

CONVERGENCE RATES OF GRADIENT DESCENT AND MM ALGORITHMS FOR GENERALIZED BRADLEY-TERRY MODELS

BY MILAN VOJNOVIC^{*}, SEYOUNG YUN[†] AND KAIFANG ZHOU^{*}

London School of Economics^{} and KAIST[†]*

We show tight convergence rate bounds for gradient descent and MM algorithms for maximum likelihood estimation and maximum a posteriori probability estimation of a popular Bayesian inference method for generalized Bradley-Terry models. This class of models includes the Bradley-Terry model of paired comparisons, the Rao-Kupper model of paired comparisons with ties, the Luce choice model, and the Plackett-Luce ranking model. Our results show that MM algorithms have same convergence rates as gradient descent algorithms up to constant factors. For the maximum likelihood estimation, 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 Bayesian inference, 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. We propose a simple, first-order acceleration method that resolves the slow convergence issue.

1. Introduction. According to the Bradley-Terry model of paired comparisons [5], comparisons of pairs of items are independent and the outcome of each comparison of items i and j is one of the items winning with

$$(1.1) \quad \Pr[i \text{ wins against } j] = \frac{\theta_i}{\theta_i + \theta_j}$$

where θ_i is a positive-valued parameter representing the strength of item i .

The Bradley-Terry model dates back to 1929, from the early work of Zermelo [49], and was subsequently popularized by the work of Bradley and Terry [5, 4]. David [10] covers early references. The Bradley-Terry model was studied by many, e.g. [15, 11, 12, 43], and is covered by classic books on categorical data analysis [1]. The original Bradley-Terry model has been extended in various directions, e.g. to allow for tie outcomes (e.g. the Rao-Kupper model [40]), choice (the Luce choice model [31]) and ranking outcomes (the Plackett-Luce ranking model [37]) for comparison sets of two

Keywords and phrases: Rank aggregation, Generalized Bradley-Terry model, Maximum likelihood estimation, Bayesian inference, Convex optimization, Gradient descent, MM algorithm, Rate of convergence

or more items, as well as group comparisons (e.g. [22, 21]). These models are commonly referred to as generalized Bradley-Terry models. They can be derived from suitably defined latent variable models, where each item is associated with a latent performance random variable. This latent variable representation is in spirit of the well-known Thurstone model of comparative judgment [45]. Statistical models of paired comparisons have been a subject of intense recent research focused on characterization of the accuracy of parameter estimators and proposing new, scalable parameter estimation methods [18, 48, 19, 39, 9, 42, 47, 27, 35].

Statistical models of ranking data play an important role in applications. The Bradley-Terry model of paired comparisons underlies the design of the Elo rating system, used for rating skills of chess players [13]. Extensions to team competitions and tie outcomes were implemented in popular on-line gaming platforms, e.g. TrueSkill rating system [20]. The generalized Bradley-Terry type of models have been used for estimation of relevance of items in information retrieval applications, e.g. learning to rank [7, 30]. Statistical models of paired comparisons are used in timely applications such as evaluation of reinforcement learning algorithms [2].

An iterative optimization algorithm for maximum likelihood (ML) estimation of the Bradley-Terry model parameter vector has been known since the original work of Zermelo [49]. Lange, Hunter and Yang [29] showed that this algorithm belongs to the class of MM optimization algorithms. Here MM refers to either minorize-maximization or majorize-minimization, depending on whether the optimization problem is a maximization or a minimization problem. The book by Lange [28] provides an excellent account on MM optimization algorithms while [25] provides a tutorial. Recent work by Mairal [32] provides results on the convergence of MM algorithms.

In a seminal paper, Hunter [24] derived MM algorithms for generalized Bradley-Terry models and sufficient conditions for their convergence to maximum likelihood 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 maximum likelihood estimator is that the directed graph whose vertices correspond to items and edges represent outcomes of paired comparisons is connected. In other words, there exists no partition of items in two disjoint 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 [8], 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. Specifically, the prior distribution is a product-form distribution with $\text{Gamma}(\alpha, \beta)$ marginal distributions, with the shape parameter $\alpha \geq 1$ and the rate parameter $\beta > 0$. Unlike to the ML estimation case, the MAP estimator is guaranteed to exist, for any observed data.

The MM algorithms for fitting generalized Bradley-Terry models are implemented in popular CRAN packages `BradleyTerry2` [46] and `BradleyTerryScalable` [26], for ML and MAP estimation, respectively. They are also implemented in Python packages such as `Choix` [33].

While the conditions for convergence of MM algorithms for generalized Bradley-Terry models are well known, to the best of our knowledge, not much was known about their convergence rates for either ML or MAP estimation. In this paper, we address this by providing tight characterizations of convergence rates. Our results identify key properties of input comparison data that determine convergence rates, and in the case of MAP estimation, how the convergence rates depend on parameters of the prior distribution. Our results reveal that popular MM algorithms used for MAP estimation for generalized Bradley-Terry models can have a slow convergence for certain values of parameters of the prior distribution. We address this by proposing a new, first-order acceleration method. We summarize our main contributions in some more details as follows.

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 rates of convergence that differ only in constant factors. We provide explicit characterizations of the rate of convergence bounds that provide insights into the key properties of the observed comparison data that determine the rate of convergence.

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. Specifically, two key properties are: (a) the 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) [14], 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} .

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

$$(1.2) \quad T^{\text{ML}} = O\left(\frac{d(\mathbf{M})}{a(\mathbf{M})} \log(1/\epsilon)\right).$$

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

$$(1.3) \quad T^{\text{MAP}} = O\left(\left(1 + \frac{d(\mathbf{M})}{\beta}\right) \log(1/\epsilon)\right)$$

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 (more vague) 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 resolves the non-existence of a maximum likelihood estimator when the graph of paired comparisons is disconnected, it can have 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 boils down to the classic MM algorithm for ML estimation, and in this case the convergence bound (1.2) holds, while for the MAP estimation, the convergence time grows arbitrarily large as β approaches 0 from above.

We propose a simple first-order acceleration method for the MAP estimation whose convergence time satisfies

$$(1.4) \quad T_{\text{Acc}}^{\text{MAP}} = O\left(\min\left\{\frac{d(\mathbf{M})}{a(\mathbf{M})}, \frac{d(\mathbf{M})}{\beta}\right\} \log(1/\epsilon)\right).$$

This acceleration method resolves the slow convergence issue of standard MM algorithms for the MAP estimation for generalized Bradley-Terry models. Note that this accelerated method does not have a discontinuity at $\beta = 0$ with respect to the rate of convergence: as β approaches 0 from above, the convergence time bound corresponds to that of the MM algorithm for ML

estimation. The acceleration method applies a transformation of the parameter vector in each iteration of the gradient descent or MM algorithm in a way that ensures moving in directions in which the objective function has certain smoothness and strong convexity properties. The acceleration method is derived by using a theoretical framework that may be of general interest. This framework can be applied to different statistical models of ranking data and prior distributions for Bayesian inference of these models.

