'+c\mathbf{1}$ for some $\mathbf{x}' \in \mathcal{X}_0$ and $c \in \mathbb{R}$, by conditions (C1) and (C2) and the definition of the γ-PL inequality (2.4), it follows that if γ-PL inequality holds on $\mathcal{X}_0$, it holds as well on $\mathcal{X}$. Conditions (C1) and (C2) are satisfied by negative log-likelihood functions for generalized Bradley-Terry models.

Since the negative log-likelihood function of Bradley-Terry model satisfies conditions (C1) and (C2) of Lemma 3.2, combining with Lemma 3.1, we observe that it satisfies the γ-PL inequality on $\mathcal{W}_\omega$ with $\gamma = c_\omega a(\mathbf{M})$. Furthermore, by Lemma 3.1, the negative log-likelihood function is μ-smooth on $\mathbb{R}^n$ with $\mu = d(\mathbf{M})/2$. Combining these facts with Theorem 3.1, we obtain the following corollary:

Corollary 3.1 (gradient descent)**.** *Assume that $\mathbf{w}^*$ is the maximum likelihood estimate in $\mathcal{W}_\omega = \{\mathbf{w} : \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$, and that $\mathbf{w}^{(t+1)}$ is according to gradient descent algorithm with step size $\eta = 2/d(\mathbf{M})$. Then, if $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$, we have $\ell(\mathbf{w}^*) - \ell(\mathbf{w}^{(t+1)}) \leq \left(1 - 2c_\omega \frac{a(\mathbf{M})}{d(\mathbf{M})}\right)(\ell(\mathbf{w}^*) - \ell(\mathbf{w}^{(t)}))$.*

The result in Corollary 3.1 implies a linear convergence with the rate of convergence bound $1 - 2c_\omega a(\mathbf{M})/d(\mathbf{M})$. Hence, we have the following convergence time bound:

$$T = O\left(\frac{d(\mathbf{M})}{a(\mathbf{M})}\log(1/\epsilon)\right). \tag{3.2}$$

It is well known that the log-likelihood function for the Bradley-Terry model of paired comparisons is minorized by function

$$\underline{\ell}(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^n \sum_{j \neq i} \underline{\ell}_{ij}(\mathbf{x}; \mathbf{y}), \tag{3.3}$$

where

$$\underline{\ell}_{ij}(\mathbf{x}; \mathbf{y}) = d_{i,j}\left(x_i - \frac{e^{x_i} + e^{x_j}}{e^{y_i} + e^{y_j}} - \log\left(e^{y_i} + e^{y_j}\right) + 1\right).$$

The minorization surrogate function utilizes the elementary facts that $\log(x) \leq x - 1$ and that equality holds if, and only if, $x = 1$ to break $\log(e^{x_i} + e^{x_j})$ terms in the log-likelihood functions.

The classic MM algorithm is derived for the surrogate function (3.3). This surrogate function has the following property.

Lemma 3.3. *For all $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$, $\underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x}) \geq -\frac{\delta}{2}\|\mathbf{x} - \mathbf{y}\|^2$ where $\delta = \frac{1}{2}e^{2\omega}d(\mathbf{M})$.*

Proof of Lemma 3.3 is provided in Appendix A.4.

The optimal point that maximizies the minorization function admits a simple closed form, which yields the classic MM algorithm, e.g., Ford (1957); Hunter (2004), for the ML estimation: for $i = 1, 2, \ldots, n$,

$$w_i^{(t+1)} = \log\left(\frac{\sum_{j=1}^n d_{i,j}}{\sum_{j=1}^n \frac{m_{i,j}}{e^{w_i^{(t)}} + e^{w_j^{(t)}}}}\right). \tag{3.4}$$

By Theorem 3.2 and Lemmas 3.1, 3.2, and 3.3, we have the following corollary:

Corollary 3.2 (MM)**.** *Assume that $\mathbf{w}^*$ is the maximum likelihood estimate in $\mathcal{W}_\omega = \{\mathbf{w} : \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$, and that $\mathbf{w}^{(t+1)}$ is according to the MM algorithm. Then, if $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$, we have $\ell(\mathbf{w}^*) - \ell(\mathbf{w}^{(t+1)}) \leq \left(1 - 2c'_\omega \frac{a(\mathbf{M})}{d(\mathbf{M})}\right)(\ell(\mathbf{w}^*) - \ell(\mathbf{w}^{(t)}))$ where $c'_\omega = 1/[(e^{-\omega} + e^{\omega})^2(1 + e^{2\omega})]$.*

From Corollaries 3.1 and 3.2, we observe that both gradient descent and MM algorithms have the rate of convergence bound of the form $1 - ca(\mathbf{M})/d(\mathbf{M})$ for some constant $c > 0$. The only difference is the value of constant c. This shows that both gradient descent and MM algorithm have a linear convergence, and the convergence time bound (3.2).