We present numerical evaluation results for the convergence rate of different iterative optimization algorithms using input data comparisons from a collection of real-world datasets. These results demonstrate the extent of the slow convergence issue of the existing MM algorithms for MAP estimation and show significant speed ups achieved by our accelerated MM algorithms.

Our theoretical results are established by using the framework of convex optimization analysis and spectral theory of matrices. In particular, the convergence rate bounds are obtained using the concepts such as smoothness, strong convexity, and Polyak-Loyasiewicz condition. We derived accelerated iterative optimization algorithms based on a general approach that may be of independent interest. This approach amounts to transforming the parameter estimator in each iteration such that certain conditions hold for the gradient vector and the Hessian matrix of the objective function. In particular, the transformation ensures orthogonality of the gradient vector to a vector that is an eigenvector of the Hessian matrix, which corresponds to an eigenvalue of small value. For generalized Bradley-Terry models, this transformation is simple to implement.

Organization of the paper. In Section 2, we introduce various instances of generalized Bradley-Terry models, ML and MAP model fitting objectives, gradient descent and MM algorithms, and some key concepts of convex optimization that we use in the paper. Section 3 contains our main results on the characterization of convergence rates of gradient descent and MM algorithms, where for simplicity of exposition, we focus only on the Bradley-Terry model of paired comparisons. Section 4 presents our accelerated algorithm for MAP estimation. In Section 5 we show how our results extend to other instances of generalized Bradley-Terry model. Section 6 presents numerical results. We discuss our results and conclude in Section 7. All our proofs are provided in Appendix (accompanying supplementary material).

2. Preliminaries. In this section we introduce several instances of generalized Bradley-Terry models, define maximum likelihood and maximum a posteriori probability estimation, and define some basic concepts from convex optimization analysis that we use throughout the paper.

2.1. Generalized Bradley-Terry models.

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

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

where $\mathbf{w} = (w_1, w_2, \dots, w_n)^\top \in \mathbb{R}^n$ are model parameters.

Rao-Kupper model of paired comparisons with ties. The Rao-Kupper model is such that each paired comparison of items i and j has three possible outcomes: either $i \succ j$ or $j \succ i$ or $i \equiv j$ (tie). The model is defined by the probability distribution of outcomes that is given by

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

and

$$\Pr[i \equiv j] = \frac{(\theta^2 - 1)e^{w_i}e^{w_j}}{(e^{w_i} + \theta e^{w_j})(\theta e^{w_i} + e^{w_j})}$$

where $\mathbf{w} = (w_1, w_2, \dots, w_n)^\top \in \mathbb{R}^n$ and $\theta \geq 1$ are model parameters.

The larger the value of parameter θ , the more mass is put on the tie outcome. For the value of parameter $\theta = 1$, the model corresponds to the Bradley-Terry model for paired comparisons.

Luce choice model. The Luce choice model is a natural generalization of the Bradley-Terry model of paired comparisons to comparison sets of two or more items. For any given comparison set $S \subseteq N = \{1, 2, \dots, n\}$ of two or more items, the outcome is a choice of one item $i \in S$ (an event we denote as $i \succeq S$) which occurs with probability

$$\Pr[i \succeq S] = \frac{e^{w_i}}{\sum_{j \in S} e^{w_j}}$$

where $\mathbf{w} = (w_1, w_2, \dots, w_n)^\top \in \mathbb{R}_n$ are model parameters.

We will use the following definitions and notation. Let T be the set of ordered sequences of two or more items from N such that for each $y = (y_1, y_2, \dots, y_k) \in T$, y_1 is an arbitrary item and y_2, \dots, y_k are sorted in lexicographical order. We can interpret each $y = (y_1, y_2, \dots, y_k) \in T$ as a

choice of item y_1 from the set of items $\{y_1, y_2, \dots, y_k\}$. According to the Luce's choice model, the probability of outcome y is given by

$$\Pr[Y = (y_1, y_2, \dots, y_k)] = \frac{e^{w_{y_1}}}{\sum_{j \in y} e^{w_j}}.$$

We denote with d_y the number of observed outcomes y in the input data. For each $y \in T$, let $|y|$ denote the number of items in y .

Plackett-Luce ranking model. The Plackett-Luce ranking model is a model of full rankings: for each comparison set of items $S \subseteq N = \{1, 2, \dots, n\}$, the set of possible outcomes contains all possible permutations of items in S . The distribution over possible outcomes is defined as follows. Let T be the set of all possible permutations of subsets of two or more items from N . Each $y = (y_1, y_2, \dots, y_k) \in T$ corresponds to a permutation of the set of items $S = \{y_1, y_2, \dots, y_k\}$. The probability of outcome y is given by

$$\Pr[Y = (y_1, y_2, \dots, y_k)] = \frac{e^{w_{y_1}}}{\sum_{j=1}^k e^{w_{y_j}}} \frac{e^{w_{y_2}}}{\sum_{j=2}^k e^{w_{y_j}}} \cdots \frac{e^{w_{y_{k-1}}}}{\sum_{j=k-1}^k e^{w_{y_j}}}$$

where $\mathbf{w} = (w_1, w_2, \dots, w_n)^\top \in \mathbb{R}^n$ are model parameters.

The model has an intuitive explanation as a sampling of items without replacement proportional to the item weights e^{w_i} . The Plackett-Luce ranking model corresponds to the Bradley-Terry model of paired comparisons when the comparison sets consist of two items. We denote with d_y the number of observed outcomes y in the input data.

2.2. ML and MAP estimation. Let $\ell : \mathbb{R}^n \rightarrow \mathbb{R}$ be the log-likelihood function of the model under consideration. The maximum likelihood estimator \mathbf{w}^{ML} is defined as follows

$$\mathbf{w}^{\text{ML}} = \arg \max_{\mathbf{w} \in \mathbb{R}^n} \ell(\mathbf{w}).$$

For generalized Bradley-Terry models, we can express the log-likelihood function as

$$\ell(\mathbf{w}) = \sum_{y \in T} d_y \log(p_{\mathbf{w}}(y))$$

where T is the set of all possible outcomes of comparisons, d_y is the number of observations of y in the input data and $p_{\mathbf{w}}(y)$ is probability of observing outcome y according to the underlying model with parameter vector \mathbf{w} .

We will consider the *negative log-likelihood function* f , defined as $f(\mathbf{w}) = -\ell(\mathbf{w})$ for $\mathbf{w} \in \mathbb{R}^n$. Indeed, $\mathbf{w}^{\text{ML}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} f(\mathbf{w})$.

The MAP estimation is defined by following the Bayesian approach under which \mathbf{w} is assumed to be a random variable with a given prior distribution p_0 . Conditional on the observed data D , the posterior distribution of \mathbf{w} is given by the Bayes' formula:

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})p_0(\mathbf{w})}{p(D)}.$$

The *log-aposteriori probability function*, defined as $\rho(\mathbf{w}) = \log(p(\mathbf{w} \mid D))$, can be written as

$$\rho(\mathbf{w}) = \ell(\mathbf{w}) + \ell_0(\mathbf{w}) + \text{const}$$

where ℓ is the log-likelihood function and ℓ_0 is the *prior log-likelihood function* $\ell_0(\mathbf{w}) = \log(p_0(\mathbf{w}))$.

The maximum aposteriori probability estimator is defined as

$$\mathbf{w}^{\text{MAP}} = \arg \max_{\mathbf{w} \in \mathbb{R}^n} \rho(\mathbf{w}).$$

2.3. Gradient descent and MM algorithms. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function. The gradient descent algorithm for minimizing function f in a general form is defined by

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \mathbf{B}_t^{-1} \nabla f(\mathbf{x}^{(t)})$$

where η_t is the step size, \mathbf{B}_t is an invertible matrix, and $\nabla f(\mathbf{x})$ is the gradient vector of f at point \mathbf{x} . If \mathbf{B}_t is the identity matrix, we have the standard gradient descent algorithm. On the other hand, if $\mathbf{B}_t = \nabla^2 f(\mathbf{x}^{(t)})$, where $\nabla^2 f(\mathbf{x})$ is the Hessian matrix of f at \mathbf{x} , we have the Newton's algorithm. The standard gradient descent algorithm is a first-order method because it requires only access to the value of $\nabla f(\mathbf{x})$ for a query point \mathbf{x} .

In this paper, we only consider the standard gradient algorithm with constant step size η , we refer to as gradient descent algorithm, i.e.

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

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:

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

For maximizing a function f , we can analogously define the MM algorithm as minimization of a surrogate function g that *minorizes* function f . In

particular, we consider a majorization surrogate function for minimization of a convex function, and a minorization surrogate function for maximization of a concave function.

The log-likelihood function of the Bradley-Terry model of paired comparisons is minorized by function

$$(2.3) \quad \underline{\ell}(\mathbf{x}; \mathbf{y}) = \sum_{i=1}^n \sum_{j \neq i} d_{i,j} \left(x_i - \frac{e^{x_i} + e^{x_j}}{e^{y_i} + e^{y_j}} - \log(e^{y_i} + e^{y_j}) + 1 \right)$$

where $d_{i,j}$ is the number of observed paired comparisons such that $i \succ j$. Here, we utilize the facts that $\log(x) \leq x - 1$ and the equality holds if and only if $x = 1$ to break $\log(e^{x_i} + e^{x_j})$ terms in the log-likelihood functions.

The optimal point that maximizes the minorization function admits a simple closed form, which yields the classic MM algorithm [15, 24] for the ML estimation: for $i = 1, 2, \dots, n$,

$$(2.4) \quad e^{w_i^{(t+1)}} = \left(\sum_{j \neq i} d_{i,j} \right) \left(\sum_{j \neq i} \frac{m_{i,j}}{e^{w_i^{(t)}} + e^{w_j^{(t)}}} \right)^{-1}.$$

Similar surrogate functions and MM algorithms are known for generalized Bradley-Terry models, which are provided in Section 5.

2.4. Convex optimization analysis. Here we review some basic concepts of convex optimization analysis that are used throughout the paper.

Strong convexity. Function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be γ -strongly convex on \mathcal{X} if it satisfies the following subgradient inequality:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{y}) - \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|^2, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

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} . If f is twice differentiable, then the eigenvalues of the Hessian of f are larger than or equal to γ , i.e. $\nabla^2 f(\mathbf{x}) \succeq \gamma \mathbf{I}_n$ for all $\mathbf{x} \in \mathcal{X}$.

Smoothness. Function f is said to be μ -smooth on \mathcal{X} if its gradient vector ∇f is μ -Lipschitz on \mathcal{X} , i.e.

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

If f is twice differentiable, then μ -smoothness is equivalent to the largest eigenvalue of the Hessian of f being smaller than or equal to μ at any point, i.e. $\nabla^2 f(\mathbf{x}) \preceq \mu \mathbf{I}_n$, for all $\mathbf{x} \in \mathcal{X}$.

For any function f that is μ -smooth on \mathcal{X} , the following property holds (see e.g. Lemma 3.4 [6]):

$$(2.5) \quad |f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Polyak-Lojasiewicz (PL) inequality [38]. Function f is said to satisfy the Polyak-Lojasiewicz inequality on \mathcal{X} if there exists $\gamma > 0$ such that

$$(2.6) \quad f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\gamma} \|\nabla f(\mathbf{x})\|^2 \text{ for all } \mathbf{x} \in \mathcal{X}$$

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} .

Linear convergence. A sequence $f(\mathbf{x}^{(0)}), f(\mathbf{x}^{(1)}), \dots$ is said to converge *linearly* to $f(\mathbf{x}^*)$ as t goes to infinity, if there exists $c \in (0, 1)$ such that

$$\lim_{t \rightarrow \infty} \frac{f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*)}{f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)} = c$$

where $f(\mathbf{x}^*)$ is the optimum value. The value c is referred to as *the rate of convergence*. We define $\bar{c} \in (0, 1)$ as *the rate of convergence bound*, if for some $t_0 \geq 0$,

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*) \leq \bar{c}(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)) \text{ for all } t \geq t_0.$$

Note that any rate of convergence bound is an upper bound on the rate of convergence.

For a convergent sequence $f(\mathbf{x}^{(0)}), f(\mathbf{x}^{(1)}), \dots$ to $f(\mathbf{x}^*)$ and a tolerance parameter $\epsilon \in (0, 1]$, let T be the smallest integer $t \geq 0$ such that $|f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)| \leq \epsilon$. We refer to T as the ϵ -convergence time, or simply the convergence time. If $f(\mathbf{x}^{(0)}), f(\mathbf{x}^{(1)}), \dots$ is a linearly convergent sequence with the rate of convergence bound \bar{c} , then

$$T = O\left(\frac{1}{1 - \bar{c}} \log(1/\epsilon)\right).$$

3. Convergence rates for the Bradley-Terry model. In this section, we provide the rate of convergence characterizations 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 smoothness parameter of the surrogate function of the MM algorithm. The rate of convergence bounds are then derived for the Bradley-Terry model by applying these general theorems.

3.1. *General convergence theorems.* It is a well-known fact that gradient descent algorithm (2.1) with a suitable choice of the step size η has linear convergence with the rate of convergence $1 - \gamma/\mu < 1$, for any γ -strictly convex and μ -smooth function. This result is due to Nesterov [36] and a simple proof can be found in Boyd and Vandenberghe [3], Chapter 9.3. The linear convergence and the rate of convergence bound follow from the following fact, which we note for future reference.

THEOREM 3.1 (gradient descent). *Suppose f is a convex function that is μ -smooth on \mathcal{X}_μ and that satisfies the γ -PL inequality on \mathcal{X}_γ . 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 theorem that establishes linear convergence of MM algorithms for any function that satisfies certain smoothness and strong convexity conditions and a surrogate function g that satisfies a smoothness condition with respect to f .

THEOREM 3.2 (MM). *Suppose f is a convex function that is μ -smooth on \mathcal{X}_μ 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.