3.3 Maximum a posteriori probability estimation

We consider the MAP estimation of the model parameters, following the Bayesian inference formulation proposed by Caron and Doucet (2012), under which the prior distribution of product-form with marginal distributions such that e^{w_i} has a Gamma distribution with the shape parameter α and the rate parameter β.

The log-a posteriori probability function can be written as $\rho(\mathbf{w}) = \ell(\mathbf{w}) + \ell_0(\mathbf{w}) + \text{const}$ where ℓ is the log-likelihood function given by (3.1) and ℓ_0 is the log-likelihood of the prior distribution given by

$$\ell_0(\mathbf{w}) = \sum_{i=1}^n \left((\alpha - 1)w_i - \beta e^{w_i}\right). \tag{3.5}$$

Note that for $\alpha = 1$ and $\beta = 0$, the log-a posteriori probability function corresponds to the log-likelihood function.

Lemma 3.4. *The negative log-a posteriori probability function for the Bradley-Terry model and the Gamma(α, β) prior distribution is γ-strongly convex and μ-smooth on $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$ with $\gamma = e^{-\omega}\beta$ and $\mu = \frac{1}{4}\lambda_n(\mathbf{L_M}) + e^{\omega}\beta$.*

Proof of Lemma 3.4 is provided in Appendix A.6.

By Theorem 3.1 and Lemma 3.4, we have the following corollary:

Corollary 3.3 (gradient descent)**.** *Suppose that $\mathbf{w}^*$ is the maximum a posteriori point in $\mathcal{W}_\omega = \{\mathbf{w} \in$*

$\mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$ *and* $\mathbf{w}^{(t+1)}$ *is according to gradient descent algorithm (2.1) with step size* $\eta = 2/(d(\mathbf{M}) + 2\beta e^\omega)$*. Then, if* $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$*, we have*

$$\frac{\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t+1)})}{\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)})} \leq 1 - \frac{2e^{-\omega}\beta}{d(\mathbf{M}) + 2e^\omega \beta}.$$

The result in Corollary 3.3 implies a linear convergence with the convergence time bound

$$T = O\left(\left(1 + \frac{d(\mathbf{M})}{\beta}\right)\log(1/\epsilon)\right). \tag{3.6}$$

This bound can be arbitrarily large by taking parameter β to be small enough. In Section 3.4, we show a simple instance for which this bound is tight. Hence, the convergence time for MAP estimation can be arbitrarily slow, and much slower than for the ML case.

We next consider the MM algorithm for the MAP problem of Caron and Doucet Caron and Doucet (2012). This MM algorithm is derived for the minorant surrogate function $\underline{\rho}$ of function ρ, which is defined as

$$\underline{\rho}(\mathbf{x}; \mathbf{y}) = \underline{\ell}(\mathbf{x}; \mathbf{y}) + \ell_0(\mathbf{x})$$

where $\underline{\ell}(\mathbf{x}; \mathbf{y})$ is the minorant surrogate function of the log-likelihood function (3.3) and ℓ_0 is the prior log-likelihood function (3.5).

The iterative updates of the MM algorithm are defined by, for $i = 1, 2, \ldots, n$,

$$w_i^{(t+1)} = \log\left(\frac{\alpha - 1 + \sum_{j\neq i} d_{i,j}}{\beta + \sum_{j \neq i} \frac{m_{i,j}}{e^{w_i^{(t)}} + e^{w_j^{(t)}}}}\right).$$

Note that this iterative optimization algorithm corresponds to the classic MM algorithm for ML estimation (3.4) when $\alpha = 1$ and $\beta = 0$.