The smoothness condition imposed on the surrogate function g in Theorem 3.2 is related to the notion of the first-order surrogate functions introduced by Mairal [32]. A surrogate function g is said to be a *first-order surrogate function* of f on \mathcal{X} if g is a surrogate function of f on \mathcal{X} and the error function defined by $h(\mathbf{x}; \mathbf{y}) = g(\mathbf{x}; \mathbf{y}) - f(\mathbf{x})$ is μ_0 -smooth on \mathcal{X} for some $\mu_0 > 0$. If g is a first-order surrogate function of f on \mathcal{X} with parameter μ_0 , then $g(\mathbf{x}; \mathbf{y}) - f(\mathbf{x}) \leq \frac{\mu_0}{2} \|\mathbf{x} - \mathbf{y}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ (Lemma 2.3 [32]).

Thus, requiring that g is a first-order surrogate function of f on \mathcal{X}_μ with parameter δ is a sufficient condition for the condition on g in Theorem 3.2.

A set of sufficient conditions for linear convergence of MM algorithms was found by Mairal [32] (Proposition 2.7 therein). These conditions require that g is a first-order surrogate function of f . The result can be summarized as follows. Suppose that f is γ -strongly convex on \mathcal{X}_γ and g is a first-order surrogate function of f on \mathcal{X}_μ with parameter μ_0 . Let $\mathbf{x}^{(t+1)}$ be according to the MM algorithm. Then, if $\mathbf{x}^{(t)} \in \mathcal{X}_\gamma$ and $\mathbf{x}^{(t+1)} \in \mathcal{X}_\mu$, then

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*) \leq c(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*))$$

where

$$c = \begin{cases} \frac{\mu_0}{\gamma}, & \text{if } \gamma > 2\mu_0 \\ 1 - \frac{\gamma}{4\mu_0}, & \text{if } \gamma \leq 2\mu_0. \end{cases}$$

The rate of convergence bounds derived from Theorem 3.2 can be tighter than those derived from Proposition 2.7 in [32]. We provide a more detailed discussion in Appendix A.3.

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 essentially the same rate of convergence bound as gradient descent algorithm, which differs only for a constant factor.

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 3.1. *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 3.1 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 (2.6), 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.

3.2. *Maximum likelihood estimation.* For the Bradley-Terry model, the log-likelihood function can be written as:

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

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. We use $\lambda_i(\mathbf{A})$ to denote the i -th smallest eigenvalue of matrix \mathbf{A} .

LEMMA 3.2. *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*

$$\gamma = c_{\omega} \lambda_2(\mathbf{L}_{\mathbf{M}}) \text{ and } \mu = \frac{1}{4} \lambda_n(\mathbf{L}_{\mathbf{M}})$$

where $c_{\omega} = 1/(e^{-\omega} + e^{\omega})^2$.

Proof of Lemma 3.2 is provided in Appendix A.4.

By Lemma 3.2, the smoothness parameter μ is proportional to the largest eigenvalue of the Laplacian matrix $\mathbf{L}_{\mathbf{M}}$. By Gershgorin circle theorem, e.g. Theorem 7.2.1 in [16], we have $\lambda_n(\mathbf{L}_{\mathbf{M}}) \leq 2d(\mathbf{M})$. 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 $\lambda_n(\mathbf{L}_{\mathbf{M}})$. This is a tight characterization up to constant factors. When \mathbf{M} is a graph adjacency matrix, then $\lambda_n(\mathbf{L}_{\mathbf{M}}) \geq d(\mathbf{M}) + 1$ [17]. In the context of comparisons, $d(\mathbf{M})$ has an intuitive interpretation as the maximum number of observed paired comparisons involving an item.

Since the log-likelihood function (3.1) satisfies conditions (C1) and (C2) of Lemma 3.1, combining with Lemma 3.2, we observe that the negative log-likelihood function satisfies the γ -PL inequality on \mathcal{W}_{ω} with $\gamma = c_{\omega} a(\mathbf{M})$. Furthermore, by Lemma 3.2, 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 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:

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

We next consider the classic MM algorithm given in (2.4), which can be written as: for $i = 1, 2, \dots, n$,

$$(3.3) \quad w_i^{(t+1)} = \log\left(\sum_{j=1}^n d_{i,j}\right) - \log\left(\sum_{j=1}^n \frac{m_{i,j}}{e^{w_i^{(t)}} + e^{w_j^{(t)}}}\right).$$

The classic MM algorithm is derived for the surrogate function $\underline{\ell}(\mathbf{x}; \mathbf{y})$ given by (2.3). This surrogate function has the following smoothness 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.5. As a side remark, we note that $\underline{\ell}$ is a first-order surrogate function of ℓ on $[-\omega, \omega]^n$ with δ as defined in Lemma 3.3. We discuss this in more details in Appendix A.6.

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 which in general admits different values. This shows that both gradient descent and MM algorithm have linear convergence, and convergence time bound (3.2).

3.3. *Maximum a posteriori probability estimation.* Following Caron and Doucet [8], we consider the MAP estimation of \mathbf{w} with the prior distribution of product-form and marginal distributions such that e^{w_i} has a Gamma distribution with the shape parameter α and the inverse scale parameter β ,

$$\mathcal{G}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

The log-aposteriori 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 that is given as follows

$$(3.4) \quad \ell_0(\mathbf{w}) = \sum_{i=1}^n ((\alpha - 1)w_i - \beta e^{w_i}).$$

For the choice of parameters $\alpha = 1$ and $\beta = 0$, the log-aposteriori probability function corresponds to the log-likelihood function.

LEMMA 3.4. *The negative log-aposteriori probability function for the Bradley-Terry model and the $\text{Gamma}(\alpha, \beta)$ prior is γ -strongly convex and μ -smooth on $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : \|\mathbf{w}\|_\infty \leq \omega\}$ with*

$$\gamma = e^{-\omega}\beta \text{ and } \mu = \frac{1}{4}\lambda_n(\mathbf{L}_\mathbf{M}) + e^\omega\beta.$$

Proof of Lemma 3.4 is provided in Appendix 3.4.

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 aposteriori 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*

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

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

$$(3.5) \quad T = O\left(\left(1 + \frac{d(\mathbf{M})}{\beta}\right) \log(1/\epsilon)\right).$$

This bound can be arbitrarily large by taking parameter β to be small enough. We will show later 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 of Caron and Doucet [8]. 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 ℓ in (2.3) and ℓ_0 is the prior log-likelihood function (3.4).

The MM algorithm can be written as: for $i = 1, 2, \dots, n$,

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

Indeed, this iterative optimization algorithm corresponds to the classic MM algorithm for ML estimation (3.3) 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*

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

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 (3.5).

Tightness of the rate of convergence bound. 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 maximum a posteriori estimate of parameter \mathbf{w}^* can be computed in an explicit form given as follows:

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

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

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

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:

$$(3.6) \quad s^{(t+1)} = \frac{m + 2(\alpha - 1)}{m + \beta s^{(t)}} s^{(t)}.$$

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 the prior $\text{Gamma}(\alpha, \beta)$ distribution.

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

$$(3.7) \quad \rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)}) = 2(\alpha - 1)(a^{(t)} - \log(1 + a^{(t)})).$$

From (3.6), 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 (3.7), $\rho(\mathbf{w}^*) - \rho(\mathbf{w}^{(t)}) = (\alpha - 1)(a^{(t)})^2(1 + o(1))$ for large t , and thus

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

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

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

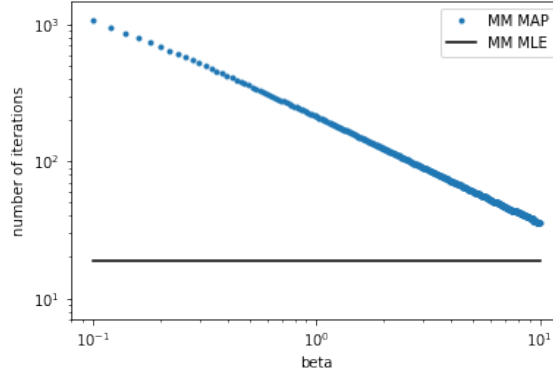


FIG 1. *Illustrative numerical example: number of iterations until convergence versus parameter β . The MM algorithm can be slower for the MAP objective for several orders of magnitude than for the ML objective. The smaller the value of parameter β , the slower the convergence for MAP.*

The rate of convergence is approximately $1 - 4(\alpha - 1)/m$ for small $\alpha - 1$. By taking value of α such that $\alpha - 1 = c\beta$ 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 the tightness of the rate of convergence bound in Corollary 3.4.

An illustrative numerical example. Consider an instance with 10 items with each distinct pair of items compared 10 times and the input data generated according to the Bradley-Terry model of paired comparisons with parameter \mathbf{w} such that $w_1 = \dots = w_5 = -1/2$ and $w_6 = \dots = w_{10} = 1/2$. We run the MM algorithm with random initial value until the smallest $t \geq 0$ such that $\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\|_\infty \leq \epsilon$, for $\epsilon = 0.0001$. The results in Figure 1 show that the MM algorithm for the MAP estimation can be slower for several orders of magnitude than for the ML estimation.

4. Accelerated MAP inference. In this section, we present an acceleration method for gradient descent and MM algorithms for MAP estimation. This acceleration method amounts to performing a transformation of the parameter estimator in each iteration in a way that guarantees certain properties of the objective function along the trajectory of the iterative optimization algorithm.

Let $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a given mapping. We define the Π -transformed gradient descent algorithm by the iteration:

$$(4.1) \quad \mathbf{x}^{(t+1)} = \Pi(\mathbf{x}^{(t)} - \eta f(\mathbf{x}^{(t)})).$$

Similarly, we define the Π -transformed MM algorithm by the iteration:

$$(4.2) \quad \mathbf{x}^{(t+1)} = \Pi(\arg \min_{\mathbf{x}} g(\mathbf{x}; \mathbf{x}^{(t)})).$$

4.1. General convergence theorems. Assume that f is a convex function that satisfies the following two conditions on a convex set \mathcal{X} that contains optimum point \mathbf{x}^* , for a given vector \mathbf{d} :

- (F1) f is μ -smooth on \mathcal{X} ;
- (F2) f satisfies the γ -PL inequality on $\mathcal{X}_0 = \mathcal{X} \cap \{\mathbf{x} \in \mathbb{R}^n : \nabla f(\mathbf{x})^\top \mathbf{d} = 0\}$.

Assume that Π satisfies the following two conditions:

- (P1) $f(\Pi(\mathbf{x})) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$;
- (P2) $\nabla f(\Pi(\mathbf{x}))^\top \mathbf{d} = 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

Condition (P1) means that applying the transformation Π at a point cannot increase the value of function f . Condition (P2) means that at any Π -transformed point, the gradient of function f is orthogonal to vector \mathbf{d} .

We have the following two theorems.

THEOREM 4.1 (Gradient descent). *Assume that f satisfies (F1) and (F2), Π satisfies (P1), and $\eta = 1/\mu$. Let $\mathbf{x}^{(t+1)}$ be according to the Π -transformed gradient descent algorithm (4.1). Then, if $\mathbf{x}^{(t)} \in \mathcal{X}_0$ and $\mathbf{x}^{(t+1)} \in \mathcal{X}$, 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 4.1 is provided in Appendix A.8.

THEOREM 4.2 (MM). *Assume that f satisfies (F1) and (F2), Π satisfies (P1), and g is a majorant surrogate function of f such that $g(\mathbf{x}; \mathbf{y}) - f(\mathbf{x}) \leq \frac{\delta}{2} \|\mathbf{x} - \mathbf{y}\|^2$. Let $\mathbf{x}^{(t+1)}$ be according to the Π -transformed MM algorithm (4.2). Then, if $\mathbf{x}^{(t)} \in \mathcal{X}_0$, we have*

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

We next present a lemma which will be instrumental in showing that the PL condition in (F1) holds for the MAP estimation problem. Let $\mathbf{P}_{\mathbf{d}} = \mathbf{I} - \frac{1}{\|\mathbf{d}\|^2} \mathbf{d} \mathbf{d}^\top$, for a vector $\mathbf{d} \in \mathbb{R}^n$ such that $\mathbf{d} \neq \mathbf{0}$. Note that $\mathbf{P}_{\mathbf{d}}$ is the projection matrix onto the space orthogonal to \mathbf{d} .

LEMMA 4.1. *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex, twice-differentiable function. Let \mathcal{X} be a convex set, $\mathcal{X}_0 = \mathcal{X} \cap \{\mathbf{x} \in \mathbb{R}^n : \mathbf{d}^\top \nabla f(\mathbf{x}) = 0\}$ for a given vector \mathbf{d} , and $\mathbf{x}^* \in \mathcal{X}_0$ be a minimizer of f . Assume that*

- (A1) $\nabla^2 f(\mathbf{x}) \succeq \mathbf{A}_{\mathcal{X}}$ for all $\mathbf{x} \in \mathcal{X}$, for a positive semidefinite $\mathbf{A}_{\mathcal{X}}$, and
 (A2) $\mathbf{u}^\top \mathbf{A}_{\mathcal{X}} \mathbf{v} = 0$ for $\mathbf{u} = (\mathbf{I} - \mathbf{P}_{\mathbf{d}})\mathbf{z}$ and $\mathbf{v} = \mathbf{P}_{\mathbf{d}}\mathbf{z}$, for all $\mathbf{z} \in \mathbb{R}^n$.

Then, f satisfies the γ -PL inequality on \mathcal{X}_0 for all $\gamma \leq \gamma_0$ with

$$\gamma_0 := \min_{\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} : \mathbf{d}^\top \mathbf{x} = 0} \frac{\mathbf{x}^\top \mathbf{A}_{\mathcal{X}} \mathbf{x}}{\|\mathbf{x}\|^2}$$

where note that γ_0 is the smallest eigenvalue of $\mathbf{A}_{\mathcal{X}}$ corresponding to an eigenvector orthogonal to \mathbf{d} .

Proof of Lemma 4.1 is provided in Appendix A.9.