Since $\underline{\rho}(\mathbf{x}; \mathbf{y}) - \rho(\mathbf{x}) = \underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x})$, by Lemma 3.3, we have

Lemma 3.5. *For all* $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$*,* $\underline{\rho}(\mathbf{x}; \mathbf{y}) - \rho(\mathbf{x}) \geq -\frac{\delta}{2}\|\mathbf{x} - \mathbf{y}\|^2$ *where* $\delta = \frac{1}{2}e^{2\omega}d(\mathbf{M})$*.*

By Theorem 3.2 and Lemmas 3.4 and 3.5, we have the following corollary:

Corollary 3.4 (MM)**.** *Suppose that* $\mathbf{w}^*$ *is the maximum a posteriori point in* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega\}$ *and* $\mathbf{w}^{(t+1)}$ *is according to the MM algorithm. Then, if* $\mathbf{w}^{(t)} \in \mathcal{W}_\omega$*, we have*

$$\frac{\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t+1)})}{\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)})} \leq 1 - \frac{2e^{-\omega}\beta}{(1 + e^{2\omega})d(\mathbf{M}) + 2e^\omega\beta}.$$

From Corollaries 3.2 and 3.4, we observe that both gradient descent and MM algorithm have the rate of convergence bound $1 - \Omega(\beta/(\beta + d(\mathbf{M})))$, and hence both have linear convergence and convergence time bound (3.6). The bound is tight as shown for a simple example in Section 3.4.

3.4 Tightness of the rate of convergence bound in Corollary 3.4

We show that the bound in Corollary 3.4 is tight for a simple case of two items. Let m denote the number of paired comparisons. Note that $d(\mathbf{M}) = m$ and $a(\mathbf{M}) = 2m$. Let d_1 and d_2 denote the number of paired comparisons won by items 1 and 2, respectively.

The MM algorithm iterates $\mathbf{w}^{(t)}$ are such that

$$e^{w_i^{(t+1)}} = \frac{d_i + \alpha - 1}{m + \beta s^{(t)}} s^{(t)}, \text{ for } i = 1, 2$$

where $s^{(t)} = e^{w_1^{(t)}} + e^{w_2^{(t)}}$. From this, observe that $s^{(t)}$ evolves according to the following autonomous nonlinear dynamical system:

$$s^{(t+1)} = \frac{m + 2(\alpha - 1)}{m + \beta s^{(t)}} s^{(t)}. \tag{3.7}$$

The limit point of $s^{(t)}$ as t goes to infinity is $2(\alpha - 1)/\beta$. Note that $(\alpha - 1)/\beta$ is the mode of Gamma(α, β).

Now, let us define $a^{(t)}$ by $s^{(t)} = [2(\alpha - 1)/\beta](1 + a^{(t)})$. Note that $a^{(t)}$ goes to 0 as t goes to infinity. By a tedious but straightforward calculus, we can show

$$\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)}) = 2(\alpha - 1)(a^{(t)} - \log(1 + a^{(t)})). \tag{3.8}$$

From (3.7), note that $1/a^{(t)}$ evolves according to a linear dynamical system, which allows us to derive the solution for $a^{(t)}$ in the explicit form given as follows:

$$a^{(t)} = \left(\frac{1}{1 - \frac{2(\alpha-1)}{\beta s^{(0)}}}\left(1 + \frac{2(\alpha - 1)}{m}\right)^t - 1\right)^{-1}.$$

From (3.8), $\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)}) = (\alpha - 1)(a^{(t)})^2(1 + o(1))$ for large t, and thus

$$\begin{aligned} &\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)}) \\ =&(\alpha - 1)\left(1 - \frac{2(\alpha - 1)}{\beta s^{(0)}}\right)^2 \left(1 + \frac{2(\alpha - 1)}{m}\right)^{-2t} (1 + o(1)). \end{aligned}$$

It follows that the rate of convergence of the log-a posteriori probability function is given as follows:

$$\lim_{t\to\infty} \frac{\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t+1)})}{\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)})} = \left(1 + \frac{2(\alpha - 1)}{m}\right)^{-2}.$$

The rate of convergence is approximately $1 - 4(\alpha - 1)/m$ for small $\alpha - 1$. By taking value of α such that $\alpha - 1 =$

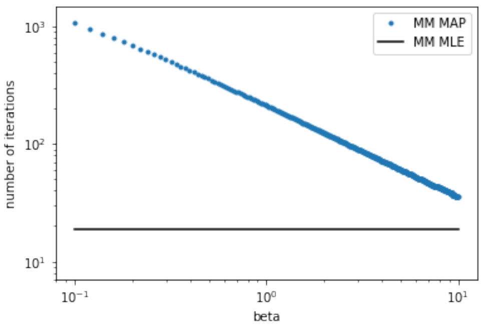


Figure 1: An illustrative numerical example: number of iterations until convergence versus parameter β. The MM algorithm can be slower for the MAP objective for several orders of magnitude than for the ML objective. The smaller the value of parameter β, the slower the convergence for MAP.