4.2. *Convergence rate for the Bradley-Terry model.* Let $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : \|\mathbf{w}\|_\infty \leq \omega\}$ and Π be defined as

$$(4.3) \quad \Pi(\mathbf{x}) = \mathbf{x} + c(\mathbf{x})\mathbf{1}$$

where

$$(4.4) \quad c(\mathbf{x}) = \log \left(\frac{\alpha - 1}{\beta} n \right) - \log \left(\sum_{i=1}^n e^{x_i} \right).$$

We first discuss conditions (P1) and (P2). For a MAP problem with the log-likelihood function ℓ and the prior log-likelihood function ℓ_0 , let f be the negative log-aposteriori probability function $f(\mathbf{w}) = -(\ell(\mathbf{w}) + \ell_0(\mathbf{w}))$. Suppose that $\nabla \ell(\mathbf{w})^\top \mathbf{1} = 0$, for all $\mathbf{w} \in \mathbb{R}^n$, a condition that holds for all models considered in this paper. Then, with $\mathbf{d} = \mathbf{1}$, (P1) and (P2) are equivalent to, respectively, (a) $\rho(\mathbf{x}) \leq \rho(\Pi(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^n$ and (b) $\nabla \ell_0(\Pi(\mathbf{x}))^\top \mathbf{1} = 0$, for all $\mathbf{x} \in \mathbb{R}^n$.

We show that Π satisfies (P1) and (P2), which is stated in the following lemma.

LEMMA 4.2. *For all $\mathbf{w} \in \mathbb{R}^n$,*

$$(4.5) \quad \rho(\Pi(\mathbf{w})) \geq \rho(\mathbf{w})$$

and

$$(4.6) \quad \nabla \rho(\Pi(\mathbf{w}))^\top \mathbf{1} = 0.$$

Proof of Lemma 4.2 is provided in Appendix A.10.

We next show that f satisfies conditions (F1) and (F2). Condition (F1) holds because we have already shown that f is μ -smooth on \mathcal{W}_ω with $\mu = d(\mathbf{M})/2 + e^\omega\beta$. Condition (F2) can be shown to hold by Lemma 4.1 as follows. Assume $\mathbf{d} = \mathbf{1}$. Note that we have $\nabla^2(f(\mathbf{w})) \succeq \mathbf{A}_{\mathcal{W}_\omega}$, for all $\mathbf{w} \in \mathcal{W}_\omega$, where $\mathbf{A}_{\mathcal{W}_\omega} = c_\omega \mathbf{L}_\mathbf{M} + e^{-\omega}\beta \mathbf{I}$. The assumptions of Lemma 4.1 hold: (A1) holds because, clearly, $\mathbf{A}_{\mathcal{W}_\omega}$ is a positive semidefinite matrix, and (A2) holds because $\mathbf{u}^\top \mathbf{L}_\mathbf{M} \mathbf{v} = 0$ (which follows from $\mathbf{L}_\mathbf{M} \mathbf{1} = \mathbf{0}$) and $\mathbf{u}^\top \mathbf{I} \mathbf{v} = \mathbf{u}^\top \mathbf{v} = 0$ (because \mathbf{u} and \mathbf{v} are orthogonal). Note that we have

$$\gamma_0 = c_\omega \lambda_2(\mathbf{L}_\mathbf{M}) + e^{-\omega}\beta.$$

Hence, by Lemma 4.1, it follows that f satisfies condition (F2) with $\gamma = c_\omega \lambda_2(\mathbf{L}_\mathbf{M}) + e^{-\omega}\beta$.

From Theorem 4.1, we have the following corollary:

COROLLARY 4.1 (Gradient descent). *Suppose that iterates are according to the Π -transformed gradient descent (4.1) for the negative log-aposteriori probability function of the Bradley-Terry model of paired comparisons, with product-form prior distribution such that $e^{w_i} \sim \text{Gamma}(\alpha, \beta)$, $\alpha > 1$ and $\beta > 0$, and $\eta = 2/(\|\mathbf{M}\|_\infty + 2e^\omega\beta)$. Then, we have*

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

where $c_\omega = 1/(e^{-\omega} + e^\omega)^2$.

From Theorem 4.2, we have the following corollary:

COROLLARY 4.2 (MM). *Suppose that iterates are according to the Π -transformed MM (4.1) for the negative log-aposteriori probability function of the Bradley-Terry model of paired comparisons, with product-form prior distribution such that $e^{w_i} \sim \text{Gamma}(\alpha, \beta)$, $\alpha > 1$ and $\beta > 0$. Then, we have*

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

where $c_\omega = 1/(e^{-\omega} + e^\omega)^2$.

PROOF. Condition (F1) holds for $-\rho$ because we have already shown that $-\rho$ is μ -smooth with $\mu = d(\mathbf{M})/2 + e^\omega\beta$ on \mathcal{W}_ω and $\delta = e^{2\omega}d(\mathbf{M})/2$. Condition (F2) holds by Lemma 4.1 with $\gamma = c_\omega a(\mathbf{M}) + e^{-\omega}\beta$. Conditions (P1) and (P2) hold by (4.5) and (4.6) in Lemma 4.2, respectively. \square

Algorithm 1 Accelerated MM algorithm

```

1: Initialization:  $\epsilon, \theta, \theta^{\text{prev}}$ 
2: while  $\|\theta - \theta^{\text{prev}}\|_\infty > \epsilon$  do
3:    $\theta^{\text{prev}} \leftarrow \theta$ 
4:   for  $i = 1, 2, \dots, n$  do
5:      $\theta_i^{\text{temp}} = \left( \alpha - 1 + \sum_{j \neq i} d_{i,j} \right) \left( \beta + \sum_{j \neq i} \frac{m_{i,j}}{\theta_i + \theta_j} \right)^{-1}$  ▷ standard MM update
6:   end for
7:   for  $i = 1, 2, \dots, n$  do
8:      $\theta_i = \frac{\alpha-1}{\beta} n \frac{\theta_i^{\text{temp}}}{\sum_{j=1}^n \theta_j^{\text{temp}}}$  ▷ acceleration rescaling
9:   end for
10: end while

```

Note that in the limit of small β , the convergence rate bounds in Corollaries 4.2 and 4.2 correspond to the bounds for the ML estimation in Corollaries 3.1 and 3.2, respectively. From Corollaries 4.2 and 4.2, it follows that for accelerated gradient descent and accelerated MM algorithm, the convergence time satisfies

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

For the Bradley-Terry model of paired comparisons with parametrization $\theta = (\theta_1, \dots, \theta_n)^\top$, where $\theta_i = e^{w_i}$ for $i = 1, 2, \dots, n$, the transformation Π given by (4.3) is equivalent to a simple rescaling as shown in a procedural form in Algorithm 1.

The illustrative numerical example revisited. We ran the accelerated MM algorithm for our illustrative example, and obtained the results shown in Figure 2. By comparing with the corresponding results obtained by the MM algorithm with no acceleration, shown in Figure 1, we observe that the acceleration resolves the slow convergence issue and that it can yield a significant reduction of the convergence time.

5. Generalized Bradley-Terry models. The results presented in previous 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 characterizations 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 B.