$c\beta$ for a constant $c > 0$ such that $||\mathbf{w}^*||_\infty \leq \omega$, we have the rate of convergence $1 - \Theta(\beta/d(\mathbf{M}))$. This establishes tightness of the rate of convergence bound in Corollary 3.4.

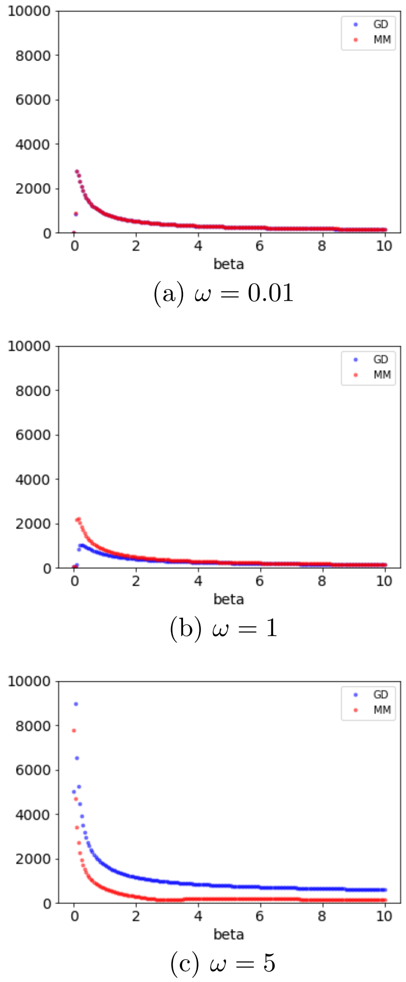
4 Numerical results

4.1 Simple illustrative numerical examples

Consider an instance with 10 items with each distinct pair of items compared 10 times and the input data generated according to Bradley-Terry model with parameters such that $w_1 = \cdots = w_5 = -\omega$ and $w_6 = \cdots = w_{10} = \omega$, for a parameter $\omega > 0$. We consider the convergence time T defined as the smallest T such that $||\mathbf{w}^{(T)} - \mathbf{w}^{(T-1)}||_\infty \leq \epsilon$, for a parameter $\epsilon > 0$. In our experiments, we set $\epsilon = 0.0001$.

The results in Figure 1, obtained for $\omega = 1/2$, show that MM algorithm for MAP estimation can be slower than for ML estimation for several orders of magnitude.

We further evaluate convergence time of gradient descent and MM algorithms for different values of parameter ω, for each distinct pair of items compared 100 times. The results in Figure 2 show the number of iterations versus the value of parameter β for gradient descent and MM algorithms, for different values of parameter ω. We observe that for small enough value of ω, convergence times of gradient descent and MM algorithms are nearly identical. Both algorithms have convergence time increasing by decreasing the value of parameter β for $\beta > 0$. We also observe a discontinuity in convergence time for $\beta = 0$ (MLE case) being smaller than for some small positive value of β (MAP case).



(a) $\omega = 0.01$

(b) $\omega = 1$

(c) $\omega = 5$

Figure 2: Number of iterations versus parameter β for gradient descent and MM algorithms, for different values of parameter ω.

The discontinuity at $\beta = 0$ originates from the fact that the log-likelihood function has infinitely many solutions for $\beta = 0$, but has a unique solution whenever $\beta > 0$. Consider a simple illustrative example: $f(x_1, x_2) = (x_1 - 1)^2 + \theta(x_2 - 1)^2$, for a parameter $\theta \geq 0$. Then, the GD converges to the unique solution $(1, 1)$ very slowly when θ is close to 0. When $\theta = 0$, however, we just need to find the minimum point of $(x_1 - 1)^2$ which can be solved in a few iterations.

4.2 Real-world datasets

In this section, we present evaluations of convergence times of gradient descent and MM algorithms for three different real-world datasets. The three datasets are described as follows.

GIFGIF This dataset contains user evaluations of digital images by paired comparisons with respect to different metrics, such as amusement, content, and happiness. The dataset was collected through an online

Table 1: Dataset properties.