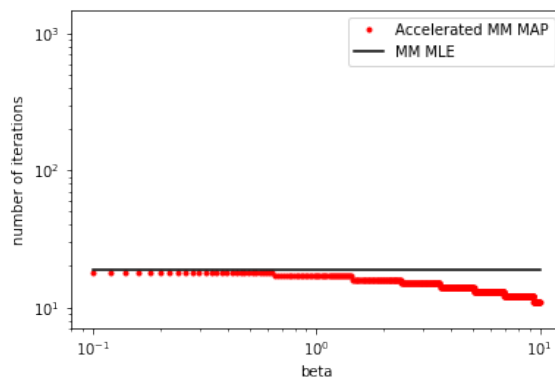


FIG 2. *The illustrative example revisited: accelerated MM resolves the convergence issue for MAP estimation: it has faster or equal convergence than for ML estimation.*

6. Numerical results. In this section we present evaluation of convergence times of MM algorithms for different generalized Bradley-Terry models using different real-world datasets. Our goal is to provide empirical validation of some of the hypotheses derived from our theoretical analysis. Overall, our numerical results validate that (a) the convergence of the MM algorithm for MAP estimation can be much slower than for ML estimation, (b) MM algorithm for MAP estimation has convergence time that increases as parameter β of the prior distribution decreases, and (c) a significant reduction of the convergence time can be achieved by the accelerated MM algorithm defined in Section 4.

6.1. Datasets. We consider three datasets, which vary in the type of data, size and sparsity. 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 web service by the MIT Media Lab as part of the PlacePulse project [41]. 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. We did this 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 a Kaggle chess rat-

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	0.671
GIFGIF: C (sample)	734	256	28	0.569
GIFGIF: H (sample)	1040	251	23	1.357
Chess (full)	65,030	8,631	155	0
Chess (sample)	13,181	985	135	1.773
NASCAR	64,596	83	1,507	39.338

ings competition [44]. Each observation contains information for a match between two players including unique identifiers of the two 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 a web page maintained by Hunter [23]. This dataset was previously used for evaluation of MM algorithms for the Plackett-Luce ranking model by Hunter [24] as more recently by Caron and Doucet [8]. We used this dataset to evaluate convergence times of MM algorithms for the Plackett-Luce ranking model.

We show a summary of some key properties for 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.

6.2. Experimental results. We evaluated convergence times defined as the number of iterations that an algorithm takes until a convergence criteria is reached. We use standard convergence criteria based on the difference of successive parameter vector estimates. Specifically, the convergence time T is defined as the smallest integer $t > 0$ such that $\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|_\infty \leq \epsilon$, for

TABLE 2
Number of iterations for MM algorithm and accelerated MM algorithm (AccMM).

Dataset	Algorithm	$\beta = 0$	0.01	0.1	1	10
GIFGIF: A (full)	MM	MLE	572	125	70	16
	AccMM	non-existent	509	123	42	13
GIFGIF: C (full)	MM	MLE	733	150	49	13
	AccMM	non-existent	551	93	37	13
GIFGIF: H (full)	MM	MLE	1,127	149	98	21
	AccMM	non-existent	1,044	159	51	18
GIFGIF: A (sample)	MM	145	854	177	26	7
	AccMM		125	81	22	7
GIFGIF: C (sample)	MM	130	694	151	39	9
	AccMM		111	78	36	9
GIFGIF: H (sample)	MM	216	1,234	237	38	8
	AccMM		146	72	26	8
Chess (full)	MM	MLE	2,217	581	113	33
	AccMM	non-existent	2,291	302	49	25
Chess (sample)	MM	121	122	91	74	19
	AccMM		117	93	48	16
NASCAR	MM	11	695	971	58	10
	AccMM		11	11	10	6
NASCAR ($\epsilon = 10^{-5}$)	MM	14	1,528	2,069	105	16
	AccMM		14	14	12	7
NASCAR ($\epsilon = 10^{-6}$)	MM	17	2,362	3,223	157	23
	AccMM		17	16	14	8
NASCAR ($\epsilon = 10^{-8}$)	MM	22	4,029	5,544	261	36
	AccMM		22	21	18	11

a given parameter $\epsilon > 0$, with initial value $\mathbf{w}^{(0)} = \mathbf{0}$. In our experiments, we used 10^{-4} as the default value for parameter ϵ . For NASCAR dataset, we also present results for several other values of parameter ϵ to demonstrate how the convergence times change with this parameter. In our experiments, we also evaluated convergence times in real processor time units. We observed that they validate all the claims that we had made based on the convergence times in the number of iterations, and hence we do not further discuss them.

The numerical results are presented in Table 2. The results are presented for MM algorithm and accelerated MM algorithm and different values of parameter β . In all our experiments we set the value of parameter α such that $1 - \alpha = \beta$, which corresponds to fixing the mode of the Gamma prior distribution to value 1. The value of parameter $\beta = 0$ corresponds to ML estimation. In Table 2, we highlighted some values with boldface for which significant speed ups are achieved by either using the MM algorithm for ML estimation or using the accelerated MM algorithm for MAP estimation.

For GIFGIF datasets, we observe that in all cases, the convergence time increases as the value of parameter β decreases for $\beta > 0$. For the values considered, this increase can be up to two orders of magnitude. When the ML estimate exists (for sampled data), we observe that the MM algorithm for ML estimation converges much faster than the MM algorithm MAP estimation for sufficiently small values of parameter β . We also observe that a significant reduction of the convergence time can be achieved by the accelerated MM algorithm. This reduction can be for as much as to nearly 10% of the convergence time of the MM algorithm without acceleration. These empirical results validate claims suggested by our theoretical results.

For Chess datasets, all the observations derived from using the GIFGIF datasets remain to hold.

For NASCAR dataset, we show results for different values of parameter ϵ , including the default value of 10^{-4} . Again, all the observations made for GIFGIF and Chess datasets remain to hold. It is noteworthy that the MM algorithm for ML estimation converges much faster than for MAP estimation for sufficiently small values of parameter β . This is especially emphasised for smaller values of parameter ϵ . For the cases considered, this can be for as much as three orders of magnitude. Similarly, the accelerated MM algorithm converges much faster than standard MM algorithm with no acceleration.

7. Discussion. We have shown that for all generalized Bradley-Terry models considered in this paper, both gradient descent algorithm and MM algorithm for ML estimation have linear convergence with the convergence time bound $O(d(\mathbf{M})/a(\mathbf{M}))$. We can interpret this bound to be proportional to the condition number of the Laplacian matrix $\mathbf{L}_\mathbf{M}$. We have also shown that for all generalized Bradley-Terry models considered in this paper, both gradient descent and MM algorithm for the MAP estimation with the product-form prior distribution with $\text{Gamma}(\alpha, \beta)$ marginal distributions, the convergence time is linear with the convergence time bound $O(d(\mathbf{M})/\beta)$. This bound 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 $\mathbf{L}_\mathbf{M}$. Our results identify a slow convergence issue of standard gradient descent and MM algorithms for MAP estimation, which occurs for small values of parameter β . The smaller values of parameter β correspond to more vague prior distribution. Our results identify a discontinuity of the convergence time at the point $(\alpha, \beta) = (1, 0)$, which corresponds to the ML estimation. The accelerated method for MAP estimation resolves the slow convergence issue, yielding a convergence time that is bounded by the best of what is achieved for ML and MAP estimation.