Dataset	m	n	$d(\mathbf{M})$	$a(\mathbf{M})$
GIFGIF: A (full)	161,584	6,123	83	0
GIFGIF: C (full)	108,126	6,122	56	0
GIFGIF: H (full)	225,695	6,124	153	0
GIFGIF: A (sample)	702	252	15	0.67
GIFGIF: C (sample)	734	256	28	0.57
GIFGIF: H (sample)	1040	251	23	1.36
Chess (full)	65,030	8,631	155	0
Chess (sample)	13,181	985	135	1.77
NASCAR	64,596	83	1,507	39.34

web service by the MIT Media Lab as part of the PlacePulse project by Rich et al. (2018). This service presents the user with a pair of images and asks to select one that better expresses a given metric, or select neither. The dataset contains 1,048,576 observations and covers 17 metrics. We used this dataset to evaluate convergence of MM algorithms for Bradley-Terry model of paired comparisons, for the three aforementioned metrics.

Chess This dataset contains game-by-game results for 65,030 matches among 8,631 chess players. The dataset was used in the Kaggle chess ratings competition by Sonas (2010). Each observation contains information for a match between two players including unique identifiers of the players, information about which one of the two players played with white figures, and the result of the match, which is either win, loss, or draw. This dataset has a large degree of sparsity. We used this dataset to evaluate convergence of the Rao-Kupper model of paired comparisons with ties.

NASCAR This dataset contains auto racing competition results. Each observation is for an auto race and contains the ranking of drivers in increasing order of their finish times in the race. The dataset is available from the web page maintained by Hunter (2003). This dataset was previously used for evaluation of MM algorithms for the Plackett-Luce ranking model by Hunter (2004) as more recently by Caron and Doucet (2012). We used this dataset to evaluate convergence times of MM algorithms for the Plackett-Luce ranking model.

We summarise some of the key properties of each dataset in Table 1. We use a shorthand notation GIFGIF: A, GIFGIF: C, and GIFGIF: H to denote datasets for metrics amusement, contempt, and happiness, respectively. For full GIFGIF and Chess datasets, we can split the items into two groups such that there exists one item in one group that was not compared with any item in the other group, i.e. the algebraic connectivity of matrix $\mathbf{M}$ is zero. In this case, there exists no ML estimate. For this reason, we also consider sampled datasets for which a ML estimate exists. This subsampling was done by selecting the largest connected component of items.

Table 2: Number of iterations until ϵ-convergence for gradient descent (GD) and MM algorithms. 'n/a' indicates cases when a ML estimate does not exist.

Dataset	Algo.	$\beta = 0$	0.01	1	10
GIFGIF: A (full)	GD	n/a	14,965	432	46
	MM	n/a	572	70	16
GIFGIF: C (full)	GD	n/a	12,745	299	33
	MM	n/a	733	49	13
GIFGIF: H (full)	GD	n/a	26,512	792	77
	MM	n/a	1,127	98	21
GIFGIF: A (sample)	GD	516	1,914	83	12
	MM	145	854	26	7
GIFGIF: C (sample)	GD	769	2,055	121	18
	MM	130	694	39	9
GIFGIF: H (sample)	GD	434	2,452	100	25
	MM	216	1,234	38	8
Chess (full)	GD	n/a	36,598	725	64
	MM	n/a	2,217	113	33
Chess (sample)	GD	529	552	314	57
	MM	121	122	74	19
NASCAR	GD	291	1,518	140	30
	MM	11	695	58	10

Numerical results presented in Table 2 validate the following observations derived from our theoretical results: (a) the convergence time increases by decreasing the value of parameter β for $\beta > 0$, which can be for a substantial amount, and (b) there is a discontinuity in the convergence time being much smaller for $\beta = 0$ (MLE case) than for a small value $\beta > 0$.

5 Conclusion

We have shown that for the ML parameter estimation of generalized Bradley-Terry models, gradient descent and MM algorithms have a linear convergence and the convergence time bound $O(d(\mathbf{M})/a(\mathbf{M}))$, which we can interpret as the condition number of the Laplacian matrix $\mathbf{L_M}$. We have further shown that for the MAP parameter estimation with the product-form prior with Gamma(α, β) marginal distributions, gradient descent and MM algorithms have a linear convergence and the convergence time bound $O(d(\mathbf{M})/\beta + 1)$, which is shown to be tight. For any fixed value $\beta > 0$, the convergence time bound is proportional to the maximum eigenvalue of the Laplacian matrix $\mathbf{L_M}$. Our results identify a slow convergence of gradient descent and MM algorithms for MAP estimation, which occurs for small values of β (less informative prior). Our results also reveal a discontinuity of the convergence time for MAP estimation at $(\alpha, \beta) = (1, 0)$, which is a point corresponding to ML estimation.