TABLE 3

The values of parameters $d(\mathbf{M})$ and $a(\mathbf{M})$ for some examples of matrix \mathbf{M} that is the adjacency matrix of a graph G with n vertices

G	$d(\mathbf{M})$	$a(\mathbf{M})$	$\frac{d(\mathbf{M})}{a(\mathbf{M})}$
complete	$n - 1$	n	$\Theta(1)$
star	$n - 1$	1	$\Theta(n)$
circuit	2	$2(1 - \cos(\pi/n)) = \frac{4\pi^2}{n^2}(1 + o(1))$	$\Theta(n^2)$
path	2	$2(1 - \cos(2\pi/n)) = \frac{\pi^2}{n^2}(1 + o(1))$	$\Theta(n^2)$

For illustration purposes, in Table 3 we show values of $d(\mathbf{M})$ and $a(\mathbf{M})$ for some examples of matrix \mathbf{M} that is adjacency matrix of a graph. We observe that when each distinct pair is compared the same number of items (complete graph case), the convergence time is $T = O(\log(1/\epsilon))$. For other cases considered, the convergence time is $T = O(n^c \log(1/\epsilon))$, for $c \geq 1$.

We can derive an upper bound on the convergence time which depends only on some well-known properties of a graph associated with matrix \mathbf{M} . Let \mathbf{A} be adjacency matrix of graph G which has edge (i, j) if, and only if, $m_{i,j} > 0$. Let $r = \bar{m}/\underline{m}$ where $\bar{m} = \max_{i,j} m_{i,j}$ and $\underline{m} = \min\{m_{i,j} : m_{i,j} > 0\}$. Let $d(n)$ be the maximum degree and $D(n)$ be the diameter of G . Then, for both gradient descent algorithm and MM algorithm for ML estimation, we have the convergence time bound:

$$(7.1) \quad T = O(rd(n)D(n)n \log(1/\epsilon)).$$

Obviously, this implies the convergence time bound $T = O(rn^3 \log(1/\epsilon))$ for any connected graph G , which follows by trivial facts $d(n) \leq n$ and $D(n) \leq n$. The bound in (7.1) is derived by using the known lower bound on the algebraic connectivity of a Laplacian matrix $\lambda_2(\mathbf{L}_\mathbf{A}) \geq 4/(nD(n))$, e.g. see Theorem 3.4 in [34]. We provide details of derivations in Appendix A.13.

SUPPLEMENTARY MATERIAL

Proofs and extensions to generalized Bradley-Terry models

(.). In the supplementary material, we provide proofs of theorems and lemmas for the Bradley-Terry model of paired comparisons, and show how the convergence rate results can be extended to generalized Bradley-Terry models.

References.

- [1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics, 2 edition, 2002.
- [2] D. Balduzzi, K. Tuyls, J. Pérolat, and T. Graepel. Re-evaluating evaluation. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS '18)*, Montreal, Canada, December 2018.

- [3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [4] R. A. Bradley. Rank analysis of incomplete block designs: Ii. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4):502–537, 1954.
- [5] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. method of paired comparisons. *Biometrika*, 39(3/4):324–345, Dec 1952.
- [6] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, Nov. 2015.
- [7] C. J. Burges, R. Ragno, and Q. 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.
- [8] F. Caron and A. Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *J. Comp. Graph. Statist.*, 21(1):174–196, 2012.
- [9] Y. Chen and C. 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, Lille, France, 07–09 Jul 2015.
- [10] H. A. David. *The Method of Paired Comparisons*. Charles Griffin and Company, London, 1963.
- [11] O. 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.
- [12] O. Dykstra, Jr. Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. *Biometrics*, 16(2):176–188, 1960.
- [13] A. E. Elo. *The Rating of Chessplayers*. Ishi Press International, 1978.
- [14] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.
- [15] L. R. Ford. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.
- [16] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 4 edition, 2013.
- [17] 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.
- [18] J. Guiver and E. Snelson. Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*, Montreal, Canada, 2009.
- [19] B. Hajek, S. Oh, and J. 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.
- [20] R. Herbrich, T. Minka, and T. Graepel. TrueskillTM: A Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS '06)*, pages 569–576. Canada, 2006.
- [21] T.-K. Huang, C.-J. Lin, and R. C. Weng. Ranking individuals by group comparisons. *J. Mach. Learn. Res.*, 9:2187–2216, 2008.
- [22] T.-K. Huang, R. C. Weng, and C.-J. Lin. Generalized Bradley-Terry Models and multi-class probability estimates. *J. Mach. Learn. Res.*, 7:85–115, Dec. 2006.
- [23] D. R. Hunter. MATLAB code for Bradley-Terry models, <http://personal.psu.edu/drh20/code/btmatlab/>, 2003.
- [24] D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 32(1):384–406, 2004.
- [25] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

- [26] E. Kaye and D. Firth. BradleyTerryScalable: Fits the Bradley-Terry Model to Potentially Large and Sparse Networks of Comparison Data. <https://cran.r-project.org/web/packages/BradleyTerryScalable/>.
- [27] A. Khetan and S. Oh. Computational and statistical tradeoffs in learning to rank. In *Advances in Neural Information Processing Systems 29*, pages 739–747. 2016.
- [28] K. Lange. *MM Optimization Algorithms*. SIAM, 2016.
- [29] 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.
- [30] H. Li. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool, 2011.
- [31] R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.
- [32] J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.
- [33] L. Maystre. Choix: Inference algorithms for models based on Luce’s choice axiom. In *GitHub*: <https://github.com/lucasmaystre/choix>, 2018.
- [34] R. Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197-198:143 – 176, 1994.
- [35] S. Negahban, S. Oh, and D. Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.
- [36] Y. Nesterov. Gradient methods for minimizing composite objective functions. *Math. Program.*, 140:125–161, 2012.
- [37] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.
- [38] B. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3:864–878, 12 1963.
- [39] A. Rajkumar and S. 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, Beijing, China, 22–24 Jun 2014.
- [40] 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.
- [41] T. Rich, K. Hu, and B. Tome. GIFGIF, <http://gifgif.media.mit.edu>, 2018.
- [42] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *J. Mach. Learn. Res.*, 17(1):2049–2095, Jan. 2016.
- [43] G. Simons and Y.-C. 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.
- [44] J. Sonas. Kaggle competition: Chess ratings - elo versus the rest of the world, <https://www.kaggle.com/c/chess>, 2010.
- [45] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.
- [46] H. Turner and D. Firth. Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software, Articles*, 48(9):1–21, 2012.
- [47] M. Vojnovic and S.-Y. 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.
- [48] F. Wauthier, M. Jordan, and N. 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.

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

DEPARTMENT OF STATISTICS
LONDON SCHOOL OF ECONOMICS
HOUGHTON STREET
LONDON, WC2A 2AE
UNITED KINGDOM
E-MAIL: m.vojnovic@lse.ac.uk
E-MAIL: k.zhou3@lse.ac.uk

DEPARTMENT OF INDUSTRIAL AND SYSTEM ENGINEERING
KAIST
291 DAEHAK-RO, YUSEONG-GU
DAEJEON
SOUTH KOREA
E-MAIL: yunseyoung@kaist.ac.kr