Acknowledgements

Yun would like to acknowledge the support from Institute of Information & communications Technology Planning & Evaluation (IITP) grant No. 2019000075 and National Research Foundation of Korea (NRF) grant No. 2019028324 funded by the government of Korea (MSIT).

References

Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, pages 70–79, 2018.

Alan Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics, 2 edition, 2002.

David Balduzzi, Karl Tuyls, Julien Pérolat, and Thore Graepel. Re-evaluating evaluation. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS '18)*, 2018.

Vivek S Borkar, Nikhil Karamchandani, and Sharad Mirani. Randomized kaczmarz for rank aggregation from pairwise comparisons. In *2016 IEEE Information Theory Workshop (ITW)*, pages 389–393. IEEE, 2016.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

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

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4): 231–357, 2015.

Christopher J. Burges, Robert Ragno, and Quoc V. Le. Learning to rank with nonsmooth cost functions. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS '06)*, pages 193–200. 2006.

F. Caron and A. Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *J. Comp. Graph. Statist.*, 21(1):174–196, 2012.

Yuxin Chen and Changho Suh. Spectral MLE: Top-K rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning (ICML '15)*, pages 371–380, 2015.

Otto Dykstra, Jr. A note on the rank analysis of incomplete block designs – applications beyond the scope of existing tables. *Biometrics*, 12(3):301–306, 1956.

Otto Dykstra, Jr. Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. *Biometrics*, 16(2):176–188, 1960.

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

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.

L. R. Ford. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 4 edition, 2013.

R. Grone, R. Merris, and V. Sunder. The Laplacian spectrum of a graph. *SIAM Journal on Matrix Analysis and Applications*, 11(2):218–238, 1990.

John Guiver and Edward Snelson. Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*, Montreal, Canada, 2009.

Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS '14)*, pages 1475–1483, Montreal, Canada, 2014.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill$^{\mathrm{TM}}$: A Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS '06)*, pages 569–576. Canada, 2006.

Tzu-Kuo Huang, Ruby C. Weng, and Chih-Jen Lin. Generalized Bradley-Terry Models and multi-class probability estimates. *J. Mach. Learn. Res.*, 7:85–115, December 2006.

Tzu-Kuo Huang, Chih-Jen Lin, and Ruby C. Weng. Ranking individuals by group comparisons. *J. Mach. Learn. Res.*, 9:2187–2216, 2008.

David R Hunter. MATLAB code for Bradley-Terry models, 2003. URL `http://personal.psu.edu/drh20/code/btmatlab/`.

David R Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 32(1): 384–406, 2004.

David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1): 30–37, 2004.

Ella Kaye and David Firth. BradleyTerryScalable: Fits the Bradley-Terry Model to Potentially Large and Sparse Networks of Comparison Data,

2017. URL https://cran.r-project.org/web/packages/BradleyTerryScalable/.

Ashish Khetan and Sewoong Oh. Computational and statistical tradeoffs in learning to rank. In *Advances in Neural Information Processing Systems 29*, pages 739–747. 2016a.

Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016b.

K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

Kenneth Lange. *MM Optimization Algorithms*. SIAM, 2016.

Hang Li. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool, 2011.

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

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Lucas Maystre. Choix: Inference algorithms for models based on Luce's choice axiom, 2018. URL https://github.com/lucasmaystre/choix.

Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. In *Advances in neural information processing systems*, pages 172–180, 2015.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

Yuri Nesterov. Gradient methods for minimizing composite objective functions. *Math. Program.*, 140: 125–161, 2012.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

Boris Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3:864–878, 12 1963.

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning (ICML '14)*, pages 118–126, Bejing, China, 22–24 Jun 2014.

P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

Travis Rich, Kevin Hu, and Basheer Tome. GIFGIF-mapping the empotional language of gifs, 2018. URL http://gifgif.media.mit.edu.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal Machine Learning Research*, 17 (1):2049–2095, January 2016.

Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.

Jeff Sonas. Kaggle competition: Chess ratings - elo versus the rest of the world, 2010. URL https://www.kaggle.com/c/chess.

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

Heather Turner and David Firth. Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software, Articles*, 48(9):1–21, 2012.

Milan Vojnovic and Se-Young Yun. Parameter estimation for generalized Thurstone choice models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML '16)*, pages 498–506, 2016.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning (ICML '13)*, pages 109–117, Atlanta, Georgia, USA, 17–19 Jun 2013.

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